TUM School of Engineering and Design
Technische Universität München

Deep Learning for Cloud Removal in Spaceborne Earth Observation

Patrick Ebel

Vollständiger Abdruck der von der TUM School of Engineering and Design der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr.-Ing. Liqiu Meng

Prüfer*innen der Dissertation:

1. Prof. Dr.-Ing. habil. Xiaoxiang Zhu
2. Prof. Dr.-Ing. habil. Michael Schmitt
3. Prof. Dr. Konrad Schindler

Die Dissertation wurde am 16.05.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Engineering and Design am 14.09.2023 angenommen.

"Über den Wolken muss die Freiheit wohl grenzenlos sein.
Alle Ängste, alle Sorgen, sagt man,
Blieben darunter verborgen und dann
Würde, was uns groß und wichtig erscheint,
Plötzlich nichtig und klein."

— Reinhard Mey, Über den Wolken (1974)

Abstract

The natural occurrence of clouds, haze, and other atmospheric disturbances poses a persistent obstacle for passive spaceborne sensors to seamlessly observe our planet. Concretely, most optical satellites can only receive sun-radiated signals reflected from Earth's surface at daytime and if the weather allows to. Due to its fundamental and long-standing nature, the restoration of satellite images by reconstructing cloud-covered pixels is an established problem in signal processing and remote sensing. As of more recently, solutions based on deep neural networks tackle the problem in a data-driven manner. While accomplishing substantial improvements over prior approaches, the novel techniques still suffer from drawbacks and limitations. First, their need for copious amounts of training data is oftentimes not met. Moreover, the common focus on narrowly defined regions of interest is in contrast with the diversity of Earth and the ambition to provide a generally applicable solution for cloud removal. Finally, existing techniques may inaccurately reconstruct a given satellite image, but thus far there is a lack in methods for obtaining indications of potential error at inference time. These are among the key challenges that cloud removal is posing today, and which are subsequently addressed herein. Specifically, this dissertation promotes a better and more faithful reconstruction of optical satellite images by making the following five key contributions:

- **Handling clouds in practice.** To begin with, the practical implications of haze, clouds and cloud shadow in the context of an established remote sensing use case are investigated. For this sake, the effects of clouds on optical satellite image scene classification are systematically explored. The performances and confidences of neural networks commonly utilized for this sake are assessed, and the outcomes are analyzed via a dedicated interpretability analysis. This serves to investigate the importance of carefully handling cloud-covered observations, and points towards further directions of handling the problem in practice.
- **Data and Benchmarks for cloud removal in optical satellite imagery.** To promote cloud removal with the aim of making it applicable on real data in the wild and anywhere on Earth, two datasets are gathered for training and evaluating solutions. First, *SEN12MS-CR* is a large-scale, global and all-season dataset of paired radar and (non-)cloudy multi-spectral optical satellite image triplets for mono-temporal cloud removal. Second, to support the endeavor of multi-temporal cloud removal, *SEN12MS-CR-TS* is curated. SEN12MS-CR-TS builds upon SEN12MS-CR to provide time series data for high quality image reconstruction. The benefits and value of both datasets are highlighted by the subsequent contributions, which extensively rely and build upon this initial effort.

- **Methods for mono-temporal multi-sensor cloud removal.** Building on the curated SEN12MS-CR dataset, a total of three different deep neural architectures fusing radar and optical observations for mono-temporal cloud removal are proposed: First, a multi-modal *residual network* architecture that encourages the preservation of cloud-free pixels through a custom loss designed for satellite image reconstruction. Second, a *Generative Adversarial Network* which learns a bi-directional relationship between optical and radar sensors to restore cloud-obscured pixels. Finally, a *visual transformer* that locally fuses both sensors and reconstructs multi-spectral observations via a global attention mechanism. Altogether, these architectures advance mono-temporal cloud removal and demonstrate the benefits of multi-sensor fusion for satellite image reconstruction.
- **Methods for multi-temporal multi-sensor cloud removal.** Historical data are a precious source of information to better reconstruct the rich multi-spectral information contained in optical satellite imagery. Hence, two solutions are proposed using a time series approach to satellite image reconstruction: First, a *sequence-to-point* model that learns to integrate cloud-free information over a time series of cloudy optical data. Second, a deep prior network to address the *sequence-to-sequence* problem of translating cloudy to clear time series while preserving temporal resolution. As in the mono-temporal case, it is shown how auxiliary radar observations are facilitating the image reconstruction problem.
- **Calibrated uncertainty predictions for cloud removal.** The aforementioned contributions allow for assessing cloud removal approaches in a general purpose setting by measuring grand average performances of image reconstruction. Yet, practical applications may necessitate goodness estimates on a sample-by-sample basis. To address this need, the final contribution of this thesis is in developing UnCRtainTS, a novel multi-temporal multi-sensor architecture introducing uncertainty estimation to multi-spectral satellite image reconstruction. Experimental evaluations show that UnCRtainTS learns well-calibrated uncertainty predictions and that uncertainty-based filtering allows for risk-sensitive control of the empirical image reconstruction error.

In sum, this thesis provides a complete treatment on the topic of cloud removal in optical satellite imagery. Following a more detailed summary of the individual contributions outlined above, the work closes by providing a conclusion, further outlook on the subject and proposals for future research on the topic.

Zusammenfassung

Das natürliche Auftreten von Wolken, Dunst und anderen atmosphärischen Störungen stellt ein Hindernis für die lückenlose Beobachtung unseres Planeten mittels passiven satellitengestützten Sensoren dar. Konkret bedeutet dies, dass die meisten optischen Satelliten die von der Erdoberfläche reflektierten Sonnenstrahlen nur tagsüber empfangen können und wenn das Wetter es zulässt. Aufgrund der grundlegenden und langjährigen Natur des Problems, ist die Wiederherstellung von Satellitenbildern mittels der Rekonstruktion wolkenbedeckter Pixeln ein etabliertes Aufgabenfeld in der Signalverarbeitung und Fernerkundung. In jüngster Zeit wurden Lösungen auf der Grundlage von tiefen neuronalen Netzen entwickelt, welche das Problem datengetrieben angehen. Obwohl die neuen Techniken gegenüber früheren Ansätzen erhebliche Verbesserungen erzielen, leiden sie immer noch unter Nachteilen und Einschränkungen. Erstens wird ihr Bedarf an großen Mengen von Trainingsdaten oft nicht genügend gedeckt. Darüber hinaus steht der übliche Fokus auf räumlich eng definierte Areale im Gegensatz zur Vielfalt der Erde, ihrer unterschiedlichen Bodenbedeckung und dem Bestreben, eine allgemein anwendbare Lösung für die Wolkenentfernung zu entwickeln. Schließlich ist es möglich, dass die vorhandenen Techniken ein bestimmtes Satellitenbild ungenau rekonstruieren—aber bisher fehlt es an Methoden, um Hinweise auf mögliche Fehler zum Zeitpunkt der Inferenz zu erhalten. Dies sind einige der wichtigsten Herausforderungen, die die Wolkenentfernung heutzutage mit sich bringt, und die im Folgenden behandelt werden. Diese Dissertation leistet einen Beitrag zu einer besseren und getreueren Rekonstruktion von optischen Satellitenbildern, indem sie folgende fünf Schlüsselbeiträge liefert:

- **Der Umgang mit Wolken in der Praxis.** Anfangs werden die Auswirkungen von Dunst, Wolken und Wolkenschatten im Zusammenhang eines etablierten Anwendungsfall der Fernerkundung untersucht. Zu diesem Zweck wird der Effekt von Wolken auf die Klassifizierung von Szenen in optischen Satellitenbildern systematisch erforscht. Die Leistungen und die Konfidenz von üblicherweise zu diesem Zweck eingesetzten neuronalen Netzen werden bewertet und die Ergebnisse werden mittels einer netzwerkbasierten Interpretierbarkeitsanalyse gedeutet. Dies dient dazu, die Wichtigkeit eines sorgfältigen Umgangs mit wolkenbedeckten Beobachtungen zu untersuchen, und weist letztlich auf weitere Richtungen für den Umgang mit dem Problem in der Praxis hin.
- **Daten und Benchmarks zur Wolkenentfernung in Satellitenbildern.** Um Wolkenentfernung auf reale Daten überall auf der Welt zu ermöglichen, werden zwei Datensätze für die Entwicklung neuer Methoden erstellt. Erstens, *SEN12MS-CR*, ein groß angelegter, globaler und ganzjähriger Datensatz von gepaarten Radar-

und (nicht-)bewölkten multispektralen optischen Satellitenbild-Tripletts für die mono-temporale Wolkenentfernung. Zweitens wurde *SEN12MS-CR-TS* kuratiert, um den multi-temporalen Ansatz der Wolkenentfernung zu fördern. SEN12MS-CR-TS baut auf SEN12MS-CR auf, um Zeitreihendaten für eine hochwertige Bildrekonstruktion bereitzustellen. Die Vorteile und der Wert beider Datensätze werden in den nachfolgenden Beiträgen deutlich, die weitgehend darauf aufbauen.

- **Methoden zur mono-temporalen Multi-Sensor-Wolkenentfernung.** Aufbauend auf dem kuratierten SEN12MS-CR Datensatz werden insgesamt drei verschiedene tiefe neuronale Architekturen vorgeschlagen, die Radar- und optische Beobachtungen zur mono-temporalen Wolkenentfernung zusammenführen. Zuerst eine multimodale *Residualnetzwerkarchitektur*, welche die Erhaltung wolkenfreier Pixel durch eine speziell für die Satellitenbildrekonstruktion entwickelte Kostenfunktion fördert. Zweitens ein *generatives adversariales Netzwerk*, das die bidirektionale Beziehung zwischen optischen und Radarsensoren lernt, um wolkenverhangene Pixel wiederherzustellen. Zuletzt ein *visueller Transformer*, der beide Sensoren lokal fusioniert und multispektrale Beobachtungen über einen globalen Aufmerksamkeitsmechanismus rekonstruiert. Allesamt setzen diese Architekturen neue Standards für die mono-temporale Wolkenentfernung und demonstrieren die Vorteile der Multisensor-Fusion für die Rekonstruktion von Satellitenbildern.
- **Methoden zur multi-temporalen Multisensor-Wolkenentfernung.** Historische Daten sind eine wertvolle Informationsquelle, um optischen Satellitenbilder und deren reichhaltige multispektrale Information besser zu rekonstruieren. Es werden zwei Lösungen vorgeschlagen, die einen Zeitreihenansatz zur Rekonstruktion von Satellitenbildern verwenden: Erstens ein *Sequenz-zu-Punkt-Modell*, das lernt, wolkenfreie Information über eine Zeitreihe von bewölkten optischen Daten zu integrieren. Zweitens ein Deep-Prior-Netz, dass das *Sequenz-zu-Sequenz-Problem* löst, bewölkte in klare Zeitreihen unter Beibehaltung deren zeitlicher Auflösung zu übersetzen. Wie im mono-temporalen Fall zeigt sich, dass zusätzliche Radarbeobachtungen die Bildrekonstruktion erleichtern.
- **Kalibrierte Unsicherheitsvorhersagen für die Wolkenentfernung.** Bisherige Ansätzen der Wolkenentfernung werden lediglich anhand der durchschnittlichen Qualität ihrer Bildrekonstruktion gemessen. Praktische Anwendungen können jedoch Güteabschätzungen einzelner Stichproben erfordern. Um diesen Bedarf zu decken, wird UnCRtainTS entwickelt, eine neue multi-temporale Multisensorarchitektur, die eine Einschätzung pixelweiser Unsicherheiten der Rekonstruktion multispektraler Satellitenbilder liefert. Experimentelle Auswertungen zeigen, dass UnCRtainTS gut kalibrierte Unsicherheitsvorhersagen erlernt und eine risikosensitive Kontrolle des empirischen Bildrekonstruktionsfehlers ermöglicht.

Allesamt liefert diese Arbeit eine umfassende Behandlung des Themas der Wolkenentfernung in optischen Satellitenbildern. Nach einer umfangreicheren Zusammenfassung der Beiträge schließt die Arbeit mit einer Schlussfolgerung, einem weiteren Ausblick und Vorschlägen für künftige Forschungen zu ihrem Thema.

Acknowledgements

Working throughout the doctoral process and towards this dissertation has been an exciting, joyful and at times challenging journey. I am glad for all the experiences and friendships I was able to make in this period of my life. It won't be feasible to reiterate it all in here or to thank every of the numerous people who supported me along the way, but I wish to pronounce my gratitude to some very special ones in particular:

First, I would like to thank my supervisors, who have been with me from the start until the end and whose efforts are at the core of our contributions. Specifically, I wish to thank Prof. Dr. Xiaoxiang Zhu for her outstanding and unconditional support of our work, her building of an enabling research group and a stimulating environment that made our contributions possible, and for providing me the opportunity to work under her guidance in this community. I also wish to thank Prof. Dr. Michael Schmitt for always being there for me, his encouraging and consistently positive outlook as well as his expertise he so kindly shared. Without the combined inspiration and support of my supervisors on both a professional and personal level, this dissertation would not have been feasible.

Second, I wish to express my gratitude to all the fellow researchers I had the chance to become friends with or collaborating with. This includes my co-authors, office mates and all the colleagues I had the pleasure of getting to know during my time at TUM, in its surroundings or even abroad. I'm glad for all our time together, throughout our shared efforts and discussions I learned how to conduct research. Particularly, I would like to thank Prof. Dr. Jan Dirk Wegner and Dr. Vivien Sainte Fare Garnot, who hosted me during my research stay in Zurich, and all the lovely people I was able to meet during the visit. This was a memorable experience and it was a joy to meet you.

Third, I wish to thank Prof. Dr. Konrad Schindler acting as an external examiner and the third member of my thesis committee. Furthermore, I would like to extend my thanks to Prof. Dr. Liqiu Meng, taking up the role of the chair of my doctoral defense. I very much appreciate their availability, time and interest in assessing this work.

Finally, I wish to thank my parents, family and friends. I am endlessly thankful for your love, encouragement and support that made all these efforts feasible.

Patrick Ebel
Munich, May 2023

Contents

List of Figures

List of Tables

Acronyms

AOI	Area of Interest.
BN	Batch Norm.
CNN	Convolutional Neural Network.
CPU	Central Processing Unit.
ESA	European Space Agency.
FLOPS	Floating Point Operations Per Second.
GAN	Generative Adversarial Network.
GEE	Google Earth Engine.
GPU	Graphical Processing Unit.
MAE	Mean Absolute Error.
MBConv	MobileNet Convolutional block.
MSE	Mean Squared Error.
PSNR	Peak Signal to Noise Ratio.
ReLU	Rectified Linear Unit.
RGB	Red Green Blue.
ROI	Region of Interest.
S1	Sentinel-1.
S2	Sentinel-2.
SAR	Synthetic Aperture Radar.
SAR2OPT	SAR-to-Optical.
SSIM	Structural Similarity.
SVM	Support Vector Machine.
VGG16	Visual Geometry Group 16 Network.

1 Introduction

Remote sensing is the measurement of the properties and characteristics of an object of interest, at a distance and without any requirements of a direct physical contact. As such, remote sensing allows for an analysis of Earth at a global scale. Specifically, multi-spectral satellites provide data about the molecular material composition of our planet's surface, at an extent unmatched by any other sensors. Yet, one of the main challenges of a seamless monitoring of our planet via optical satellite imagery is the existence of clouds, which obscure the line of sight between the sensor and a region of interest. This obstacle poses a fundamental limitation for passive optical spaceborne sensors, which are conventionally employed to provide multi-spectral observations at an otherwise high availability. The central goal of this dissertation is in developing high-quality and reliable automated image reconstruction approaches to provide analysis-ready multi-spectral data at any time—even, and particularly, in the presence of clouds.

Due to cloud-occluded vision posing a long-standing and fundamental problem for the remote sensing community, there readily exist numerous established approaches to cloud removal which share the core motivations of this work. However, the established solutions are oftentimes tailored to specific regions of interest, may not provide sufficient image restoration quality or lack in trustworthiness and reliability. In short, they do not yet meet the goals outlined below. Yet, these prior contributions in many respects serve as valuable starting points for the work at hand. Reaching the outlined aims by building on these existing solutions for the following motivations is what this thesis aspires to.

1.1 Motivation

Thanks to ongoing technological advancements, such as ever-new sensors in combination with products becoming more or easier available to scientific staff and industry clients alike, Earth observation has entered a golden age [1, 2]. One ambition of spaceborne remote sensing at a global scale is to offer high quality measurements of any location on our planet at any given time. This accomplishment—at least for the workhorse of remote sensing; optical satellites—is fundamentally impeded by the natural occurrence of haze, clouds and other atmospheric disturbances. Passive optical sensors, making up the bulk of imaging instruments operational in the field [3, 4], measure a region's molecular material composition by receiving the sun's radiation reflected from the surface area of a region of interest. This path of transmission however is interrupted in the presence of clouds. On average, about 67 % of our planet and 55 % of its land surface is covered by

clouds [5]. The issue is even more pronounced for regions close to the equator (such as rainforests [6]) and during meteorological winter season, when cloud coverage can persist for extended periods [5]. The principal motivation of this dissertation is in resolving this fundamental shortcoming of spaceborne optical satellites by developing an automated approach to faithfully and reliably reconstruct cloud-covered satellite images.

The consequences of cloud coverage directly affect subsequent applications relying on clear multi-spectral imagery. For instance, the interpretability of obscured images by a human observer is severely impaired—dense clouds cover the underlying ground area and cloud shadows darken any affected pixels. Consequently, another motivation for cloud removal is to restore human interpretability by recovering the hidden information. Moreover, noisy data likewise affects the computer-assisted automatization of remote sensing downstream tasks. Most automated processing is developed on satellite images acquired under ideal conditions, free of clouds and any other noise, such that data at inference time is assumed to likewise reflect these ideal conditions. By addressing the primary issue of cloud coverage in optical satellite imagery, such strict requirements may be relaxed and the detrimental effects of clouds on Earth observation may be eased—which serves as a final motivation for this thesis.

1.2 Objectives

The primary objective of this dissertation is to provide general-purpose means to make multi-spectral optical satellite imagery usable where and whenever needed—even, and particularly, in the presence of clouds, haze and other atmospheric disturbances. To achieve said goal, this dissertation builds on modern data-driven computer-based image reconstruction techniques, by which the information in cloud-covered and otherwise noisy pixels can be recovered. The developed methodology is meant to be general-purpose in the sense of being applicable to every region of interest on Earth, without being confined to any particular kinds of land cover or any specific meteorological season.

A secondary goal is in providing a rich and general framework for developing and evaluating such methodology. Contemporary machine learning approaches are able to achieve state-of-the-art performances, but rely on extensive amounts of diverse and high quality data. Therefore, an aim of this thesis is in providing a curated dataset for training data-intense deep neural networks on the task of cloud removal and allowing them to generalize to any other region of interest. Likewise, the performance of any method should be quantified such that it is indicative of translating to any other yet unseen region of interest. Hence, any curated dataset is required to reflect the richness of our planet in a diverse test split for benchmarking. The availability of such datasets and benchmarks will subsequently promote the development of better cloud removal approaches, assessed and compared with one another in a more competitive environment.

A final aim of this dissertation is in making cloud removal techniques more safe and reliable. Remote sensing is a safety-critical domain, with many of its applications ne-

cessitating precise, reliable and trustworthy data to draw conclusions on. First, it is instructive to raise awareness about the impacts of clouds on common remote sensing applications. Second, the process of cloud removal itself needs to become more reliable. Currently, it is common practice for any satellite image reconstruction method to solely be evaluated in terms of its *average* reconstruction goodness, which provides an overall indicator of its quality but only a coarse guide of its reliability. However, practical applications may rely on predictors of quality on a *per-sample* basis, at inference time and indicative of any potential reconstruction errors. This critical need is currently not met by any existing approaches and filling this gap is the last goal of this dissertation.

1.3 Dissertation Outline

The structure of this dissertation is as follows: Chapter 1 serves as an introduction to the topic, outlining the motivation and goals. Chapter 2 provides background knowledge, such as the underlying methodology and remote sensing specifics. In chapter 3, related work is reviewed such that the contributions of this work can be considered in their context. In chapter 4, a summary of the publications constituting this cumulative dissertation are provided. Finally, chapter 5 provides a summary and further outlook on the subject. Appended are publications constituting this cumulative dissertation.

Furthermore, the interested reader is referred to the website of `https://patrickTUM.github.io/cloud_removal/`, which serves as a project page accompanying the publications constituting this work. In particular, it features the benchmark tables reported herein at the time of writing this thesis, but additionally allows to include and communicate future results as well as distributing any related source code.

Figure 1.1: QR code referencing to `https://patrickTUM.github.io/cloud_removal/`. The website provides code accompanying the research herein, and may communicate update benchmarking results.

2 Theoretical Background

This chapter serves to familiarize the reader with core concepts of this work, and to convey an appreciation for the problems that haze or clouds pose for optical Earth observation—but also what it takes to address these challenges. Accordingly, the first section of this chapter introduces key concepts from the broader machine learning literature as well as architectures that served as starting points or building blocks for the methods proposed herein. This includes convolutional and generative architectures, both temporal and spatial attention mechanisms, as well as likelihood optimization for uncertainty quantification in machine learning. What follows in the second section is to convey an understanding of clouds and their properties, the characteristics of radar plus optical spaceborne imaging and how the former causes problems to the latter. Finally, background knowledge on image reconstruction for remote sensing is communicated, wherein the cloud removal problem at the heart of this thesis is defined in combination with important metrics for image reconstruction quality assessment.

2.1 Deep Learning at a Glance

Deep learning, originating from the study of artificial neural networks [7, 8, 9, 10], is too broad a field to be covered in depth within the scope of this dissertation. Therefore, this section focuses on providing a brief overview of specific topics and methods relevant to follow the work at hand. In particular, this includes convolutional architectures and Generative Adversarial Networks introduced in section 2.1.1, attention-based networks that are covered in section 2.1.2 and likelihood-based uncertainty optimization as outlined in section 2.1.2. For further background knowledge, the interested reader is referred to the popular textbooks of [11, 12, 13].

2.1.1 Convolutional Neural Networks and Generative Adversarial Networks

Convolutional Architectures. Following the seminal success of AlexNet [14] in the ImageNet competition [15], deep neural networks established themselves as state-of-the-art on various computer vision benchmarks. While their architectures may vastly differ, a commonality shared among many of them is the usage of learnable filters. A common implementation of learnable filters is in terms of the discrete convolution operation. The discrete convolution operation over two-dimensional tensors I, K is defined as

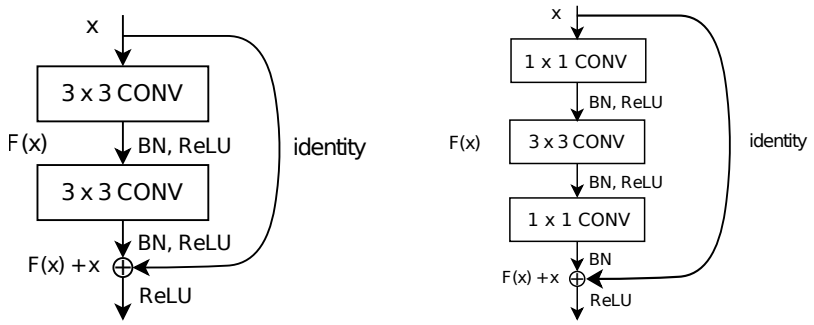


Figure 2.1: Residual connections. Left: residual block [22]. Right: MBConv block [24]. Both configurations utilize residual mappings, but the MBConv block decomposes the multi-channel convolution layers into separate multi-channel point-wise and per-channel depth-wise operations. Figures adapted from [22].

$$(I * g)(i, j) = \sum_a \sum_b I(a, b)K(i - a, j - b) \tag{2.1}$$

where i, j pertain to indices of the respective dimensions, shifted by integer steps a, b [13]. K is conventionally referred to as a filter or kernel of learnable parameters, optimized according to some cost function and applied onto image I. In practice, kernels may integrate information over more than one input channel, and involve adjustments such as padding or a varying stride [16]. Furthermore, implementations that slide along more than two dimensions are referred to as 3D convolutions and may be helpful for purposes such as video inpainting [17, 18], where spatio-temporal context is required.

Influential convolutional neural network architectures relevant for this thesis are: VGG16 [19], that advanced deeper architectures and is commonly used as an encoder for style or perceptual losses in the image generation context [20]. U-Net [21], which is an encoder-decoder architecture processing information along two separate 'where' and 'what' pathways. Residual networks [22] and variations thereof such as MobileNets [23, 24], which model differences within the input-output mapping via residual connections. Finally, Generative Adversarial Networks [25], consisting of a tandem of networks oftentimes involving U-Nets, to learn sampling from a high-dimensional target distribution of data. Of particular importance for this dissertation are residual architectures and Generative Adversarial Networks, so they are subsequently introduced in greater detail.

Residual networks are an influential architecture in computer vision, and for satellite image reconstruction in remote sensing [26, 27]. At the heart of residual networks is the residual block, consisting of two or more convolutional layers constituting a mapping $F(x)$ and a residual connection. While the two convolutional layers process information in the usual sequential manner, the residual connection fast-forwards the input tensor x and adds it onto the output of $F(x)$, such that the resulting mapping is $F(x)+x$. Hence, the network only models the residual change applied to x, making it easier to learn the

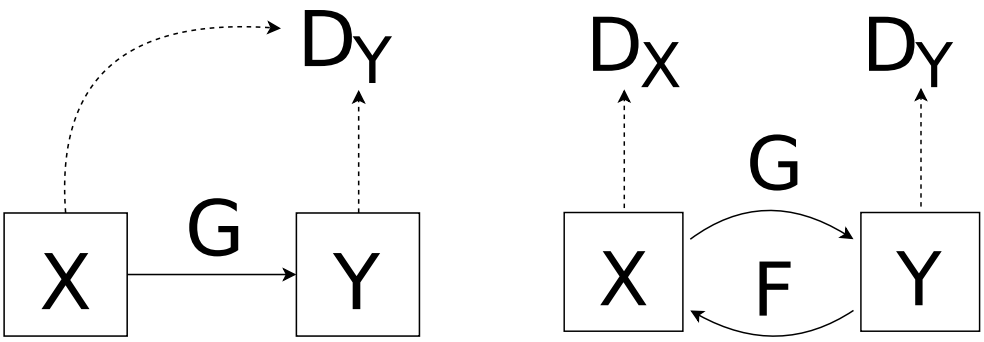


Figure 2.2: GAN architectures. Left: Conditional GAN [31], consisting of a tandem of a generator G and a discriminator D. Right: Cycle GAN [32] features two generators G and F mapping between two domains, plus two discriminators D_X and D_Y to classify real versus generated data in each domain. Figures adapted from [32].

identity mapping and facilitating optimization [28]. The residual block structure is illustrated on the left in Figure 2.1. In practice, residual connection are particularly appealing if only little modifications need to be applied—such as in the case of removing semi-transparent haze [29] or, as in the context of this thesis, for removing clouds.

A variation on the residual block, referred to as MBConv block and depicted on the right in Figure 2.1, is introduced by [24]. The MBConv building block is a parameter-efficient rearrangement of the original residual block, separating the convolutions into point-wise and depth-wise operations. First applying $[1 \times 1]$ kernels across all channels followed by $[H \times W]$ kernels on a channel-wise basis reduces the filters' dimensionality and thus the model's overall memory requirements. The efficiency of MBConv blocks is beneficial in particular for networks that may otherwise consume prohibitive amounts of memory, while the cost in performance compared to conventional spatio-spectrally operating convolution blocks is minimal [23, 24, 30].

Generative Adversarial Networks learn a generative model of an distribution implicitly described by training samples, such that target data can be drawn from this distribution [33]. The classical GAN architecture is a tandem consisting of a generator G and a discriminator network D_Y as depicted on the left of Figure 2.2. Both networks are in competition with one another, where it is the generator's task to synthesize realistic samples from the target distribution Y and the discriminator's objective is to classify whether provides samples are from the empirical distribution Y or the distribution induced by G. Moreover, both networks can be conditioned on data from a source distribution X, such that the learned distribution is conditional rather than marginal [31]. In the pix2pix approach of [25], both the input and the output of a U-Net generator [21] as well as the input to the discriminator are an image, such that the network can be used for image-to-image translation. Training this tandem involves minimizing an adversarial loss [33], for which many improvements have been suggested [34, 35, 36] with LSGAN [37] being of particular interest for this thesis. It is defined as

$$\mathcal{L}_{LSGAN}(D) = \frac{1}{2}\mathbb{E}_{y\sim Y}[(D(y) - b)^2] + \frac{1}{2}\mathbb{E}_{x\sim X}[(D(G(x)) - a)^2] \tag{2.2}$$

$$\mathcal{L}_{LSGAN}(G) = \frac{1}{2}\mathbb{E}_{x\sim X}[(D(G(x)) - c)^2] \tag{2.3}$$

where x and y are samples drawn from their respective distributions X and Y, a and b refer to the labels encoding real versus generated data and c is the value of the label whose data G generates. That is, G and D are simultaneously optimized in a zero-sum game until reaching equilibrium in a game theoretical sense [33]. Conventionally, supervised training of G involves a linear combination of adversarial and L1 losses [25].

Finally, Cycle-GAN [32] introduces a cycle-consistent loss term, which encourages a bijective mapping between the domains of the input and the target distribution. That is, after translating from the input to the target and back again, the original input should be recoverable. For this purpose, Cycle-GAN uses two generators G and F as well as two discriminators D_X and D_Y in a configuration as shown on the right of Figure 2.2. Cycle consistency is then learned in both directions via an auxiliary objective function

$$\mathcal{L}_{cycle}(G, F) = \mathbb{E}_{x\sim X}[||F(G(x) - x)||_1] + \mathbb{E}_{y\sim Y}[||G(F(y) - y)||_1] \tag{2.4}$$

Notably, cycle consistency provides two benefits: First, it allows for learning a bi-directional relationship between X and Y, which may teach a tighter coupling between two related modalities compared to simple conditioning [38]. Second, cycle consistency provides pixel-level supervision to the networks without the requirement of paired images across both domains. This is particularly useful in any cases where other supervision is unavailable due to a lack in pixel-wise correspondences across domains [32] .

2.1.2 Attention

While convolutional architectures have been the convention in computer vision and recurrent architectures were the dominant paradigm in natural language processing, the recent emergence of the attention mechanism [39] has revolutionized both fields alike. Principally, attention refers to a correlation-based feature extraction mechanism with a global receptive field. It is given by

$$attention(q, k, v) = softmax(\frac{qk^T}{\sqrt{d_k}})v \tag{2.5}$$

where q denotes a $[d_k \times 1]$ query vector, k is a $[d_k \times 1]$ key vector, v denotes a $[d_v \times 1]$ vector of values and $\sqrt{d_k}$ is a scaling factor to counteract diminishing gradients. The outcome is a $[d_v \times 1]$ attention vector of the weighted values. Intuitively, the formula correlates queries of interest q by their associated keys v via dot-product, normalized the weighted vector, which is in return utilized to attend to particularly relevant values v. In

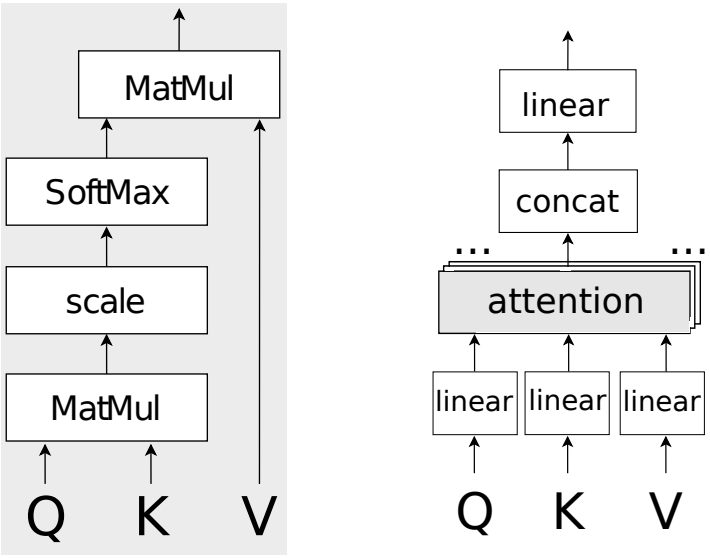


Figure 2.3: Attention mechanism, a correlative operator to weight values v by associating queries q with their closest keys k. Left: The attention operand as formalized in equation 2.5. Right: Multi-head attention, with h parallel heads of attention as shown on the left. Before each scaled dot-product attention operation, every head has a dedicated linear layer and their outputs get concatenated before being processed by a final linear transformation. Illustrations adapted from [39].

practice, q, k and v are learned according to the task at hand. Furthermore, while this operand describes a single so-called attention head, it is typically generalized to matrix operations implementing h heads in parallel. The concept is shown in Figure 2.

While attention is very expressive and neither constrained to e.g. a local view nor limited by technicalities such as fading gradients, its power may come at a cost. Specifically, its runtime of $\mathcal{O}(n^2 d)$ in sequence length n and feature dimension d can be prohibitively expensive, requiring engineering prowess to make the approach affordable nonetheless. To resolve this, recent approaches in remote sensing have e.g. proposed to compute attention on downsampled feature maps [40] or within local windows combined with global shifts [41]. Moreover, attention lacks the inductive bias of convolutions, which may be beneficial in common computer vision tasks. Therefore, an attention-based model has to learn previously hard-wired properties such as translation equivariance from the data, which requires ample training resources.

Within the scope of this dissertation, of particular importance are temporal attention as formalized above and originally proposed in [39], as well as visual attention as suggested in [42]. As visual attention also suffers from a computational complexity quadratic in the sequence length, which poses a hurdle for applying it to image tensors, the attention operation of equation 2.5 is mostly applied in modified forms. In a simple adaptation [42], images are divided into smaller patches that are separately encoded to summarize several pixels, and handled as individual tokens by the established mechanism. A more

recent modification, of particular relevance for the visual transformer in this thesis, implements attention on hierarchical feature maps by merging patches in deeper layers. Furthermore, attention is computed within windows and, to maintain global interactions, windows are shifted systematically in subsequent layers to vary neighborhoods [41]. This briefly summarizes the approaches to attention most relevant for this thesis.

2.1.3 Uncertainty Quantification in Machine Learning

Modeling probability distributions via neural networks is an established approach in machine learning [12, 43, 44]. Uncertainty quantification is an established technique in safety-critical applications, such as biomedical imaging [45]. Specifically, with respect to image reconstruction, there is an increase in awareness of the challenges of solving inverse problems [46] and uncertainty-based solutions to address these in a risk-aware manner [47, 48, 49, 50].

Uncertainty can be distinguished into epistemic and aleatoric uncertainty [51]. Epistemic or model uncertainty originates from the uncertainty in model weights, which are due to randomness at initialization and stochastics during the training process. For neural networks, it may be estimated via e.g. deep ensembles [52] or monte-carlo dropout [53]. Aleatoric or data uncertainty is due to noise in the data or its labels [54, 55, 56, 57, 58]. While epistemic uncertainty may be explained away under optimal conditions and in the limit of infinite data, aleatoric uncertainty is inherent to the data at hand.

Within the scope of this dissertation, aleatoric uncertainty is learned to be predicted by a deep neural network trained via a negative log-likelihood loss [12]

$$\mathcal{L}_{NLL}(x, \theta) = -\sum_{j=1}^{n} log(\Psi(x, \theta)) \ . \tag{2.6}$$

such that the reggression loss evaluating a single variable μ is now a cost function on the variables θ of a parametric noise distribution Ψ. For its simplicity and generality, a Gaussian noise assumption is made such that the aleatoric uncertainty on the reconstructed pixel is modeled with a K-variate Normal distribution centered at the predicted value $\hat{\boldsymbol{y}}_j$ and with positive definite covariance matrix $\boldsymbol{\Sigma}$:

$$\mathcal{N}(\boldsymbol{y}_j|\hat{\boldsymbol{y}}_j, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}(2\pi)^{\frac{K}{2}}} exp\left(-\frac{1}{2}\|\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j\|_M\right) \ , \tag{2.7}$$

with $\|.\|_M$ the Mahalanobis distance, defined as:

$$\|\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j\|_M = (\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j)^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j) \ . \tag{2.8}$$

Subsequently, the negative log likelihood loss writes as:

$$\mathcal{L}_{NLL}(\boldsymbol{y_j}|\hat{\boldsymbol{y}}_{\boldsymbol{j}}, \boldsymbol{\Sigma}) \propto \sum_{j=1}^{n} log(|\boldsymbol{\Sigma}_j|) + \|\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j\|_M \,. \tag{2.9}$$

Complementary, epistemic uncertainty in this thesis is described by an ensemble of M models. For this sake, $|M|$ neural networks are trained with different weight initializations and under differing batch draws, as originally suggested in [52]. The individual Gaussian fits are then combined to approximate a single, unimodal and centered Normal distribution. For mean estimates $\hat{y}^m$ and variance predictions $(\sigma^m)^2$ of the collection of members $m = 1, ..., |M|$, the ensembled predictions are then given by

$$\hat{\boldsymbol{y}}^M = \frac{1}{M} \sum_{m=1}^{M} \hat{\boldsymbol{y}}^m \tag{2.10}$$

$$(\sigma^M)^2 = \frac{1}{M} \sum_{m=1}^{M} (\sigma^m)^2 + \frac{1}{M} \sum_{m=1}^{M} (\hat{\boldsymbol{y}}^m)^2 - (\hat{\boldsymbol{y}}^M)^2 \tag{2.11}$$

Recently, uncertainty quantification became a trending topic in remote sensing [59], with applications to e.g. biomass or flood hazard monitoring [60, 61, 62].

2.2 Clouds: Definition & Properties

Clouds are defined as a visible mass of drops of water, suspended in the atmosphere [63]. Their physical properties are constituted by about 100 million of such droplets within each cubic meter of air, and every droplet being about 10 micrometers in size. It's these properties which pose an obstacle for passive spaceborne sensors, because the mass of tiny particles scatters light in any direction and makes the cloud opaque. Moreover, clouds are associated with precipitation such as rain or snow, which itself can obscure a satellite's view. The extent to which clouds block the view depends on their *optical thickness*, which may range from constituting solely a filmy and semi-transparent layer to a dense and total occlusion of all that lays below [64, 65, 66]. Optical thickness, together with the shape and location of clouds allows to categorized them into different types. Notably, the primary discrepancy between haze and clouds is their altitude, so both terms may often be used interchangeably. The nuances in the nature and severity of coverage are remarkable in the context of image reconstruction, as it differentiates cloud-covered pixels from other kinds of noise that may be more dichotomous in nature [67, 68, 69], and conveys the diversity of the problem that optical satellites face.

Clouds are a naturally occurring weather phenomenon. They form in conditions of the atmosphere being saturated by water vapor, through condensation or evaporation. That is, the air reaches maximum humidity and holds as much suspended water droplets or ice particles as it can [70]. All in all, the global cloud fraction averages to about 67 % [71].

The coverage over land is at 55 %, with distinct seasonal and geographical variation [5]. In particular, the cloud fraction is elevated during meteorological winter months and over equatorial regions. That is, some land cover such as rainforests [6] may be obscured by clouds throughout most of the year because of their geographical location, or for extended periods due to seasonality. In such cases, optical sensors are constrained and the remote sensing practitioner may refer to alternative information sources.

2.3 Basics of SAR and Optical Imaging

This section gives a brief overview of optical and radar remote sensing, with its two imaging modalities being of a central importance in this thesis. In both cases the focus is on spaceborne sensors measuring the interaction of electromagnetic radiation with Earth's surface. Yet, the two sensor modalities differ in many other regards, among which are the recorded wavelengths and the measured physical versus molecular properties, the discrepancy in active compared to passive sensor usage, and their difference in viewpoint directions. Figure 2.4 provides an illustration of the differences of both modalities and the challenges of interpreting SAR imagery. While the multi-sensory observations are clearly pertaining to multiple views of a common region of interest, the multimodal pairings differ considerably and offer complementary sources of information [72].

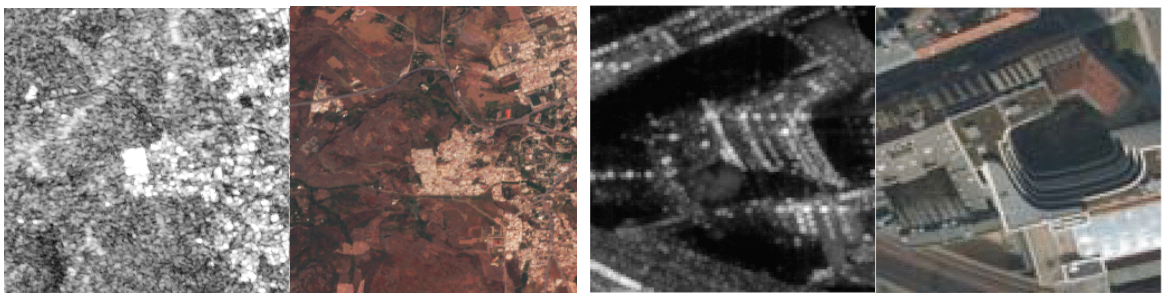


Figure 2.4: Imaging differences between optical and radar sensors. Left: co-registered Sentinel-1 and Sentinel-2 observations, visualizations from the data of [73]. Right: corresponding TerraSAR-X satellite measurements and aerial UltraCAM photography, plots from [74]. In both cases, the differences between modalities are apparent. The discrepancies get more apparent at higher spatial resolution.

2.3.1 Optical Imaging

Optical satellite sensors measure electromagnetic waves between circa 10^{-8} to 10^{-6} meters length, including the visible light range and beyond. In practice, most optical satellite sensors measure solar lights, reflected off the surface of Earth that is being imaged. While there exist active optical spaceborne systems emitting radiant energy themselves [75], the majority of satellite missions mount passive sensors capturing sun-emitted reflections. The optical sensor's passivity implies they only function during daytime and,

most critically, their view can not penetrate haze or clouds. Given the frequency of cloud coverage, this poses a fundamental limitation of optical satellite sensors.

Optical sensors make up the majority of spaceborne imaging [4], featured by missions such as Landsat, Modis/Aqua constellation, Planet Labs Doves, Airbus SPOT 6/7, Airbus Pléiades or, of primary importance for this thesis, Sentinel-2 from ESA's Copernicus mission. Sentinel-2 measures reflective radiance in 13 spectral bands at down to 10 meters resolution, with a revisit time of 5 days. Figure 2.5 illustrates the spectrum of Sentinel–2, ranging from visible bands to short-wavelength infrared.

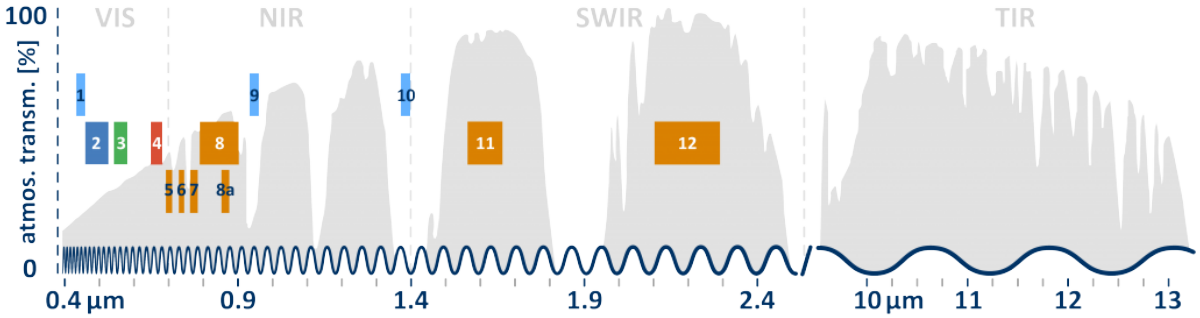


Figure 2.5: Spectral sensitivity of ESA's Sentinel-2 satellites. The instrument captures spectral intensities assigned to 13 separate bands. Figure from [76].

2.3.2 Synthetic Aperture Radar

Imaging Radar sensors measure microwaves of about 10^{-2} meters wavelength, about four orders of magnitude larger than the visible spectrum. Synthetic Aperture Radar (SAR) refers to a moving imaging radar system whose physical antenna is synthetically prolonged along the flight path, resulting in a finer spatial resolution of the measured backscatter. Notably, SAR systems are two-part, composed of a transmitter and a receiver—whereas the aforementioned optical sensors only consist of the latter component. Importantly, the actively emitted microwaves are able to penetrate through clouds before and after backscattering off the imaged land, which makes SAR applicable independent of daytime and weather conditions. The backscattered echoes are received by a radar antenna, measuring complex-valued data containing amplitude and phase information. The recorded signal represents a measure of the imaged scene's reflectivity, as influenced by its physical and electrical properties. Further accounting of the physics underpinning SAR and its signal processing is out of the scope of this work, but the interested reader is referred to the corresponding literature [77, 78, 79].

Examples of common or representative SAR satellites are TanDEM-X, TerraSAR-X and, of primary interest in the context of this dissertation, the two satellites from the Sentinel-1 constellation. Their sensors provide two channels of polarization, pertaining to microwaves that are vertically emitted and either vertically (VV) or horizontally received (VH). Sentinel-1's wide-swath mode spatial azimuth resolution is at 20 meters

and revisits are about every 5-6 days. Figure 2.4 illustrates the discrepancies between the two modalities. Among the factors that pose challenges for interpreting SAR are its sensor-inherent speckle noise [80] and a sideways-looking view, which make radar data challenging to relate to optical imaging in particular at a high resolution.

2.4 Image Reconstruction for Remote Sensing

Following the characterization of optical satellite imagery in section 2.3.1, clouds in section 2.2, and the problems that the latter poses for the former, this section introduces the image reconstruction task as a formalization of the problem of cloud-covered satellite images. For clarity, cloud removal is defined as a regression task and defined as

TASK: CLOUD REMOVAL

Input: A potentially cloudy optical satellite image tensor $\mathcal{I}$ of dimensions $[T_\mathcal{I} \times C_I \times H \times W]$. Optionally, further data $\mathcal{J}$, such as co-registered radar satellite data or cloud masks of dimensions $[T_\mathcal{I} \times C_\mathcal{J} \times H \times W]$, where $C_\mathcal{J}$ denotes the channels of the auxiliary data $\mathcal{J}$.

Output: A cloud-free reconstruction $\hat{\mathcal{I}}$ of I with dimensions $[T_\mathcal{T} \times C_\mathcal{T} \times H \times W]$, where $\mathcal{T}$ denotes a cloud-free and co-registered optical view on the region of $\mathcal{I}$ with dimensions $[T_\mathcal{T} \times C_\mathcal{T} \times H \times W]$. $\hat{\mathcal{I}}$ is optimal if its closeness or similarity to $\mathcal{T}$ as measured under a given metric m is optimal.

For all tensors (ignoring subscripts), T pertains to the temporal dimension, C refers to the spectral channels and H, W are the spatial dimensions of height and width. If $\mathcal{J}$ is given and includes data from a secondary sensor, the cloud removal task is considered to be multi-modal or *multi-sensory*, otherwise it is a *single-sensor* setting. With the datasets part of this distribution featuring optical as well as paired radar data, the multi-sensory case is usually considered unless specified otherwise. In the case of $T_\mathcal{I} = T_\mathcal{T} = 1$ the task is termed *mono-temporal*, else it is referred to as *multi-temporal* cloud removal. In the setting of $T_\mathcal{I} > 1$ but $T_\mathcal{T} = 1$, this is referred to as *sequence-to-point multi-temporal* cloud removal. Finally, the case of $T_\mathcal{I} = T_\mathcal{T} > 1$ is referred to as *sequence-to-sequence multi-temporal* cloud removal. For brevity, unless stated otherwise, sequence-to-point multi-temporal cloud removal is commonly abbreviated as just *multi-temporal* cloud removal.

Note that depending on the context $\mathcal{T}$ may be utilized for supervision or evaluation of the goodness of $\hat{\mathcal{I}}$. Commonly, as a metric m the $\mathcal{L}_1$ or $\mathcal{L}_2$ cost functions are utilized

$$\mathcal{L}_1(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \sum_{j=1}^{n} \|\boldsymbol{x}_j - \boldsymbol{y}_j\|_1 \,, \qquad (2.12) \qquad \mathcal{L}_2(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \sum_{j=1}^{n} \|\boldsymbol{x}_j - \boldsymbol{y}_j\|_2^2 \,, \qquad (2.13)$$

where $\|.\|_1$ and $\|.\|_2$ pertain to the L1 and L2 norms respectively.

For all cloud removal methods benchmarked in this dissertation, their image reconstruction performance is quantitatively evaluated in terms of mean absolute error (MAE) analogous to equation 2.12 or root mean squares error (RMSE) (either of both, following the conventions of the considered benchmark dataset) as well as Peak Signal-to-Noise Ratio (PSNR), structural similarity (SSIM) [81] and the Spectral Angle Mapper (SAM) metric [82]. The metrics are each defined as

$$MAE(x, y) = \frac{1}{C \cdot H \cdot W} \sum_{c=h=w=1}^{C,H,W} |x_{c,h,w} - y_{c,h,w}| \tag{2.14}$$

$$RMSE(x, y) = \sqrt{\frac{1}{C \cdot H \cdot W} \sum_{c=h=w=1}^{C,H,W} (x_{c,h,w} - y_{c,h,w})^2} \tag{2.15}$$

$$PSNR(x, y) = 20 \cdot log_{10} \left(\frac{1}{RMSE(x, y)} \right) \tag{2.16}$$

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + \epsilon_1)(2\sigma_{xy} + \epsilon_2)}{(\mu_x + \mu_y + \epsilon_1)(\sigma_x + \sigma_y + \epsilon_2)} \tag{2.17}$$

$$SAM(x, y) = cos^{-1} \left(\frac{\sum_{c=h=w=1}^{C,H,W} x_{c,h,w} \cdot y_{c,h,w}}{\sqrt{\sum_{c=h=w=1}^{C,H,W} x_{c,h,w}^2 \cdot \sum_{c=h=w=1}^{C,H,W} y_{c,h,w}^2}} \right) \tag{2.18}$$

with images x, y compared via their respective pixel-values $x_{c,h,w}, y_{c,h,w} \in [0, 1]$, dimensions $C = 13$, $H = W = 256$, means μ_x, μ_y, standard deviations σ_x, σ_y, covariance σ_{xy} as well as infinitesimally small constants ϵ_1, ϵ_2 to stabilize the calculations. Both MAE and RMSE are pixel-level metrics and quantify the average discrepancy between target and predicted pixels in units of the measure of interest. PSNR quantifies the signal-to-noise ratio of the prediction as a reconstruction of the target image. SSIM is another image-wise measure that builds on PSNR and captures the structural similarity of the prediction to the target in terms of perceived change, contrast and luminance [81]. Finally, the SAM measure is a image-level metric that provides the spectral angle between the bands of two multi-channel images [82]. Combined, these metrics are meant to provide a complete and multifaceted view on a model's performance.

3 Related Work

Given that haze, clouds or other atmospheric disturbances are persistent and regularly reoccurring weather phenomena, their presence poses a long-standing problem for spaceborne Earth observation. Consequently, the past decades of research in remote sensing have yielded a sizable amount of works addressing the corresponding problem of cloud removal. This chapter serves to provide an overview of the corpus of literature on this subject. For this sake, past contributions are coarsely divided into classical techniques involving designed features and more contemporary approaches based on deep learning, which are more recent but of particular relevance for this thesis. For each of these two paradigms, techniques are categorized as operating either on mono-temporal or multi-temporal data, corresponding to the respective variants of the cloud removal task as formally introduced in chapter 2.4. The focus of this chapter is on the matter of cloud removal, but the interested reader is likewise referred to overviews of closely related tasks, such as super-resolution, missing data reconstruction or denoising [83, 84, 85, 86].

3.1 Methods Based on Hand-Crafted Features

3.1.1 Mono-temporal Approaches

Much of the early work on mono-temporal cloud removal is inspired by related works on *image inpainting* in the classical computer vision literature [87, 88, 89, 90, 91, 92]. Examples of adapting inpainting techniques for cloud removal in aerial or spaceborne optical imagery are given by [93, 94, 95], which use existing methods such as exemplar-based inpainting, multi-scale fragment transplanting or wavelet-based geometric flow propagation. More domain-specific adaptations inpaint pixel-values guided by closest spectral fit [96]. A drawback of inpainting approaches is their reliance on the availability of cloud-free source regions, that are furthermore required to be sufficiently similar to the target region in terms of their structure and texture.

Closely related are *interpolation methods*, which are capable to fill information shrouded even by thick clouds via spatial interference from neighboring or sufficiently close-by cloud-free pixels. This includes techniques based on nearest neighbors [97] or kriging [98]. A limitation of the spatial interpolation paradigm is that clouds may continuously cover large adjacent spaces such that its proximity assumption is oftentimes violated.

Besides the two aforementioned approaches that extrapolate from local information, several techniques rely on *global filtering*. For instance, the authors of [99] perform thin cloud removal via low-frequency homomorphic filtering, and [100] propose filtering cloud-associated principal components. Moreover, the closely related hyperspectral image restoration literature demonstrates many use cases of low-rank tensor decomposition methods to dissect irregular noise from the lower-dimensional signal [101, 102, 103, 104, 105]. Albeit potent, these techniques make strong assumptions on the statistics of the images and their cloud coverage—which may not always be met, such that the cloudy foreground is not sufficiently separable from the underlying land cover.

Finally, a special mention deserve early approaches to *SAR-optical data fusion*, as represented by the works of [106, 107]. The initial work of [106] proposes a cross-modal correlation approach to inpaint cloud-covered pixels with cloud-free optical intensities whose co-registered SAR recordings are closest to those SAR measurements of the pixels to be replaced. Eckart et al. [107] extend on this work by fusing radar data with information from an auxiliary optical image via the closest spectral fit approach [96]. While optical and radar sensors measure different quantities and are thus challenging to relate to one another, these initial efforts influenced subsequent deep learning approaches to SAR-to-optical domain transfer and multi-sensor data fusion [72], not at last including the models introduced in this thesis.

3.1.2 Multi-temporal Approaches

The simplest yet practically relevant approach to dealing with cloud coverage in time series is to choose the most recent, least cloudy of all available observations. While this strategy is appealing in its simplicity and defines a minimum goodness for better approaches to beat, it has its limitations: If no entirely cloud-free observation is available at any time, then the resulting prediction will as well contain noisy pixels. Workarounds to resolve this restriction may be increasing the sampling frequency or going further back in time in the hope of clear views—but cloud coverage may persist for long periods depending on geolocation and seasonality [5], so the closest cloud-free observation may only be found in outdated historical data.

As an alternative strategy, *inpainting approaches* that have been used in the mono-temporal setting can be adapted to transplant information across the spatio-temporal dimensions. For instance, the work of [108] proposes temporal inpainting via Bayesian sparse dictionary learning, which was subsequently extended by [109]. In terms of multi-sensor approaches, [110] extend the previous mono-temporal radar-optical fusion paradigm of [107] to a time series and reconstruct the satellite image by blending it with dictionary patches learned from both modalities.

Another approach that can be adjusted from the single timepoint setting to the multi-temporal scenario are *interpolation techniques*. For example, the authors of [111] introduce a bi-temporal interpolation of pixels close in space or sharing similar spectral properties. In case the interpolation technique includes explicitly inferring across the

temporal dimension, then it is commonly referred to as *temporal mosaicing*. As an example, the work of [112] performs interpolation via a spatial and temporal adjacency weighting and [113] reconstruct cloud-covered pixels via least angle regression over the temporal evolution of a dictionary of cloud-free pixels. The authors of [114] propose two approaches—once, local cloud removal via an ensemble of linear predictors and second, based on support vector machines (SVM). Further works following the SVM regression approach include [115, 116]. Mosaicing abolished the need for any entirely clear sample, but may introduce image artifacts into the end product as a result of the composition process. Furthermore, mosaicing techniques commonly rely on accurate cloud mask, yet cloud detectors themselves may be prone to imperfections [117] and subsequently impact the reconstruction quality.

More principled approaches to satellite image reconstruction involve *algebraic methods* such as variants of principal component analysis [118, 119, 120], non-negative matrix factorization [121, 122, 123], tensor decomposition [124], as well as matrix or tensor completion [125, 126, 127], originating from the broader signal processing literature. Specifically adapted for satellite image reconstruction, a compressed sensing approach to cloud removal in bi-temporal sequences is proposed in [128]. [129] implement robust matrix completion with a temporal consistency constraint, adjusted to handle even extensive cloud cover. Relatedly, [130] propose a weighted modification of low-rank tensor completion to reconstruct cloud-covered information. Based on robust principal component analysis, the authors of [131] remove clouds with a sparsity constraint. The subsequent works of [132, 133] propose low-rank tensor ring decomposition with a total variation regularization. Finally, [134] apply nonnegative matrix factorization to remove clouds based on guidance from nonlocal filters. In sum, there exist a pletora of algebraic techniques to support the endeavor of cloud removal, which still remain popular today. Yet, like many of the other aforementioned handcrafted approaches, they have in common that they are fit specifically to the region of interest. This can be a benefit for case studies, where a single area is of relevance and data is rare. The downturn is the methods need to be re-fitted anew for each other region and can't transfer knowledge from one to another. This promotes failure in inherently challenging cases, which may have otherwise been resolvable by data-driven generalization from earlier scenes.

3.2 Methods Based on Deep Neural Networks

Given their initial success in computer vision, deep neural networks soon after became the predominant approach to many problems in remote sensing, and likewise for the purpose of satellite image reconstruction. Early works adapted established architectures, oftentimes with minimal changes to the backbone, but eagerly experimented with multi-spectral information or auxiliary sensors to overcome challenges very specific to the task of cloud removal. The following two subsections provide an overview of seminal works utilizing deep neural networks for optical satellite image reconstruction in a mono-temporal or multi-temporal setting, respectively.

3.2.1 Mono-temporal Approaches

One of the first deep neural networks for cloud removal is McGAN [135], a generative architecture with a pix2pix backbone [25]. The model maps cloudy RGB and near-infrared spectral bands to a cloud-removed prediction. Alternatively to GAN, the network of [136] was among the first to implement a residual architecture for haze removal. However, a shortcoming of these seminal works and many subsequent contributions is their focus on narrowly defined regions of interest as well as their reliance of synthesized cloudy samples , which makes a generalization to different areas and real world conditions unclear. Finally, it is questionable to which extent purely optical data, including a supplementary near-infrared band, may suffice in recovering information *from* and *for* the very same sensor modality that is critically affected by cloud coverage.

Different from optical inpainting, the *SAR-to-optical (SAR2OPT)* paradigm of [6, 137, 138] takes inspiration from earlier cross-modal approaches [106, 107, 110] and translates radar satellite measurements to cloud-free optical imagery. As SAR is invariant to daylight conditions and robust to atmospheric noise [78] yet different in its measured quantities to optical sensors, the novel problem becomes that of bridging a modality gap. The inpainting aim is thus reframed into a domain transfer objective. While these early works demonstrate the feasibility of mapping radar data to optical information, both modalities principally differ in their measured quantities and not all spectral properties of land cover may be inferred from their corresponding radar backscatter, fundamentally constraining the SAR2OPT framework.

A third approach is in unifying both preceding paradigms, by combining cloudy optical inputs with co-registered SAR recordings. The principal motivation of this *data fusion approach* is that there may be complementary information contained in multiple modalities [72]. Training a deep feature extractor then includes learning to integrate multimodal representations and weighting their benefits in a data-driven manner. Among the earliest representatives of this approach is SAR-Opt-cGAN [139], whose architecture resembles that of McGAN [135] but uses 10 spectral bands of Sentinel-2 combined with paired Sentinel-1 measurements for cloud removal. Likewise, the model of [140] is among the first which combine optical with radar sensors. While their philosophy of utilizing multi-sensory information greatly influences the contributions contained in this thesis, the manner in which many of these premier efforts pursued this objective is relatively crude and oftentimes limited to a simple, initial feature stacking followed by an early fusion stage. To make more principal modifications to adopted architectures originally tried and tested on conventional camera data, and to integrate multi-sensory information in a way tailored to satellite image reconstruction, is an open challenge still left for following contributions to address.

A crucial part of the above networks is not only their architectures and any adjustments undertaken to adapt them for satellite image processing, but also the data on which they have been trained and tested. Concretely, the availability of sizable amounts of domain-specific data became of increasing importance with the advent of deep neural

networks. However, at the time of the aforementioned initial contributions it was not yet feasible or common practice to make the employed training or testing data publicly accessible. This is unfortunate, as the lack of shared data makes comparisons across models challenging. Neither was it common for models to be evaluated on publicly accessible benchmarks, due to the mere lack of any. Furthermore, the employed data typically featured synthetic clouds, generated via Perlin noise [141], alpha blending [142] or by using Rayleigh simulation [143] which does subsequently not resemble the physical and spectral characteristics of real clouds. A notable exception to these limitations mark the RICE datasets of [144], which feature paired cloudy and cloud-free aerial as well as spaceborne observations, curated via Google Earth or from the Landsat-8 mission. However, all data curated thus far are focused on narrowly-defined regions of interest, leaving the challenge of general purpose and planet-wide cloud removal still unaddressed and for future work to resolve.

3.2.2 Multi-temporal Approaches

Among the premier contributions to propose a neural network for multi-temporal cloud removal are the works of [145, 146]. The approach of [145] is outstanding, as it does not only aim to remove clouds but also other sensor artifacts such as dead scan lines [67] in a unifying manner. For this sake, the authors propose a residual architecture, with a Siamese processing of satellite images, whose features get stacked and integrated into a reconstructed satellite image. Similarly, [146] propose a network for bi-temporal cloud removal, which reconstructs cloud-covered information in one satellite image by translating co-registered representations from an auxiliary cloud free optical image and adjusting them to match the receiving image's structure and spectrum. A major restriction of these works are their need of cloud-free supplementary optical data. Furthermore, the input time series is expected to be bi-temporal, with the image to be reconstructed as the first sample and the second image being the cloud-free source of information.

These limitations are addressed by the works of [147, 148]. The model of [147] is a convolutional architecture, composed of a temporal and a spatial subnetwork as well as a final fusion module. The authors of [148] propose a similar network, but complement it with a hand-crafted feature aggregation and an iterative refinement procedure. As with the preceding contributions, both approaches rely on accurate cloud masks from an external detector without discriminating between semi-transparent haze or dense cloud coverage. Moreover, the models are trained on only a few and narrowly defined areas, with quantitative evaluations solely conducted on simulated data, as is common practice in the literature thus far.

A final mention deserves the spatio-temporal generative model called STGAN [149]. STGAN ingests a temporal sequence of cloudy Sentinel-2 images to predict a cloud-free optical image. Other than many of its successors, it is both trained and evaluated on real and globally distributed data, which is kindly made publicly accessible for further research and benchmarking purposes. Among its limitations is that the provided data

solely contains RGB and near-infrared spectral bands, such that STGAN solely operates on a reduced spectrum and without any involvement of radar data. While this design choice has its precedents in the literature [135], it may limit the resulting model's reconstruction goodness and applicability for remote sensing downstream applications. A final shortcoming is due to STGAN's architecture, which stacks and fuses temporal representations in a pairwise manner, and thus becomes impermissibly costly for any but very short input sequences.

4 Summary of Contributions

This chapter summarizes the key contributions of this dissertation. The presented works all relate to the topic of cloud removal in satellite imagery and correspond to the appended peer-reviewed publications which constitute this cumulative thesis. At first, this dissertation analyzes the impacts of clouds and the benefits of cloud removal in practice. Second, the task of curating data for future training and benchmarking of methodology is addressed, which provides the basis for all further models considered herein. Third, contributions for the mono-temporal multi-sensor cloud removal setting are summarized. Fourth, the case of multi-temporal multi-sensor cloud removal is considered. Finally, the matter of uncertainty estimation is addressed to ensure reliable and trustworthy cloud removal. In combination, these contributions are meant to provide a comprehensive overview of the state-of-the-art in cloud removal and its further directions.

4.1 The Impact of Clouds on Earth Observation

Cloud coverage is an ubiquitous and persistent issue for a global and seamless optical monitoring of our planet. Clouds and their physical properties have been researched in depth [63], as well as the statistics of their spatio-temporal distribution across the surface of Earth [5]. Yet, little attention has thus far been spent on investigating their effects on contemporary machine learning approaches, deployed in the setting of common remote sensing tasks. Moreover, the majority of curated optical satellite datasets are explicitly cleaned from clouds and remote sensing models are subsequently (pre-) trained on (predominantly) clear-view data [150, 151, 152]. This common practice, however, is in contrast to the application of networks typically trained on non-cloudy datasets to data in the wild, which may be polluted by haze or clouds. To shed light on the matter of whether and how clouds affect remote sensing in practice is the aim of this section, which serves as a thematic opening to the summary of contributions of this thesis.

Peer-reviewed publications associated with this section

- J. Gawlikowski*, **P. Ebel***, M. Schmitt, and X. X. Zhu. Explaining the effects of clouds on remote sensing scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15:9976–9986, 2022.
  * *Authors contributed equally to this work.*

The contribution of [153] analyzes the effects of haze, clouds and cloud shadow on a common remote sensing application, full scene land cover classification. For this sake, the work combines the intersection of regions in the two datasets of SEN12MS [151] and SEN12MS-CR [154] to get paired and co-registered patches of cloudy and cloud-free Sentinel-2 observations, as well as associated patch-wise single-label land cover annotations [155]. The resulting data exhibits an extent of cloud coverage that coarsely coincides with statistics observed empirically [5]. Comparing the multi-spectral fingerprints of cloudy versus cloud-free data on a land cover-wise basis reveals a drastic shift in band statistics. To investigate the effects of this distribution shift on classification performances, five architectures [22, 156, 19] frequently employed for the task [155] are trained on cleaned-up and cloud-free data, as is a common practice in research. As is shown in Figure 4.1, the networks are fitted to cloud-free data and A further analysis of issues is provided by a subsequent Grad-CAM interpretation [157] of selected samples, which shows that outlier intensities or high contrast areas such as clouds or cloud shadows oftentimes redirect a network's focus and thereby drive the misclassifications.

At last, the study analyses classification accuracy and prediction confidence as a function of the percentage of cloud cover. The primary outcome is that performance drastically decreases as coverage increases. Notably, a presence of at most 10% of cloudy pixels (which may be hard to avoid in practice, as a consequence of imperfect cloud detection [117]) readily constitutes a salient drop in whole scene classification performance.

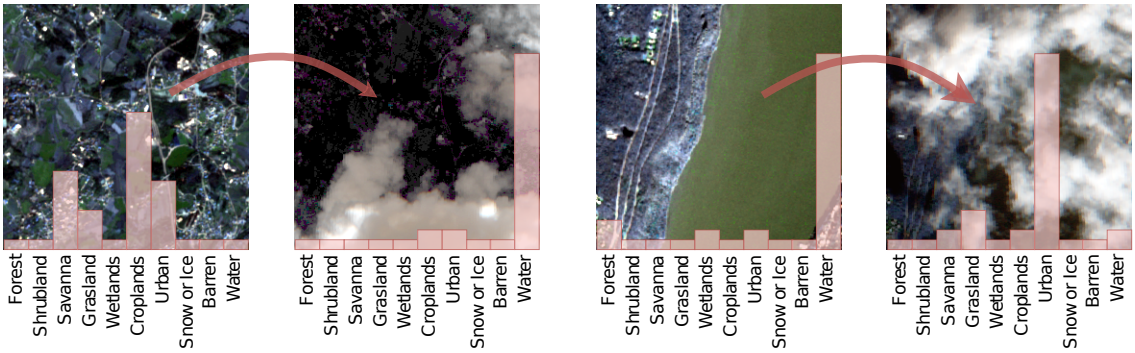


Figure 4.1: The effects of clouds on scene classification. The visualization shows two examples of clear images, cloudy images and the corresponding predicted class probabilities. In both cases the cloud-free image is classified correctly, but the cloudy version is misclassified. In the first example, much of the croplands are obscured by cloud shadow which causes the misclassification as water body at a high confidence. In the second example, the clouds cover a large range of the water but keep a part of a city visible such that the sample containing clouds is misclassified as Urban with a high conviction. The cloud coverage of the samples is 19% and 77%, respectively. Though parts of the images are still visible, the classifier's predictions are misguided by the clouds and the resulting shadows.

Furthermore, the gradual drop in performance is not reflected in terms of the network's confidence, as measured by a concentration of its logit values onto a single class. This dichotomy, at perpetually high confidences, indicates the network being unaware of the presence of any outlier pixels. At the same time, it shows the risk of deploying popular models on data in the wild, as the remote sensing practitioner may neither rely on their classifications nor on their confidence in the presence of any haze, clouds or cloud shadow left unfiltered. The outcomes of this experiment are visualized in Figure 4.2.

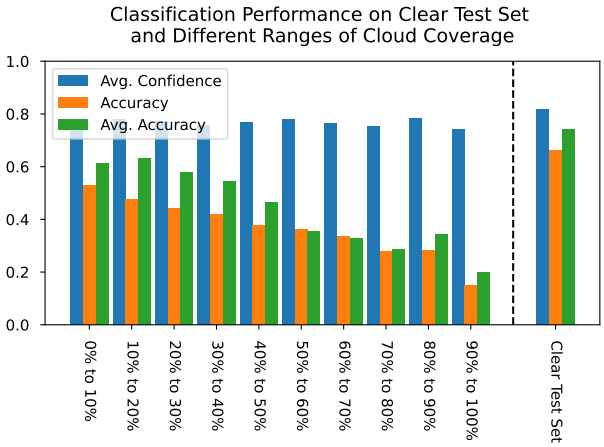


Figure 4.2: Confidently wrong. The performance of ResNet50 [22] as a function of varying ranges of cloud coverage. While accuracies detriment with increasing cloud coverage, the network's confidence remains consistently high.

Finally, while not constituting key contributions of this dissertation, the interested reader is furthermore referred to the related publications of [158, 159]. Both studies explore the practical benefits of cloud removal in the context of a subsequent downstream task, such as scene classification [158] or semantic segmentation [159].

4.2 Data & Benchmarks for Global and All-Season Cloud Removal

Early approaches to cloud removal were based on hand-crafted features inspired by the computer vision literature and classical signal processing algorithms. While these techniques may be valuable for case studies, they are specifically fit and hence geospatially constrained to particular regions of interest rather than being able to generalize to our entire planet. More recent contributions build on contemporary neural networks. Yet, early adaptations of the new paradigm oftentimes still retain the focus on selected regions of interest [145], more akin to case studies on selected areas in the geosciences. While special interests in narrowly defined areas are a common motif, a limiting factor is the lack of sufficiently large and representative datasets for cloud removal in remote sensing: Firstly, deep neural networks require large amounts of training data beyond what has been used in prior work. Besides, the curated data should be diverse enough to represent the variety of land cover on Earth and subsequently allow the trained model to generalize to any unseen places on our planet. Second, the collected data should be indicative of real conditions—in particular, clouds should be as encountered in the wild to

Table 4.1: Overview of datasets and benchmarks for cloud removal in optical aerial and satellite imagery that are publicly available. The two datasets of SEN12MS-CR and SEN12MS-CR-TS contained in this thesis (highlighted in bold) are the first in offering a global and multi-modal collection of curated optical satellite images for reconstructing cloud-covered pixels in multi-spectral remote sensing.

dataset	source	resolution	# ROI	# patches	patch size	spectral bands	SAR	time points
RICE-I [144]	Google Earth	< 15 m	1	500	512	3	✗	1
RICE-II [144]	Landsat-8	30 m	1	450	512	3	✗	1
cloudy City-OSM [162]	Google Maps	0.1 m	2	104	500	3	✗	4
STGAN [149]	Sentinel-2	10 m	945	3101	256	4	✗	3
SEN12MS-CR [154]	Sentinel-2	10 m	169	122,218	256	13	✓	1
WHUS2-CR [163]	Sentinel-2	10 m	36	17,182	256	13	✗	1
SEN12MS-CR-TS [73]	Sentinel-2	10 m	53	15,578	256	13	✓	30

capture their complex microphysical and electromagnetic properties. Finally, the aforementioned generalization capability of any candidate approach should be put to test by an equally comprehensive test split, to provide a faithful benchmark for general purpose cloud removal. These motives guide the collection of the following datasets.

Peer-reviewed publications associated with this section

- **P. Ebel**, A. Meraner, M. Schmitt, and X. X. Zhu. Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery. IEEE Transactions on Geoscience and Remote Sensing, 2020.
- **P. Ebel**, Y. Xu, M. Schmitt, and X. X. Zhu. SEN12MS-CR-TS: A remote-sensing data set for multimodal multi-temporal cloud removal. IEEE Transactions on Geoscience and Remote Sensing, 60:1–14, 2022.

In the spirit of the parent dataset SEN12MS [151], all satellite observations collected for this thesis are provided with their full multi-spectral information. That is, all datasets introduced in this dissertation consist of a collection of 13-bands level 1-C top-of-atmosphere reflectance Sentinel-2 products. The decision for level 1-C processing is made in order to leave atmospheric corrections to the cloud removal approach, rather than being taken care of by an external preprocessing pipeline. Furthermore, following earlier sensor fusion experiments [139, 160, 140] indicating the benefits of radar measurements, all datasets feature co-registered Sentinel-1 measurements to complement any Sentinel-2 images. Notably, all curated observations reflect real conditions and all paired cloudy images are natural, rather than being simulated as was conventional in many of the earlier works [135, 161, 162] reviewed in chapter 3.

An overview of publicly available cloud removal datasets is given in Table 4.1. Datasets are listed in chronological order, with the contributions of this dissertation marked in bold. Their completeness, diversity and scale place both SEN12MS-CR and SEN12MS-CR-TS at a unique position in the the satellite image reconstruction benchmarking ecosystem. The benchmarking tables reported in this section evaluate seminal cloud removal models based on the metrics defined in section 2.4. Scores are maintained and updated on `https://patrickTUM.github.io/cloud_removal/`.

Mono-temporal Cloud Removal. SEN12MS-CR, a multi-sensor dataset for mono-temporal cloud removal is proposed. SEN12MS-CR builds on the established SEN12MS dataset [151] for whole-planet land cover classification, which has previously laid the groundworks for the 2020 IEEE GRSS Data Fusion Contest for global land cover mapping [164]. As such, 169 out of the original 252 non-overlapping regions of interest are subsampled with equal distribution across all continents and meteorological seasons. The geospatial distribution of the samples areas is illustrated in Figure 4.3, with points of any color indicating a region of interest contained in SEN12MS-CR. For every region, a Sentinel-1 radar image, a co-registered cloudy as well as a paired cloud-free Sentinel-2 image are acquired within the same meteorological season to limit intermediate surface changes The resulting full-scene images have an average size of approximately 52 $\times$ 40 km^2 ground coverage, corresponding to complete-scene images of about 5200 $\times$ 4000 px^2. Each image is manually checked for any potential artifacts and subsequently translated from Partitioned into patches of size 256 $\times$ 256 px^2 with a spatial overlap of 50% between neighboring patches, yielding an average of over 700 patches per ROI. Each patch consists of a triplet of ortho-rectified, geo-referenced cloudy and cloud-free 13-band multi-spectral Sentinel-2 images Finally, each patch triples is automatically controlled for potential imaging artifacts and exclusively artifact-free patches are preserved to constitute the final cleaned-up version of SEN12MS-CR. Table 4.2 provides benchmark results on SEN12MS-CR for models at the time of writing this thesis.

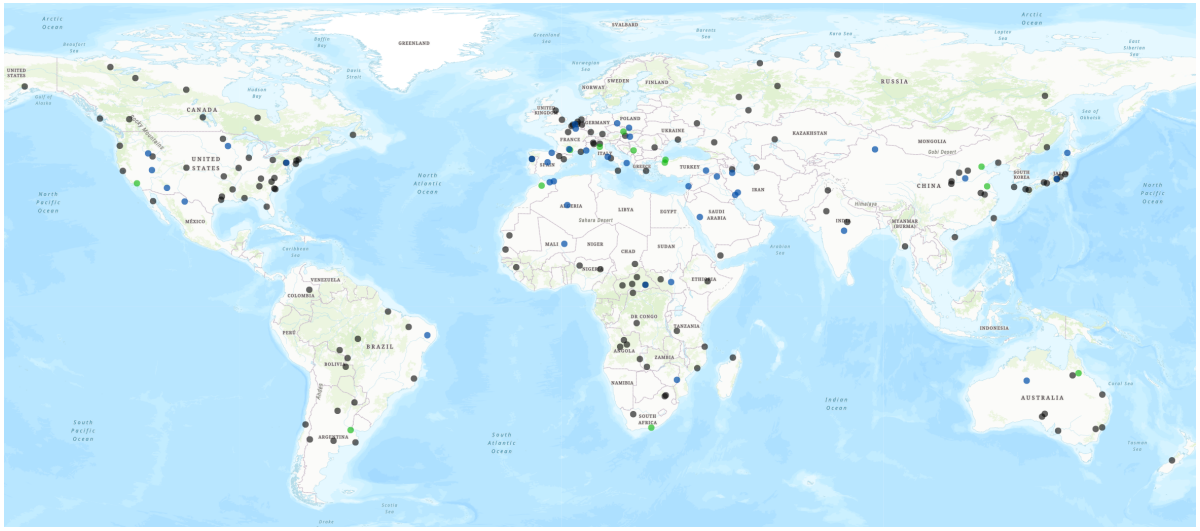


Figure 4.3: Map of data locations. SEN12MS-CR and SEN12MS-CR-TS are datasets of regions of interest samples over the whole globe and throughout all seasons.

One hallmark of SEN12MS-CR is it consisting of exclusively real world data, featuring clouds as they occur in the wild and representing all their natural characteristics as outlined in chapter 2. This novelty is in contrast to previous research, which solely conducted quantitative analysis on synthetic data—with clouds cropped out [111, 140], simulated through Perlin noise [135, 169, 161] or alpha-blended with secondary images

Table 4.2: Benchmarking on SEN12MS-CR, with data and splits as described in [154].

Method	↓ MAE	↑ PSNR	↑ SSIM	↓ SAM
McGAN [135]	0.048	25.14	0.744	15.676
SAR-Opt-cGAN [139]	0.043	25.59	0.764	15.494
SAR2OPT [6]	0.042	25.87	0.793	14.788
SpA GAN [165]	0.045	24.78	0.754	18.085
Simulation-Fusion GAN [166]	0.045	24.73	0.701	16.633
DSen2-CR [167]	0.031	27.76	0.874	9.472
GLF-CR [168]	*0.028*	*28.64*	**0.885**	*8.981*
UnCRtainTS$_{L2}$	**0.027**	**28.90**	*0.880*	**8.320**

Table 4.3: Comparing training on real versus synthesized clouds and the capability of networks trained in these manners to generalize to real world data, featuring clouds as they occur in the wild. The results show that testing on synthetic data strongly overestimates performances when compared to evaluating on real data. Furthermore, networks trained on simulated data perform worse on real data than networks trained on real data do. This indicates a gap in the realism of established simulations of clouds, missing parts of their properties outlined in chapter 2.

training data	test performance ↑ precision		↑ recall		↑ F1	
	synthetic	real	synthetic	real	synthetic	real
Perlin	0.239	0.168	0.800	**0.592**	0.368	0.262
copy	**0.692**	*0.458*	**0.856**	*0.586*	**0.766**	*0.514*
real	—	**0.564**	—	0.551	—	**0.557**

[110, 162, 170]. To investigate the merits of training and benchmarking on real data, the following experiment on SEN12MS-CR from the accompanying publication of [154] is outlined: A given single-image cloud removal neural network is trained trice: on data of real clouds, on samples of Perlin-simulated cloud noise and finally by overlaying plus alpha-blending clouds from another image. After training, the tree networks are tested on their respective simulated test data (if any) as well as on real data. This serves to check for any differences in performances. Performances for this experiment are reported in terms of the measures of [171], as utilized in [154]. The outcomes of the experiment show that there is a considerable domain gap between simulated and real cloudy data. Moving from simulated to real data deteriorates performances. Furthermore, evaluating on simulated data may grossly overestimates reconstruction performances, as compared to benchmarking on real data. This implies that any reconstruction goodness quantified on simulations is a poor indicator of the actual performances on real data. In sum, training on real data yields better performances on real data, and testing on real data avoids any kinds of misestimate of the model's actual goodness.

Finally, one concern to be addressed is that of potential intermediate land cover change between the acquisition dates of the cloudy input and the paired cloud-free target data. This deserves attention, as any pronounced discrepancy may pose a challenge for evaluating cloud-removed predictions under controlled conditions. To analyse potential surface area changes, Figure 4.4 analyzes the empirical distributions of pixel-value differences between paired cloud-free (in blue) and cloudy (in green) input pixels to their cloud-free target pixel counterparts. The residual mismatch is calculated by subtracting input pixel values minus the co-registered pixel value in the co-registered target patch, such that brighter input pixels obtain a positive residue. The respective histograms are parameterized by independently fitting two asymmetric Laplace distributions [172, 173]. The mode of the cloud-free distribution, i.e. the most probable value, is at 0.006 with the first and third quartiles at -0.09 and 0.01. That is, there is a tight correspondence between cloud-free input and output pixels. The mode and the mean of the cloudy fit are at 0.028 and 0.14, respectively. The 0.25 and 0.75 percentiles are located at -0.04 and 0.2, indicating a pronounced skew towards higher values. This matches the intuition that bright, cloud-covered input pixels have higher reflectance values compared to their paired pixels in the cloud-free target patch. To conclude, the analysis confirms the paradigm of SEN12MS-CR to benchmark on real data by controlling for intermediate changes by collecting paired data sufficiently close in time.

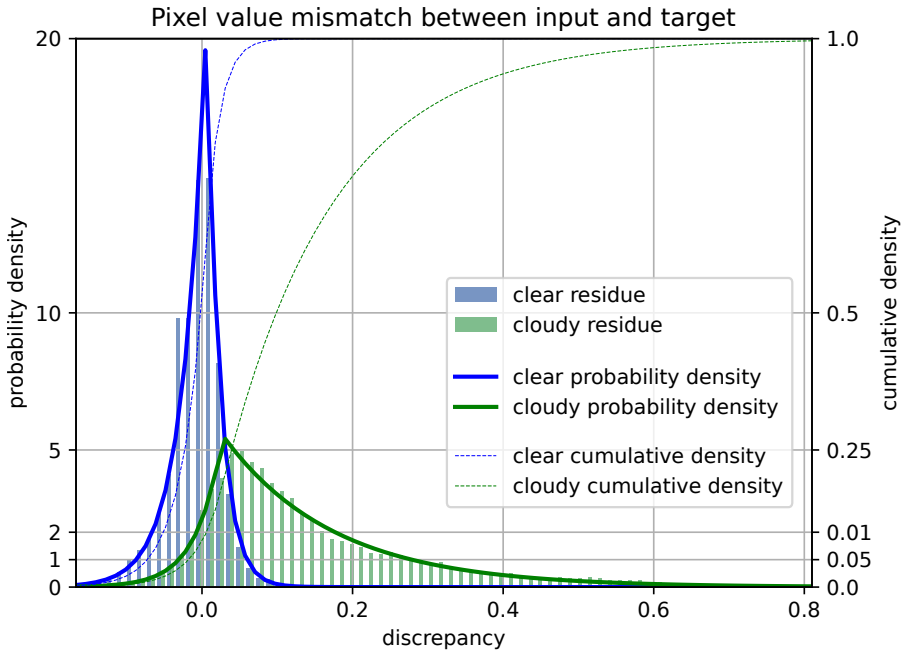


Figure 4.4: Discrepancy of pixel values between paired input and target Sentinel-2 data in SEN12MS-CR. Blue visualizations display the mismatch for cloud-free input pixels to their target image counterparts, while green visualizations pertain to the mismatch of cloudy-pixels to the respective cloud-free target pixels. The first distribution peaks at close by zero, indicating a strong correspondence between input and target pixels. The second distribution is clearly shifted off-zero, with a skew towards higher values, reflecting the typical brightness of cloudy pixels.

Multi-temporal Cloud Removal. To promote multi-sensor time series cloud removal, the dataset and benchmark of SEN12MS-CR-TS is curated. The dataset consists of 53 large scale regions of interest, which are globally distributed and cover a total surface of over $80,000\ km^2$. The selected regions are a subsample of the areas in the SEN12MS-CR precursor, such that both datasets are compatible and complementary to one another. Figure 4.3 shows the distribution of collected areas, with gray points being exclusive to SEN12MS-CR, blue points belonging to the train splits of both datasets and green pins representing test split areas of either benchmark. For each region, 30 co-registered and paired S1 and S2 full-scene images are collected, evenly spaced in time throughout the year of 2018. All full-scene images are preprocessed and sliced into $256 \times 256 px^2$ patches, and quality-controlled as for SEN12MS-CR. Figure 4.5 depicts exemplary input data, composed of paired Sentinel-1 and Sentinel-22 samples as well as cloud masks predicted via s2cloudless [174]. Finally, Table 4.4 provides benchmark outcomes on SEN12MS-CR-TS for cloud removal models at the time of writing this thesis.

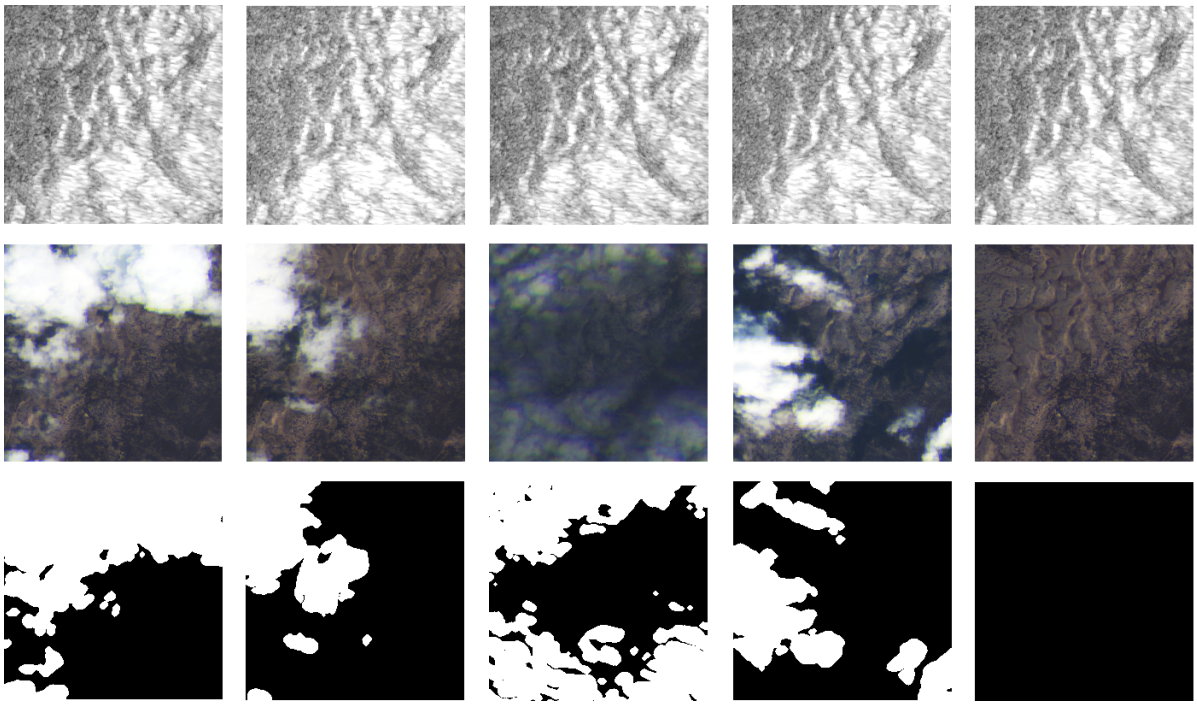


Figure 4.5: Example SEN12MS-CR-TS data. Rows: S1 data, S2 data and binary cloud masks.Columns: Samples of five different time points, four for input and one as target. The illustrations show that the observed region is affected by variable atmospheric disturbances and covered by a dynamic extent of clouds, changing over time. The detected cloud coverage of the individual input samples is 49, 23, 48 and 26 percent, and the target sample is cloud-free. While some pixels are clear at least at one point in the input sequence and may thus be inferred by integrating across time, others are cloud-covered throughout the sequence and require spatial context or cloud-robust sensor information to be reconstructed.

Table 4.4: Benchmarking on SEN12MS-CR-TS. All models are evaluated on time series of $T = 3$ inputs. For further details, please see the respective publications and [73].

Model	↓ RMSE	↑ PSNR	↑ SSIM	↓ SAM
least cloudy	0.079	—	0.815	12.204
mosaicing	0.062	**31.68**	0.811	14.324
DSen2-CR [167]	0.060	26.04	0.810	12.147
STGAN [149]	0.057	25.42	0.818	12.548
CR-TS Net [73]	*0.051*	26.68	0.836	10.657
U-TAE [40]	*0.051*	27.05	0.849	11.649
UnCRtainTS$_{L2}$	**0.049**	27.23	*0.859*	*10.168*
UnCRtainTS$_{\mathcal{N}}$	*0.051*	*27.84*	**0.866**	**10.160**

4.3 Deep Learning Methods for Mono-Temporal Multi-Sensor Cloud Removal

The scenario of translating a single cloud-covered image to a reconstructed version thereof is a data-efficient approach to the cloud removal task. However, this version of the task is nonetheless challenging, as spatio-spectral correlations with clear neighboring or distant pixels may only inform about cloud-covered information to a limited extent. This raises the question of how to facilitate inpainting of obscured pixels. Potential facilitation provides the inclusion of paired SAR measurements, which are robust to detrimental weather phenomena [78] and can thus provide valuable guidance even in the case of dense and thick cloud coverage [139, 140]. Consequently, all models associated with key publications introduced in this section follow a multimodal paradigm of integrating optical information with complementary and paired radar data. This scenario is depicted in Figure 4.6, with exemplary predictions by the network of [167]. This and any other benchmarked mono-temporal cloud removal approaches are trained and tested on SEN12MS-CR, with their scores reported in Table 4.2. What follows is a summary of the three mono-temporal cloud removal models associated with this section.

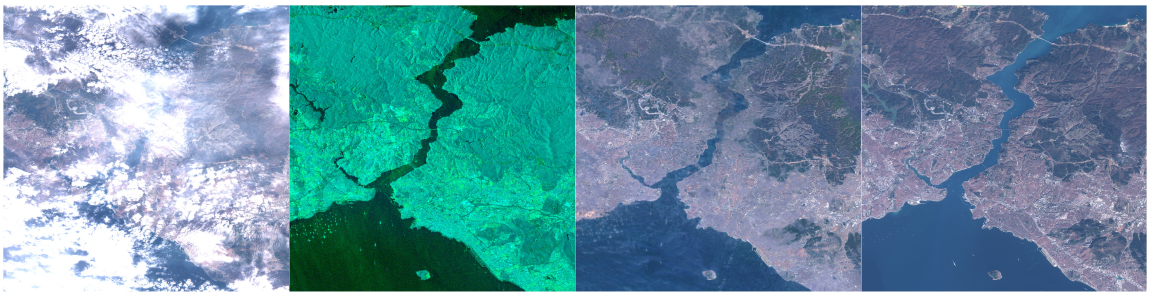


Figure 4.6: Exemplary full scene prediction. Depicted are a cloudy Sentinel-2 image, Sentinel-1 measurements, the cloud-removed prediction by DSen2-CR and a cloud-free target image of Istanbul. Despite thick cloud coverage, DSen2-CR succeeds in inferring the test scene's structure thanks to the auxiliary radar data.

Peer-reviewed publications associated with this section

- A. Meraner, **P. Ebel**, X. X. Zhu, and M. Schmitt. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. ISPRS Journal of Photogrammetry and Remote Sensing, 166:333–346, 2020
- **P. Ebel**, A. Meraner, M. Schmitt, and X. X. Zhu. Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery. IEEE Transactions on Geoscience and Remote Sensing, 2020.
- F. Xu, Y. Shi, **P. Ebel**, L. Yu, G.-S. Xia, W. Yang, and X. X. Zhu. GLF-CR: SAR-enhanced cloud removal with global–local fusion. ISPRS Journal of Photogrammetry and Remote Sensing, 192:268–278, 2022

Residual Architectures. Inspired by prior work adapting a residual backbone [22] for super-resolution [26] or hyperspectral image denoising [27] in remote sensing, DSen2-CR is proposed in [167]. DSen2-CR is noteworthy in being among the first cloud removal models to use and predict the complete spectrum of Sentinel-2 observations, as well as combining it with paired Sentinel-1 data. The network consists of stacked residual blocks and a long residual connection directly forwarding the input multi-spectral Sentinel-2 data to the output, as depicted in Figure 4.7. While prior image reconstruction networks are conventionally trained with a pixelwise L1 or L2 loss as in equations 2.12 or 2.4, a novelty of this work is the Cloud-Adaptive Regularized Loss (CARL), which teaches the reconstruction of obscured pixels while explicitly encouraging the preservation of clear areas. CARL consists of a cloud-adaptive part implementing the aforementioned distinction, as well as a target regularization part for smooth predictions with a tradeoff controlled by hyperparameter λ. It is defined as

$$\mathcal{L}_{\text{CARL}} = \overbrace{\frac{\| \mathbf{M} \odot (P - T) \; + \; (1 - \mathbf{M}) \odot (P - I) \|_1}{N_{\text{tot}}}}^{\text{cloud-adaptive part}} + \lambda \overbrace{\frac{\|P - T\|_1}{N_{\text{tot}}}}^{\substack{\text{target reg.} \\ \text{part}}}$$

where T is the cloud-free target image, P denotes the cloud-removed predicted image, I refers to the cloudy input image M pertains to a binary cloud mask as predicted via a separate cloud mask detector [175, 176] and N_{tot} is the total pixel count. Following earlier studies that an L1 loss yields sharper satellite image reconstructions than an L2 cost function [177, 26], distances within CARL are described in terms of the L1 norm. As shown by [167], training via CARL outperforms conventional pixelwise losses. Furthermore, the inclusion of Sentinel-1 radar measurements complements the multi-spectral optical inputs and yields further gains in performance. An exemplary full-scene reconstruction is illustrated in Figure 4.6. Note how the overall scene is successfully reconstructed despite dense occlusions, and the cloud-covered coastlines are clearly outlined in the radar image as well as in the subsequent cloud-removed prediction.

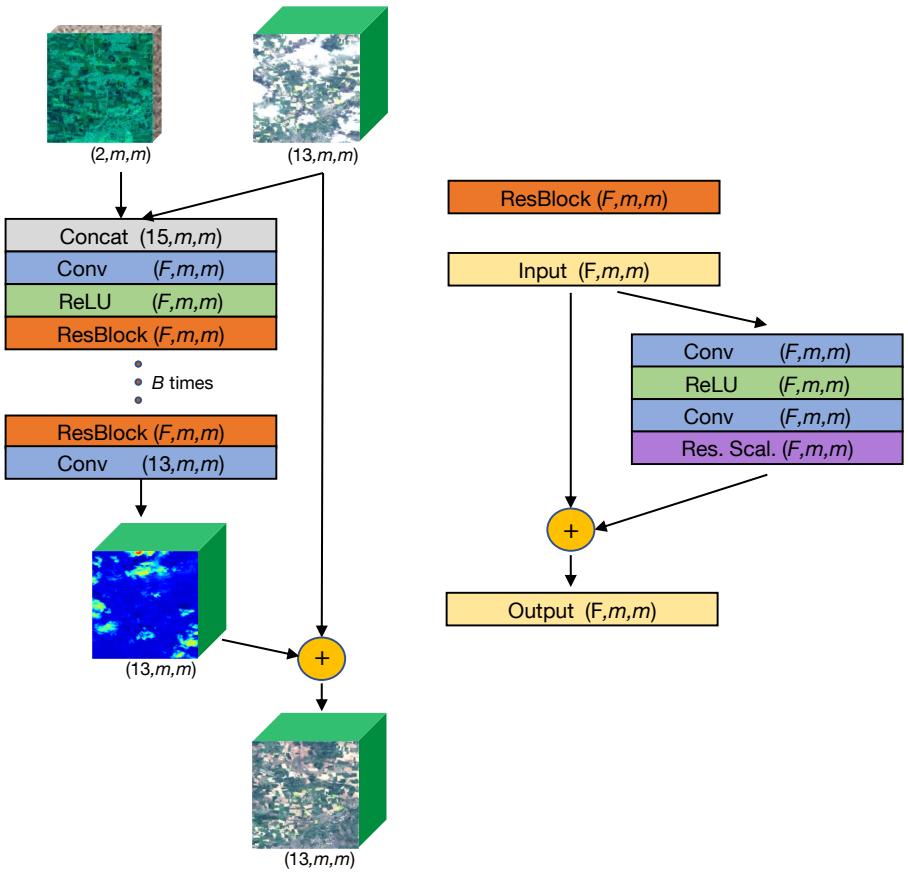


Figure 4.7: DSen2CR architecture. The network maps a concatenation of co-registered radar and cloudy optical satellite images to a cloud-free optical satellite image. The architecture of DSen2CR consists of a sequence of B stacked residual blocks, as introduced in section 2. This sequence predicts a residual modification, applied to a long skip connection passing the optical input directly to the output. In combination with the proposed CARL loss, this encourages the faithful preservation of cloud-free pixels, while only obscured image parts need to be reconstructed.

Generative Adversarial Networks. In the spirit of preceding cloud removal models [135, 139, 161], the network of [154] follows a generative adversarial approach to learn sampling from a distribution of cloud-free images. Distinct from most prior efforts, its generator is multi-modal by being conditioned on co-registered radar plus cloudy optical samples to map to cloud-free optical images. To better internalize the relation between the two sensors, the proposed model adopts the cycle-consistent setup of [32] introduced in chapter 2, which encourages learning a bi-directional mapping between both modalities. For this sake, the complete setup is composed of a generator and a discriminator networks per each of the two modalities. That is, a second generator complements the first by mapping from the cloud-free optical domain back into the radar modality, and one discriminator classifies real versus generated images per domain.

Different from any prior generative work [135, 139], the proposed network replaces the conventional U-Net backbone [21] by a generator designed specifically for optical satellite image reconstruction. Following recent residual architectures [26, 136, 27, 167], the generator features a long-skip connection bypassing the encoder-decoder processing to directly forward cloudy inputs to the output. Thereby, only residual modifications need to be learned, which an auxiliary cloud map regression task encourages to be sparse and limited to cloudy pixels while cloud-free input information is preserved. Any residual changes are processed via the encoder-decoder structure, connected by a deep bottleneck composed of residual blocks. The generator mapping paired Sentinel-1 and cloudy Sentinel-2 data to cloud-free Sentinel-2 observations is shown in Figure 4.8.

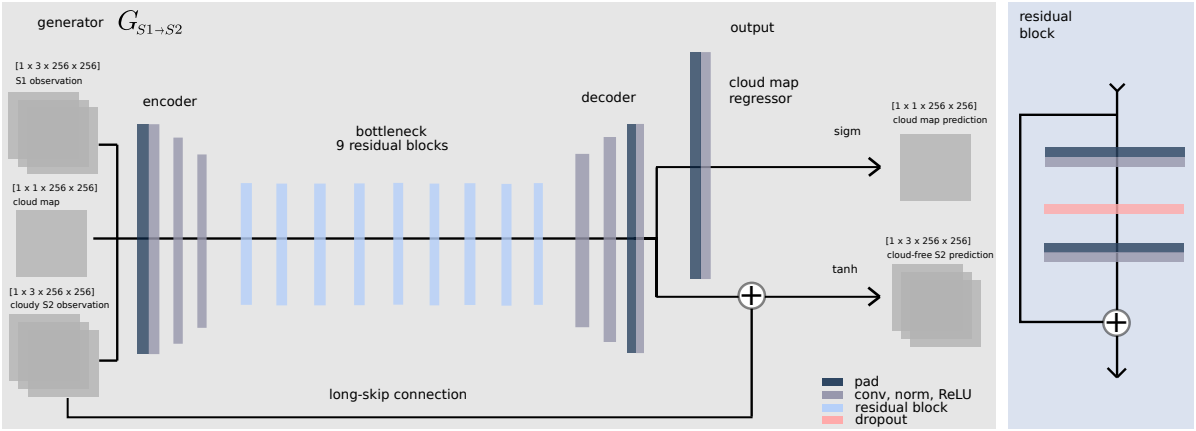


Figure 4.8: GAN architecture. Generator of the Cycle-GAN architecture in [154]. The network maps cloudy optical samples, their associated cloud maps and co-registered radar data to cloud-free optical data. The network follows an encoder-decoder structure with a deep residual bottleneck part. A long-skip connection bypasses these components and directly forwards the optical input data towards the output layer, such that only residual modifications need to be learned. To encourage sparse modifications, a shallow cloud map regression module is applied on the residual features—enforcing sparsity, such that cloud-free pixels are left mostly untouched while solely obscured pixels are modified.

The cycle-consistent cost function supervising the generators allows for pixel-level supervision even when there exists no pixel-level correspondence between domains, as the similarity between the original input and its reconstruction in the same modality is evaluated. In the work of [154], input optical images may be cloudy yet generated optical images are cloud-free, so the cycle-consistent loss is guided via a cloud mask and only evaluated over clear pixels to keep the principle intact. This encourages a faithful preservation of cloud-free pixels comparable to the CARL loss of [167] while the adversarial loss teaches removing the spectral statistics of cloudy data, with neither relying on pixel correspondence between cloud-covered inputs and clear target samples. The hypothetical infeasibility of controlling for this pixel correspondence and the possibility of intermediate land cover changes was a common criticism of benchmarking on real data, serving as an argument for rather simulating clouds instead [111, 110, 135, 169, 161, 140, 162, 170]. As the SEN12MS-CR dataset accompanying the work of [154] provides such paired cloudy and cloud-free observations, whether its close acquisition times rebut such criticism can be put to test: By interpolating from unpaired training as described above to paired supervision via a conventional L1 loss, the proposed generative network can make use of variable amounts of paired input-output data. The outcomes of this experiment show that pixel-based supervision on paired cloudy and clear data yields better image reconstruction performances compared to an unpaired approach not requiring pixel-based correspondences. This further evidences that training on paired real data of clouds in the wild is beneficial and that models can make use of the extra supervision enabled by SEN12MS-CR in practice, supporting the dataset's underlying philosophy.

Finally, the work of [154] introduced several other mentionworthy concepts into the cloud removal literature. For instance, it proposes perceptual losses for better internalizing features and style [20] of the target distribution. Moreover, real-valued cloud maps instead of binary segmentations are utilized, acknowledging the continuum of permeability in (semi-transparent) haze or clouds. Finally, as GAN can be notoriously challenging to train this work employed recent techniques for more stable a optimization, such as a more informative adversarial loss [37] plus spectral normalization [178] of the discriminator, and demonstrated their effectiveness in the mono-temporal cloud removal setting.

Visual Transformers. At last, GLF-CR [168] introduces the visual transformer paradigm to cloud removal and combines it with a global-local fusion (GLF) approach to augmenting optical with radar representations as well as the other way around. More specifically, the model builds on shifted-window transformers [41, 179], that partition input patches into smaller windows to compute visual attention in, followed by a tile shift operation that induces global interactions. Following the narrative of the two preceding mono-temporal cloud removal models, GLF-CR aims to further strengthen the learned relation between features extracted from both radar and optical sensors.

For this sake, the architecture of GLF-CR consists of two branches initially encoding sensor-specific representations separately, followed by a global and local merging of the modalities at multiple hierarchical levels.

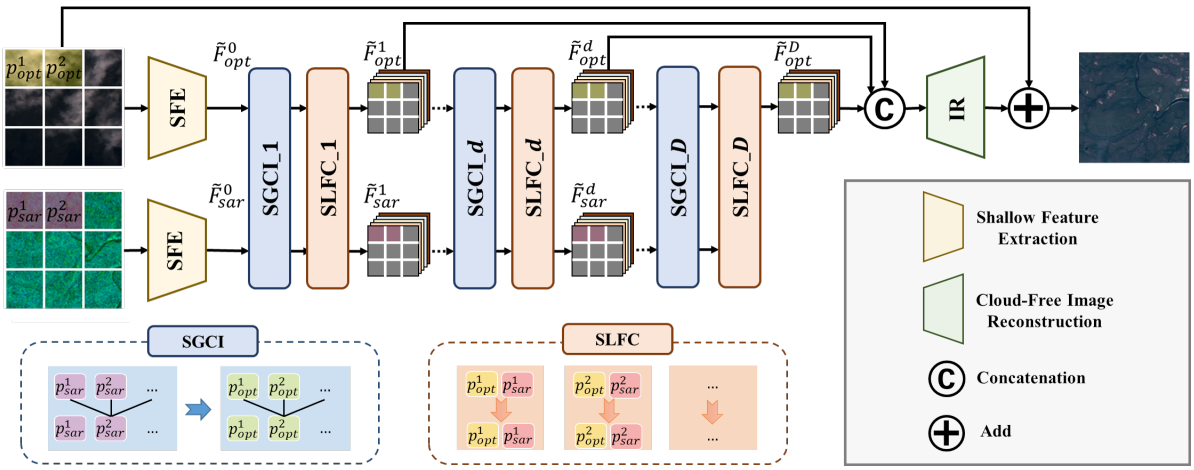


Figure 4.9: GLF-CR architecture. GLF-CR is a two-stream network, hierarchically merging information from radar and optical modalities to reconstruct cloudy pixels. For global-local fusion, features are processed by stacks of SGCI and SLFC modules mediating between both modalities. Finally, the hierarchical representations are agglomerated and integrated by a decoder for residual image reconstruction.

For its global-local fusion approach, GLF-CR introduces two new components: First, a SAR-guided global context interaction (SGCI) module and second, a SAR-based local feature compensation (SLFC) module. The idea of the SGCI block is that cloud-free and cloudy land cover can not be correlated straightforwardly due to the coverage of the latter, but spatial attention masks between their respective radar views can be used as a proxy. This way, radar data refines visual attention in the optical domain and thus guide the borrowing of features from distant clear pixels. The subsequently applied SLFC module first computes a dynamic filter on SAR features for despeckling purposes and then modulates the representations of optical data via their corresponding radar features and vice versa. Following this dual propagation mechanism, the refined features are then passed to a downsampling operator and the global-local processing is repeated at the next deeper layer. In sum, the refined features are processed by a sequence of global-local fusion stacks, whose multi-hierarchical representations are finally integrated by a residual image reconstruction decoder. A conceptual overview of the described architecture is provided in Figure 4.9.

Exemplary predictions of a mosaiced full-scene image are shown in Figure 4.10. Despite intense cloud coverage, GLF-CR can reconstruct the scene of interest at a high fidelity. Remarkably, minor details such as a dam in a lake (highlighted by an orange box) are much clearer reconstructed with the aid of the proposed SAR guidance mechanism, as compared to the single-sensor approach.

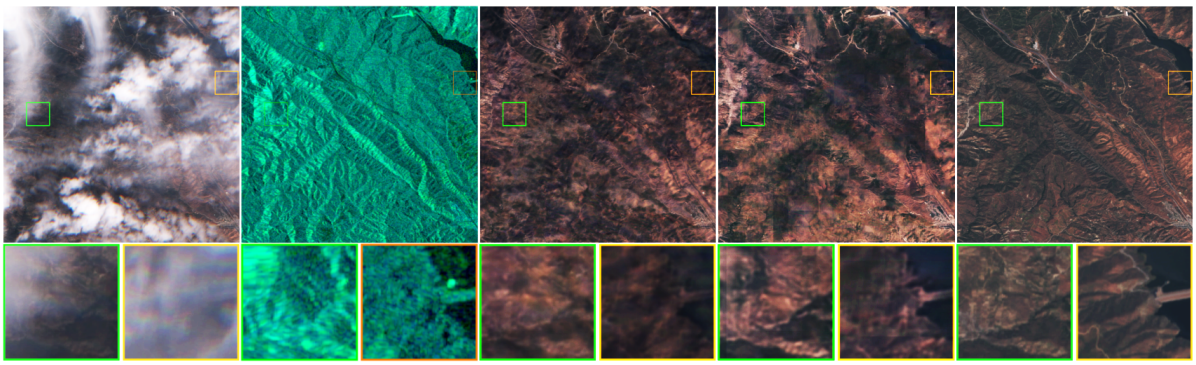


Figure 4.10: Exemplary full scene prediction. The figures show a cloudy Sentinel-2 scene, together with paired Sentinel-1 measurements, the cloud-removed prediction by GLF-CR (with and without using SAR) and a cloud-free target image of the scene. Note the intense cloud coverage, but GLF-CR still being able to faithfully reconstruct the scene. Remarkably, reconstructions of small details are considerably cleaner when using SAR versus without.

4.4 Deep Learning Methods for Multi-Temporal Multi-Sensor Cloud Removal

While mono-temporal cloud removal can use spatio-spectral correlations or radar data as a leverage to reconstruct otherwise entirely obscured information, the restriction to a single time point also brings its limitations. Specifically, with respect to multi-sensor fusion, Sentinel-1 quantifies the physical arrangement of land surface while Sentinel-2 provides insights into its molecular material compounds. As such, the sensors provide related yet complementary measurements [72]. Hence, not all properties of a multi-spectral image can be derived solely from a co-registered radar view. To incorporate further information, one may thus consider historical data. One of the earliest approaches to cloud removal is temporal mosaicing—as outlined in section 3, its underlying idea is to integrate across repeated measures and gather a cloud-removed collage of the optical images. Analogously, contemporary deep learning solutions may harness a time series of observations to arrive at better reconstructions of the regions of interest. This may serve to include information that may otherwise not be attainable from a single noisy optical view or paired radar measurements. In this sense, multi-temporal multi-sensor cloud removal may be considered as a generalization of the mono-temporal setting.

Peer-reviewed publications associated with this section

- **P. Ebel**, Y. Xu, M. Schmitt, and X. X. Zhu. SEN12MS-CR-TS: A remote-sensing data set for multimodal multi-temporal cloud removal. IEEE Transactions on Geoscience and Remote Sensing, 60:1–14, 2022.

When a time series of cloudy optical satellite images is provided, this raises the question whether the main interest is in the agglomerated information contained in the sequence, or whether the temporal dynamics of the time series should be preserved. The first aim is addressed in the sequence-to-point setting of the cloud removal task, where the desired output is a single cloud-free image prediction. The latter is covered by the sequence-to-sequence scenario, where an output sequence of the same temporal length as the input time series is expected. This section covers both the sequence-to-point and the sequence-to-sequence task for cloud removal, as introduced in the context of [73].

Sequence-to-point Cloud Removal. Addressing the challenge of multi-temporal multisensor cloud removal, the architecture of CR-TS Net initially processes each input satellite image independently and equally in parallel via weight-shared branches before merging representations across time points. CR-TS net draws inspiration from the Siamese structure of STGAN [149] as well as prior residual architectures for satellite image reconstruction [26, 167] preserving the original spatial resolution throughout all layers. Moreover, it replaces the costly cross-wise feature combination for every pair of time points [149] with more efficient 3D convolutions across both spatial and the

temporal dimension. The architecture of CR-TS Net is shown in Figure 4.11. Its more efficient design permits CR-TS Net to scale to longer input time series than was previously computationally affordable, while preserving feature maps at a high resolution for more detailed reconstructions. The network furthermore processes satellite images at their complete spectrum, and complements the optical information with radar data for multi-sensor fusion. The test scores in Table 4.4 show that CR-TS Net outperforms all preceding benchmarked approaches to multi-temporal cloud removal.

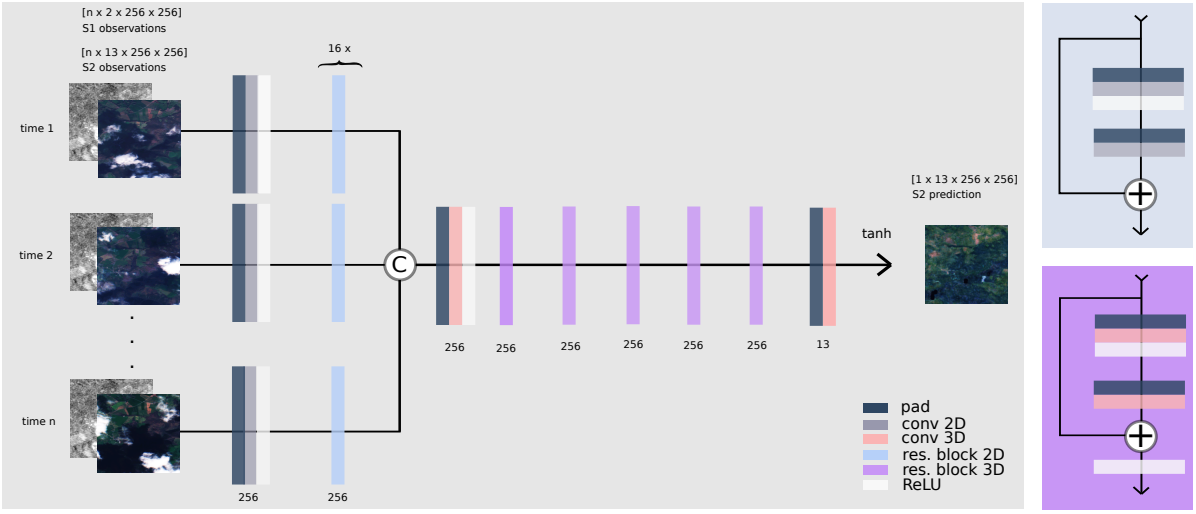


Figure 4.11: CR-TS Net architecture, designed for sequence-to-point cloud removal. The network is based on the architecture of [149] and consists of n siamese residual branches [167] doing single time point cloud removal on n individual time points. Subsequently, the feature maps are stacked in the temporal dimension and 3D convolutions are applied to integrate information across time. The output of the network is a single cloud-free image prediction.

Sequence-to-sequence Cloud Removal. Furthermore, the work of [73] considers the challenge of cloud removal over long sequences while preserving temporal resolution. Following the internal learning paradigm of [180, 18] and later successors [181, 182], a convolutional neural network is fitted directly onto the target data of interest to learn representing their signal. In this view, the observed clear pixels at any spatio-temporal coordinate are signal to be internalized, providing supervision to a tabula rasa neural network. Pixels shrouded by clouds or haze get reconstructed thanks to the inductive bias hardwired into convolutional neural networks [180]. That is, rather than generalizing across regions of interest, generalization within the internal learning framework amounts to translating information from clear pixels towards cloud-covered ones within a single region. In this sense, the sequence-to-sequence model of [73] is notably close to the classical approaches of section 3.1.

The sequence-to-sequence model of [73] is based on a 3D convolution U-Net [21], as previously considered for RGB video foreground inpainting in [18]. Exemplary qualita-

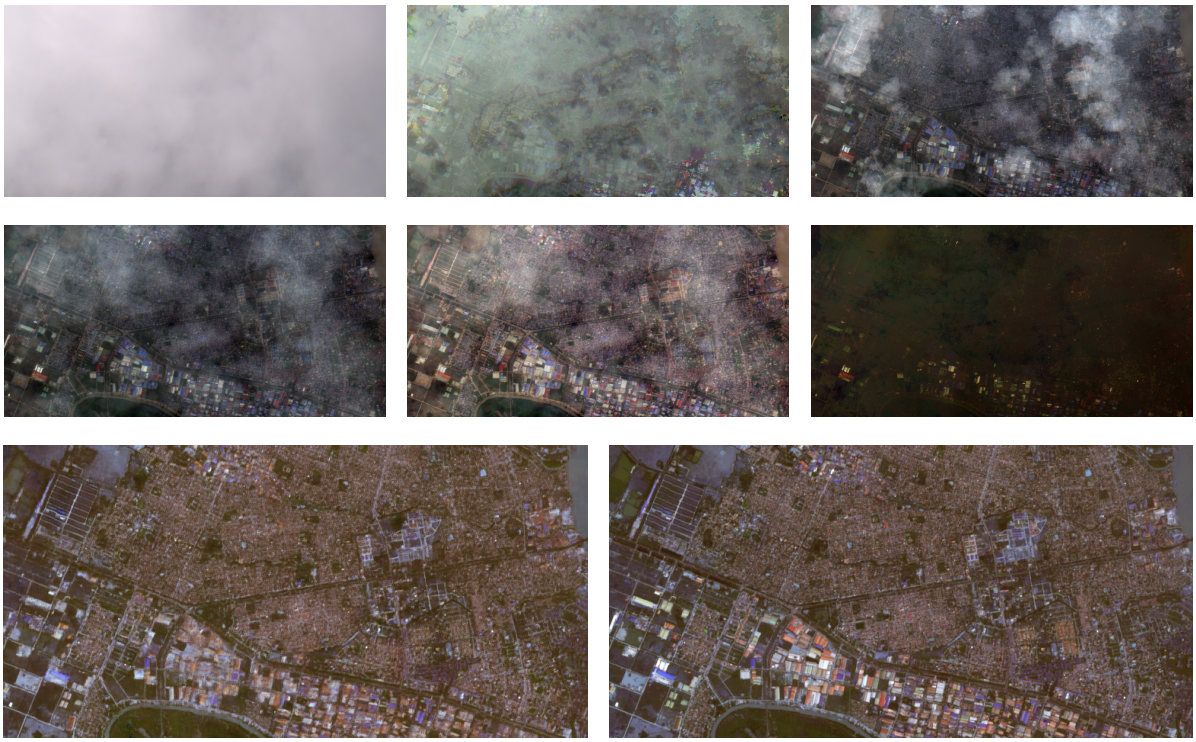


Figure 4.12: Exemplary sequence-to-sequence cloud removal results of the proposed internal learning approach, compared to baseline methods. The presented results show a cloudy image of the input sequence, the outcomes obtained via the matrix factorization or decomposition techniques of [120, 122, 121, 123, 124] as well as the prediction of the proposed model and finally the cloud-free image to be predicted. The results indicate that the presence of large and dense clouds poses a challenge for conventional methods. In comparison, the internal learning model biased via radar data achieves a close reconstruction of the cloud-free image.

tive results of the proposed approach, compared to classical signal processing baselines, are presented in Figure 4.12. Notably, the proposed model yields considerably better reconstructions than any of the baselines and provides valuable reconstruction even in the presence of intense cloud coverage. Further ablations in [73] show that conditioning the model on Sentinel-1 data rather than random noise to drive predictions provides a valuable prior to learn better scene representations, making it not entirely dissimilar to earlier SAR2OPT approaches as in [6, 137, 138].

4.5 Deep Learning Methods for Trustworthy Cloud Removal

While any sufficiently large and heterogeneous benchmarking dataset may provide overall indicators of a method's image reconstruction goodness, grand average performances provide little insight into a model's sample-by-sample trustworthiness. This is specifically problematic for risk-sensitive domains such as remote sensing, which relies on accurate observations and precise measurements. To resolve this shortcoming of current approaches, a novel model called UnCRtainTS is introduced []. UnCRtainTS is a neural network for multi-temporal and multi-sensor cloud removal, trained to predict reconstructed images and associated uncertainty maps implying potential errors alike.

Peer-reviewed publications associated with this section

- **P. Ebel**, V. Garnot, M. Schmitt, J. Wegner and X. X. Zhu. UnCRtainTS: Uncertainty Quantification for Cloud Removal in Optical Satellite Time Series. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2023.

With respect to its architecture, UnCRtainTS draws inspirations from residual networks for mono-temporal satellite image reconstruction [26, 27, 167] as well as recent attention-based approaches to vegetation monitoring [40]. Pertaining to the first, UnCRtainTS consists of a resolution preserving main branch of residual blocks to better conserve or reconstruct high-frequency spatial details. This kind of backbone is particularly beneficial when training with conventional pixelwise losses rather than a cost function dedicated for image reconstruction. With respect to the latter, UnCRtainTS makes efficient use of temporal pixel-wise attention [183] to integrate information across time points. In total, the network's architecture is composed of three main components: an encoder part that parallely processes input time points, a temporal attention and aggregation part which efficiently integrates information across time points, and finally a decoder that provides further spatio-spectral processing. The architecture is depicted in Figure 4.13. Tables 4.2 and 4.4 highlight that the backbone of UnCRtainTS performs competitive in mono-temporal as well as multi-temporal reconstruction settings alike.

To predict reconstructed images in combination with uncertainty maps, an NLL loss as introduces in equation 2.6 is utilized. This allows learning to predict the parameters of a distribution that are most likely to explain the sampled data. Due to its simplicity and its generality, it is assumed that the observations follows a Normal distribution, centered around the cloud-free target data. In this setting, the spread is indicative of the associated uncertainty. Following the multi-spectral nature of the to-be-reconstruced satellite image, a multivariate Normal distribution over the 13 bands of Sentinel-2 samples is assumed. Subsequently, the resoluting cost function with which UnCRtainTS is optimized corresponds to the Gaussian NLL loss of equation 2.9.

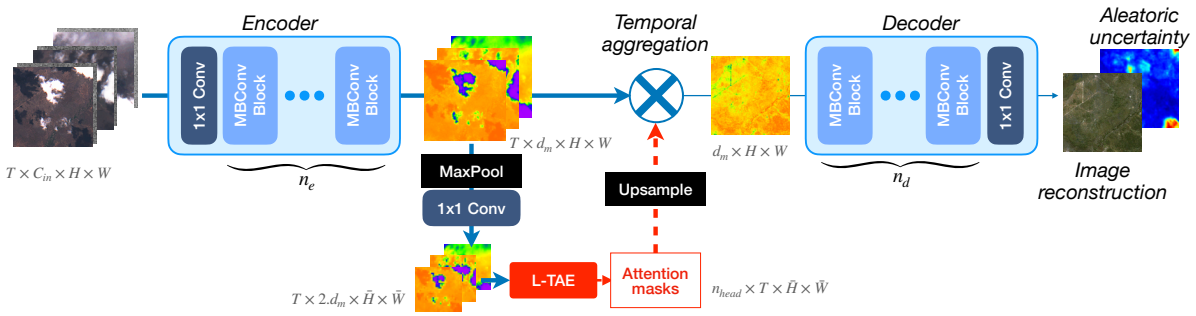


Figure 4.13: UnCRtainTS. The network consists of three main parts, applied along a main branch of MBConv blocks [24] that is processing feature maps at full input resolution: First, an *encoder* is applied in parallel to the T time points. Then, an *attention-based temporal aggregator* computes attention mask by applying an L-TAE to downsampled feature maps, used to aggregate the sequence of observations. Finally, the temporally integrated feature map is processed by a *decoding block*, yielding the image reconstruction and aleatoric uncertainty.

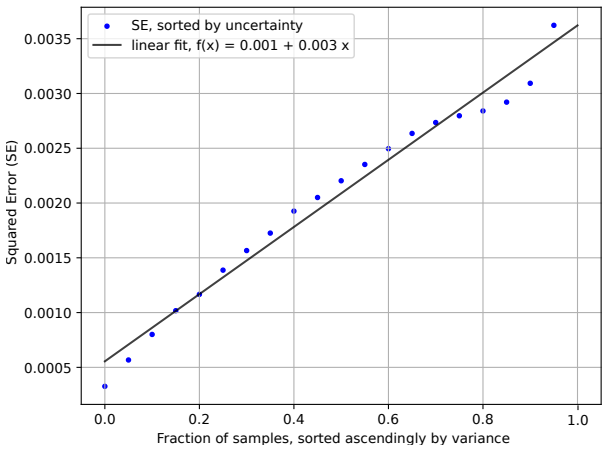
Finally, it is demonstrated that UnCRtainTS learns well-calibrated pixelwise uncertainty predictions and how these may be employed in practice: The predicted variances should be indicative of the model's empirical error. This objective is formalized in terms of the Uncertainty Calibration Error (UCE) [184]

$$UCE(e,u) = \sum_{p=1}^{P} \frac{N_p}{N} |e(B_p) - u(B_p)| \quad (4.1) \qquad u(B_p) = \sqrt{\frac{1}{N_p} \sum_{j \in B_p} \frac{1}{K} \sum_{k=1}^{K} u_j^k} \quad (4.2)$$

where $e(B_p)$ denotes the RMSE of N_p pixel predictions in bin B_p, P is the bin count and a bin's uncertainty $u(B_p)$ is given in terms of Root Mean Variance.UCE quantifies the deviation between the predicted uncertainty and the reconstruction error.

Lastly, by sorting samples according to their image-wise uncertainty, UnCRtainTS can be utilized for a fine control of the committed error, as is shown in Figure 4.14.

Figure 4.14: Controlling error on the test split by discarding the most uncertain samples, when images are ranked by their predicted variances. Discarding the top 50% of uncertain reconstructions almost halves prediction error, enabling risk management for optical satellite image reconstruction.



5 Conclusion & Outlook

5.1 Summary and Conclusion

This dissertation addressed the topic of reconstructing cloud-covered information in optical satellite images. The nature of the problem and this work's objectives were outlined in section 1. Clouds pose a long-standing challenge to the remote sensing practitioner, so the topic has garnered a correspondingly voluminous amount of publications over the years. This literature is reviewed in chapter 3. Open challenges of existing approaches were covered in section 3 of this dissertation. In chapter 4, the main contributions of this dissertation were outlined: First, investigating the effects of clouds on remote sensing applications in practice, followed by developing methods for mono-temporal as well as multi-temporal multi-sensor cloud removal, curating the required data for training and benchmarking such methodology—and at last, researching approaches to trustworthy satellite image reconstruction. Finally, this chapter provides a conclusion of the work at hand. To compactly summarize its curated insights:

- **Distribution shift.** Clouds pose a severe obstacle to established remote sensing applications, such as land cover classification. The presence of clouds or cloud shadow causes a salient distribution shift of the spectral characteristics for any land cover type. Beyond its direct effects on learned feature extration, this shift also puts to question any preprocessing pipelines with normalization based on the sufficient statistics computed on cloud-free data.

- **Confidently wrong.** The presence of clouds or cloud shadow is detrimental to scene classification performance, while the neural networks remain fairly confident of their predictions. That is, the margins at which classes get correctly or erroneously predicted over alternatives remain constantly large, even at an increasing level of noise in the observations to be classified.

- **Clouded minds.** An interpretability analysis of scene classification in the presence of clouds or cloud shadow reveals that outlier pixels drive most of a network's misclassifications. Specifically, outstandingly bright, very dark or high-contrast areas often coincide with a focus of attention, shifted away from the remainder of the scene. This indicates a need for more robust models or prior cloud removal.

- **Data needs.** Large scale and real data are required for general cloud removal. Training on synthetic data is facing a domain gap towards real data, such that

models trained on real data perform considerably better than their counterparts trained on generated data.

- **Model diversity.** Several different architectures can be used to to reconstruct cloud covered pixels—residual architectures, generative networks or attention-based models, with the more recent attention-based models performing best. The latter may implement visual attention over the spatial domain, or sequence-based attention for multi-temporal cloud removal.

- **SAR with benefits.** As demonstrated throughout the zoo of networks introduced in this thesis, including radar data to complement the cloud-covered optical observations is beneficial in providing better reconstructions. Furthermore, this may promote a better calibration of predicted uncertainty.

- **Deep learning for time series cloud removal** in globally distributed multi-spectral satellite data is feasible, and outperforms preceding approaches in the sequence-to-point as well as sequence-to-sequence setting alike. With regards to sequence-to-point approaches, longer input time series steadily improve reconstruction quality and auxiliary radar information further boosts performance. For sequence-to-sequence cloud removal, it is shown that deep neural networks following the internal learning approach outperform hand-crafted reconstruction methods which thus far have been the dominant paradigm to the task at hand.

- **Well-calibrated uncertainty estimates** that are indicative of the committed empirical error can be obtained in a pixel-wise manner for the task of multi-spectral satellite image reconstruction. Filtering reconstructions based on their aggregated predicted variances allows for a fine control of error, which may support safety-critical downstream applications.

- **More evidence** leads to better reconstructions as well as improved calibration. Notably, two manners of reaching this goal have been demonstrated to be effective: First, collecting longer time series of satellite data, where additional samples are likely cloud-free, facilitate the restoration task and provide growing evidence for better calibration. Second, the complementary information of SAR inputs for a multi-sensor approach is beneficial to improve the trustworthiness of the reconstructions

- **Stronger together.** Deep ensembles of independently trained neural networks not only provide straightforward means to quantify the epistemic uncertainty of the ensemble, but also boost further benefits in both reconstruction performance and calibration. This may offer a straightforward approach to obtaining better predictions in practice, whenever highly reliable predictions are needed.

5.2 Open Challenges

Although approaches to reconstruct cloud-covered information in optical satellite imagery have progressed considerably, there is still many open challenges for further work to resolve. According to the author of this thesis, the following may be particularly promising directions for future research to investigate:

- **Unified image reconstruction.** The topic of cloud removal is closely related to the tasks of super-resolution [26, 185, 186], image denoising [101, 102, 103, 104, 105, 27] and reconstructing missing data [67, 145]. Methodologically, these four subjects have in common that they may be unified as a single regression task, such that it is primarily the type of data and noise that tells them apart [83, 84, 85, 86]. Hence, future research should consider the interaction between the entirety of these related problems in a common framework for satellite image reconstruction.

- **Higher spatial or spectral resolution.** Furthermore, it may be promising to extend the set of optical sensors that cloud removal models are researched and developed for. Currently, the majority of methods are designed for medium-resolution multi-spectral imagery such as provided by Sentinel-2, owing to its widespread adaptation and ease of availability. However, it would as well be interesting to consider e.g. hyperspectral products [27] or very high resolution spaceborne optical sensors. Either may pose particular technical challenges, associated with the high dimensionality of images to be reconstructed or the more complex compositions and geometry apparent in high resolution imagery.

- **Sequence-to-sequence cloud removal.** Deep learning approaches to sequence-to-sequence cloud removal, while they have initially been covered in the context of the internal learning paradigm by [187, 73], deserve further attention. Specifically, adapting and conventionally training deep sequence-to-sequence reconstruction models [188] on a large dataset such as SEN12MS-CR-TS and subsequently defining a suitable manner of evaluating them is still an open task. It is a promising enterprise nonetheless, as sequence-to-sequence approaches allow to pass information across time points while preserving temporal information rather than integrating it away, as is the case for the established sequence-to-point approach. As such, sequence-to-sequence models should likewise aim for better capturing the temporal dynamics of the processed region of interest, such as changes in land cover due to seasonality or other events.

- **Self-supervised learning.** As deep learning models become increasingly expressive, the need for more data and better supervision grows likewise. Self-supervised learning emerged as a recent paradigm in remote sensing to meet these demands. As initially evidenced in [158], cloud-covered data can serve as an abundantly available source for self-supervised pretraining. Moreover, the task of cloud removal itself may be well-suited for self-supervised pretraining, which should be explored in the future. With a combined volume of over 2 terabytes of curated multi-modal

data pairings, SEN12MS-CR and SEN12MS-CR-TS are well-positioned to serve for further explorations in self-supervised learning.

- **Downstream task evaluation.** Finally, any cloud removal efforts should ultimately prove their practical benefits for common remote sensing applications. Currently, cloud removal methods are primarily evaluated in terms of image reconstruction metrics—which may correlate with their benefits for further downstream applications, but is only an indirect indicator thereof. Originally, the works of [158] and [158] explicitly explored the benefits of reconstructed satellite imagery in the context of common surveying tasks, such as land cover classification and segmentation. Beyond these initial efforts, further research may underline the value proposition of cloud removal and explore additional downstream applications, such as change detection.

5.3 Outlook

Spaceborne Earth observation becomes of ever increasing relevance for urban and environmental monitoring at scale, with more governmental bodies, public institutes and companies adapting remote sensing approaches. In the light of this development, specifically optical multi-spectral satellite data grow increasingly important. While satellite data—overcoming past legislative, technical and infrastructure hurdles—become easier available to the public in raw and preprocessed formats alike, the natural occurrence of haze, clouds and cloud shadow remains a fundamental and persistent problem.

To tackle this principal challenge for optical Earth observation, several promising solutions of both software and hardware-based nature emerged in the recent years. This begs to ask whether research on satellite image reconstruction, and cloud removal in particular, may remain as needed in the future as it has been for the past decades. Closing the thesis at hand, this section briefly contemplates the subject's future outlook.

One emerging trend on the hardware side is the growth in volume of satellite constellations, documented by the rise in launches of industrial and amateur CubeSats [189, 190, 191] This allows for lower revisit times, resulting in an increase of the frequency of image acquisition. Frequent sampling is partly motivated by circumventing clouds and the intention to coincidentally capture clear views. However, this is not for granted as large clouds can linger around for several days in a row, such that shorter revisits may only yield a surplus of cloudy observations. Moreover, current large-volume fleets mostly focus on visible optical measurements, with more costly multi- or even hyper-spectral instruments remaining rare and thus having lower revisit times [4]. Another objective of large constellations is in achieving global low-latency imaging for continuous monitoring and rapid mobilizing [192]. While short revisit times greatly benefit this matter, it foremost indicates an emerging appetite for seamless and low-latency space imagery. By making every acquired image usable and enabling seamless monitoring at any time is how cloud removal can provide value in the future of a rapidly advancing industry.

On the software side, recent publications demonstrated cloud-covered pixels in multi-spectral images can be ignored whenever irrelevant, yet certain tasks may still remain solvable [193, 40]. This raises the question of what the additional benefits may be of including an explicit cloud removal task or performing image reconstruction prior to another downstream task. Recent work confirms the preceding findings but, importantly, also highlights that cloud removal can further improve performance and enhance robustness [158]. Demonstrating the practical benefits of cloud removal in the context of established applications will be a critical objective as the field matures, and choosing suitable tasks for this matter will be a vital endeavor: In general, whenever a task safely permits integrating over one or more dimensions of the data of interest, there may readily exist alternatives to image reconstruction. Where cloud removal is without any alternatives, is whenever none of any spatio-temporal information is redundant and preserving any dimension of data is necessary.

In sum, the natural occurrence of clouds poses a fundamental issue which will continue to persist in the future. Yet, the remote sensing practitioner's repertoire has been growing substantially over the last years in order to meet an equally expanding amount of awaiting challenges. Some of these obstacles can be addressed with a dedicated setup or paradigm, and circumvent the general necessity for cloud removal. Opportunities where cloud removal can in the future demonstrate its benefits, are whenever analysis-ready, seamless or undelayed optical imagery are needed. In particular, this includes manual monitoring by human interpreters, but also automated downstream applications relying on the spatio-temporal integrity of multi-spectral information. The majority of research on optical Earth observation is conducted on idealized, carefully curated data free of any clouds, noise and other artifacts. This creates ample opportunities for cloud removal to demonstrate its benefits, as most readily existing methodology can not be deployed in practice without any further pre-processing.

What is important for the future perspectives of cloud removal and satellite image reconstruction in general, is to mature and prove their value in the remote sensing practitioner's workflow. The author of this thesis thinks that the accumulated contributions of his work have laid the foundation for continuing along this route.

"What does it mean, to see? The plain man's answer (and Aristotle's, too) would be, to know what is where by looking. In other words, vision is the process of discovering from images what is present in the world, and where it is."

— David Marr, Vision (1982)

Bibliography

[1] M. Schmitt, S. A. Ahmadi, and R. Hänsch. There is no data like more data-current status of machine learning datasets in remote sensing. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, pages 1206–1209. IEEE, 2021.

[2] Euroconsult. Euroconsult estimates that the global space economy totaled $370 billion in 2021. Euroconsult Space Economy Report., 2021.

[3] E. I. Journal. Significant supply expansion for eo industry: Data demand driven by defense and emerging markets. Euroconsult Space Economy Report., 2016.

[4] J. McDowell. Jonathan's space report. Jonathan's Space Home Page. `https://planet4589.org/space/index.html`, note = Accessed: 2023-04-01, 2017.

[5] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks. Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites. *IEEE Transactions on Geoscience and Remote Sensing*, 51(7):3826–3852, Jul 2013. `doi:10.1109/TGRS.2012.2227333`.

[6] J. D. Bermudez, P. N. Happ, D. A. B. Oliveira, and R. Q. Feitosa. SAR to optical image synthesis for cloud removal with Generative Adversarial Networks. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV–1:5–11, Sep 2018. `doi:10.5194/isprs-annals-IV-1-5-2018`.

[7] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[8] M. Minsky and S. A. Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT press, 1988.

[9] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[10] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[11] D. J. MacKay and D. J. Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[12] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[13] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL: `https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[16] V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[17] C. Wang, H. Huang, X. Han, and J. Wang. Video inpainting by jointly learning temporal structure and spatial details. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5232–5239, 2019.

[18] H. Zhang, L. Mai, N. Xu, Z. Wang, J. Collomosse, and H. Jin. An internal learning approach to video inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2720–2729, 2019.

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Y. Bengio and Y. LeCun, editors, *International Conference on Learning Representations*, 2015.

[20] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[21] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[25] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[26] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:305–319, 2018.

[27] Q. Yuan, Q. Zhang, J. Li, H. Shen, and L. Zhang. Hyperspectral image denoising employing a spatial–spectral deep residual convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):1205–1218, 2019. `doi: 10.1109/TGRS.2018.2865197`.

[28] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems*, 31, 2018.

[29] J. Li, G. Li, and H. Fan. Image dehazing using residual-based deep cnn. *IEEE Access*, 6:26831–26842, 2018.

[30] Z. Dai, H. Liu, Q. V. Le, and M. Tan. Coatnet: Marrying convolution and attention for all data sizes. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3965–3977. Curran Associates, Inc., 2021. URL: `https://proceedings.neurips.cc/paper_files/paper/2021/file/20568692db622456cc42a2e853ca21f8-Paper.pdf`.

[31] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*, 2014.

[32] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL: `https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf`.

[34] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017.

[35] Y. Mroueh and T. Sercu. Fisher GAN. *Advances in Neural Information Processing Systems*, 30, 2017.

[36] J. Zhao, M. Mathieu, and Y. LeCun. Energy-based Generative Adversarial Network. *arXiv preprint arXiv:1609.03126*, 2016.

[37] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares Generative Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.

[38] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[40] V. S. F. Garnot and L. Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4872–4881, 2021.

[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10012–10022, 2021.

[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[43] D. A. Nix and A. S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of the IEEE International Conference on Neural Networks*, volume 1, pages 55–60. IEEE, 1994.

[44] C. M. Bishop. Mixture density networks. 1994. URL: `https://publications.aston.ac.uk/id/eprint/373/`.

[45] M.-H. Laves, S. Ihler, J. F. Fast, L. A. Kahrs, and T. Ortmaier. Well-calibrated regression uncertainty in medical imaging with deep learning. In *Medical Imaging with Deep Learning*, pages 393–412. PMLR, 2020.

[46] S. Bhadra, V. A. Kelkar, F. J. Brooks, and M. A. Anastasio. On hallucinations in tomographic image reconstruction. *IEEE transactions on medical imaging*, 40(11):3249–3260, 2021.

[47] M.-H. Laves, M. Tölle, and T. Ortmaier. Uncertainty estimation in medical image denoising with Bayesian deep image prior. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, pages 81–96. Springer, 2020.

[48] M. Tölle, M.-H. Laves, and A. Schlaefer. A mean-field variational inference approach to deep image prior for inverse problems in medical imaging. In *Medical Imaging with Deep Learning*, pages 745–760. PMLR, 2021.

[49] J. Antorán, R. Barbano, J. Leuschner, J. M. Hernández-Lobato, and B. Jin. Uncertainty estimation for computed tomography with a linearised deep image prior, 2022. `arXiv:2203.00479`.

[50] H. Chung, E. S. Lee, and J. C. Ye. MR image denoising and super-resolution using regularized reverse diffusion. *arXiv preprint arXiv:2203.12621*, 2022.

[51] A. Kendall and Y. Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017.

[52] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

[53] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.

[54] H. Takahashi, T. Iwata, Y. Yamanaka, M. Yamada, and S. Yagi. Student-t variational autoencoder for robust density estimation. In *International Joint Conference on Artificial Intelligence*, pages 2696–2702, 2018.

[55] N. Skafte, M. Jørgensen, and S. Hauberg. Reliable training and estimation of variance networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[56] M. Seitzer, A. Tavakoli, D. Antic, and G. Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*, 2021.

[57] N. Ansari, H.-p. Seidel, N. V. Ferdowsi, and V. Babaei. Autoinverse: Uncertainty aware inversion of neural networks. In *Advances in Neural Information Processing Systems*, 2022.

[58] A. Stirn, H.-H. Wessels, M. Schertzer, L. Pereira, N. E. Sanjana, and D. A. Knowles. Faithful heteroscedastic regression with neural networks. *arXiv preprint arXiv:2212.09184*, 2022.

[59] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

[60] N. Lang, N. Kalischek, J. Armston, K. Schindler, R. Dubayah, and J. D. Wegner. Global canopy height regression and uncertainty estimation from gedi lidar waveforms with deep ensembles. *Remote Sensing of Environment*, 268:112760, 2022.

[61] N. Lang, W. Jetz, K. Schindler, and J. D. Wegner. A high-resolution canopy height model of the Earth. *arXiv preprint arXiv:2204.08322*, 2022.

[62] P. Chaudhary, J. P. Leitão, T. Donauer, S. D'Aronco, N. Perraudin, G. Obozinski, F. Perez-Cruz, K. Schindler, J. D. Wegner, and S. Russo. Flood uncertainty estimation using deep ensembles. *Water*, 14(19):2980, 2022.

[63] D. Spänkuch, O. Hellmuth, and U. Görsdorf. What is a cloud? toward a more precise definition? *Bulletin of the American Meteorological Society*, 2022.

[64] T. Nakajima and M. D. King. Determination of the optical thickness and effective particle radius of clouds from reflected solar radiation measurements. Part I: Theory. *Journal of Atmospheric Sciences*, 47(15):1878–1893, 1990.

[65] C. F. Bohren, J. R. Linskens, and M. E. Churma. At what optical thickness does a cloud completely obscure the sun? *Journal of the atmospheric sciences*, 52(8):1257–1259, 1995.

[66] M. D. King, S.-C. Tsay, S. E. Platnick, M. Wang, and K.-N. Liou. Cloud retrieval algorithms for MODIS: Optical thickness, effective particle radius, and thermodynamic phase. *MODIS Algorithm Theoretical Basis Document*, 1997, 1997.

[67] J. Storey, J. Lacasse, R. Smilek, T. Zeiler, P. Scaramuzza, R. Rengarajan, and M. Choate. Image impact of the Landsat 7 ETM+ scan line corrector failure. USGS Report., 2005.

[68] W. L. Barnes, X. Xiong, and V. V. Salomonson. Status of Terra MODIS and Aqua MODIS. *Advances in Space Research*, 32(11):2099–2106, 2003.

[69] P. Rakwatin, W. Takeuchi, and Y. Yasuoka. Restoration of aqua MODIS band 6 using histogram matching and local least squares fitting. *IEEE Transactions on Geoscience and Remote Sensing*, 47(2):613–627, 2008.

[70] U. Lohmann, F. Lüönd, and F. Mahrt. *An introduction to clouds: From the microscale to climate*. Cambridge University Press, 2016.

[71] R. Pincus, S. Platnick, S. A. Ackerman, R. S. Hemler, and R. J. P. Hofmann. Reconciling simulated and observed views of clouds: MODIS, ISCCP, and the limits of instrument simulators. *Journal of Climate*, 25(13):4699–4720, 2012.

[72] M. Schmitt and X. X. Zhu. Data fusion and remote sensing: An ever-growing relationship. *IEEE Geoscience and Remote Sensing Magazine*, 4(4):6–23, 2016.

[73] P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu. SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[74] Y. Wang and X. X. Zhu. The SARptical dataset for joint analysis of SAR and optical image in dense urban area. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, pages 6840–6843. IEEE, 2018.

[75] R. Dubayah, J. B. Blair, S. Goetz, L. Fatoyinbo, M. Hansen, S. Healey, M. Hofton, G. Hurtt, J. Kellner, S. Luthcke, et al. The global ecosystem dynamics investigation: High-resolution laser ranging of the Earth's forests and topography. *Science of remote sensing*, 1:100002, 2020.

[76] J. Rosentreter, K. Fenske, H. Feilhauer, M. Stellmes, and R. Rissiek. Remote sensing data analysis online course - sentinel 2. Freie Universität Berlin Blog. `https://blogs.fu-berlin.de/reseda/sentinel-2/`, note = Accessed: 2023-04-01, 2022.

[77] K. Tomiyasu. Tutorial review of synthetic-aperture radar (SAR) with applications to imaging of the ocean surface. *Proceedings of the IEEE*, 66(5):563–583, 1978.

[78] R. Bamler. Principles of synthetic aperture radar. *Surveys in Geophysics*, 21(2-3):147–157, 2000.

[79] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, and K. P. Papathanassiou. A tutorial on synthetic aperture radar. *IEEE Geoscience and remote sensing magazine*, 1(1):6–43, 2013.

[80] J. W. Goodman. Some fundamental properties of speckle. *JOSA*, 66(11):1145–1150, 1976.

[81] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, Apr 2004. `doi:10.1109/TIP.2003.819861`.

[82] F. A. Kruse, A. Lefkoff, J. Boardman, K. Heidebrecht, A. Shapiro, P. Barloon, and A. Goetz. The spectral image processing system (SIPS)-interactive visualization and analysis of imaging spectrometer data. In *AIP Conference Proceedings*, volume 283, pages 192–201. American Institute of Physics, 1993.

[83] H. Shen, X. Li, Q. Cheng, C. Zeng, G. Yang, H. Li, and L. Zhang. Missing information reconstruction of remote sensing data: A technical review. *IEEE Geoscience and Remote Sensing Magazine*, 3(3):61–85, 2015.

[84] D. Yang, Z. Li, Y. Xia, and Z. Chen. Remote sensing image super-resolution: Challenges and approaches. In *2015 IEEE international conference on digital signal processing (DSP)*, pages 196–200. IEEE, 2015.

[85] G. Tsagkatakis, A. Aidini, K. Fotiadou, M. Giannopoulos, A. Pentari, and P. Tsakalides. Survey of deep-learning approaches for remote sensing observation enhancement. *Sensors*, 19(18):3929, 2019.

[86] P. Wang, B. Bayram, and E. Sertel. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews*, page 104110, 2022.

[87] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, pages 417–424, 2000.

[88] M. Richard and M. Chang. Fast digital image inpainting. In *Proceedings of the International Conference on Visualization, Imaging and Image Processing*, pages 106–107, 2001.

[89] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher. Simultaneous structure and texture image inpainting. *IEEE Transactions on Image Processing*, 12(8):882–889, 2003.

[90] A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, 2004.

[91] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

[92] C. Guillemot and O. Le Meur. Image inpainting: Overview and recent advances. *IEEE Signal Processing Magazine*, 31(1):127–144, 2013.

[93] L. Lorenzi, F. Melgani, and G. Mercier. Inpainting strategies for reconstruction of missing data in vhr images. *IEEE Geoscience and Remote Sensing Letters*, 8(5):914–918, 2011.

[94] F. Chen, Z. Zhao, L. Peng, and D. Yan. Clouds and cloud shadows removal from high-resolution remote sensing images. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, volume 6, pages 4256–4259. Ieee, 2005.

[95] A. Maalouf, P. Carré, B. Augereau, and C. Fernandez-Maloigne. A bandelet-based inpainting technique for clouds removal from remotely sensed images. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2363–2371, 2009.

[96] Q. Meng, B. E. Borders, C. J. Cieszewski, and M. Madden. Closest spectral fit for removing clouds and cloud shadows. *Photogrammetric Engineering & Remote Sensing*, 75(5):569–576, 2009.

[97] A. C. Siravenha, D. Sousa, A. Bispo, and E. Pelaes. Evaluating inpainting methods to the satellite images clouds and shadows removing. In *International Conference on Signal Processing, Image Processing and Pattern Recognition*, pages 56–65. Springer, 2011.

[98] C. Yu, L. Chen, L. Su, M. Fan, and S. Li. Kriging interpolation method and its application in retrieval of MODIS aerosol optical depth. In *International Conference on Geoinformatics*, pages 1–6. IEEE, 2011.

[99] H. Shen, H. Li, Y. Qian, L. Zhang, and Q. Yuan. An effective thin cloud removal procedure for visible remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 96:224–235, 2014.

[100] M. Xu, X. Jia, M. Pickering, and S. Jia. Thin cloud removal from optical remote sensing images using the noise-adjusted principal components transform. *ISPRS Journal of Photogrammetry and Remote Sensing*, 149:215–225, 2019.

[101] H. Zhang, W. He, L. Zhang, H. Shen, and Q. Yuan. Hyperspectral image restoration using low-rank matrix recovery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(8):4729–4743, 2013.

[102] H. Fan, Y. Chen, Y. Guo, H. Zhang, and G. Kuang. Hyperspectral image restoration using low-rank tensor recovery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(10):4589–4604, 2017.

[103] Y. Wang, J. Peng, Q. Zhao, Y. Leung, X.-L. Zhao, and D. Meng. Hyperspectral image restoration via total variation regularized low-rank tensor decomposition. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(4):1227–1243, 2017.

[104] Y. Chen, W. He, N. Yokoya, and T.-Z. Huang. Hyperspectral image restoration using weighted group sparsity-regularized low-rank tensor decomposition. *IEEE Transactions on Cybernetics*, 50(8):3556–3570, 2019.

[105] Y.-B. Zheng, T.-Z. Huang, X.-L. Zhao, T.-X. Jiang, T.-H. Ma, and T.-Y. Ji. Mixed noise removal in hyperspectral image via low-fibered-rank regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):734–749, 2019.

[106] N. T. Hoan and R. Tateishi. Cloud removal of optical image using sar data for alos applications. experimenting on simulated alos data. *Journal of The Remote Sensing Society of Japan*, 29(2):410–417, 2009.

[107] R. Eckardt, C. Berger, C. Thiel, and C. Schmullius. Removal of optically thick clouds from multi-spectral satellite images using multi-frequency sar data. *Remote Sensing*, 5(6):2973–3006, 2013.

[108] X. Li, H. Shen, L. Zhang, H. Zhang, Q. Yuan, and G. Yang. Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning. *IEEE Transactions on Geoscience and Remote Sensing*, 52(11):7086–7098, 2014.

[109] X. Li, H. Shen, L. Zhang, and H. Li. Sparse-based reconstruction of missing information in remote sensing images from spectral/temporal complementary information. *ISPRS Journal of Photogrammetry and Remote Sensing*, 106:1–15, 2015.

[110] B. Huang, Y. Li, X. Han, Y. Cui, W. Li, and R. Li. Cloud removal from optical satellite imagery with sar imagery using sparse representation. *IEEE Geoscience and Remote Sensing Letters*, 12(5):1046–1050, 2015.

[111] X. Zhu, F. Gao, D. Liu, and J. Chen. A modified neighborhood similar pixel interpolator approach for removing thick clouds in Landsat images. *IEEE Geoscience and Remote Sensing Letters*, 9(3):521–525, 2011.

[112] B. Chen, B. Huang, L. Chen, and B. Xu. Spatially and temporally weighted regression: A novel method to produce continuous cloud-free Landsat imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 55(1):27–37, 2016.

[113] D. Cerra, J. Bieniarz, F. Beyer, J. Tian, R. Müller, T. Jarmer, and P. Reinartz. Cloud removal in image time series through sparse reconstruction from random measurements. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(8):3615–3628, 2016.

[114] F. Melgani. Contextual reconstruction of cloud-contaminated multitemporal multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 44(2):442–455, 2006.

[115] S. Benabdelkader and F. Melgani. Contextual spatiospectral postreconstruction of cloud-contaminated images. *IEEE Geoscience and Remote Sensing Letters*, 5(2):204–208, 2008.

[116] L. Lorenzi, G. Mercier, and F. Melgani. Support vector regression with kernel combination for missing data reconstruction. *IEEE Geoscience and Remote Sensing Letters*, 10(2):367–371, 2012.

[117] S. Skakun, J. Wevers, C. Brockmann, G. Doxani, M. Aleksandrov, M. Batič, D. Frantz, F. Gascon, L. Gómez-Chova, O. Hagolle, et al. Cloud mask intercomparison exercise (cmix): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. *Remote Sensing of Environment*, 274:112990, 2022.

[118] F. De la Torre and M. J. Black. Robust principal component analysis for computer vision. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 1, pages 362–369. IEEE, 2001.

[119] S. Hauberg, A. Feragen, and M. J. Black. Grassmann averages for scalable robust PCA. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3810–3817, 2014.

[120] M. Hintermüller and T. Wu. Robust principal component pursuit via inexact alternating minimization on matrix manifolds. *Journal of Mathematical Imaging and Vision*, 51(3):361–377, 2015.

[121] D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13:556–562, 2000.

[122] S. S. Bucak, B. Gunsel, and O. Gursoy. Incremental nonnegative matrix factorization for background modeling in surveillance video. In *Signal Processing and Communications Applications*, pages 1–4. IEEE, 2007.

[123] N. Guan, D. Tao, Z. Luo, and J. Shawe-Taylor. MahNMF: Manhattan non-negative matrix factorization. *arXiv preprint arXiv:1207.3438*, 2012.

[124] A. Sobral, S. Javed, S. Ki Jung, T. Bouwmans, and E.-h. Zahzah. Online stochastic tensor decomposition for background subtraction in multispectral video sequences.

In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 106–113, 2015.

[125] Y. Chen, H. Xu, C. Caramanis, and S. Sanghavi. Robust matrix completion and corrupted columns. In *International Conference on Machine Learning*, pages 873–880, 2011.

[126] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7):4324–4337, 2013.

[127] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2012.

[128] L. Lorenzi, F. Melgani, and G. Mercier. Missing-area reconstruction in multispectral images under a compressive sensing perspective. *IEEE Transactions on Geoscience and Remote Sensing*, 51(7):3998–4008, 2013.

[129] J. Wang, P. A. Olsen, A. R. Conn, and A. C. Lozano. Removing clouds and recovering ground observations in satellite image sequences via temporally contiguous robust matrix completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2754–2763, 2016.

[130] M. K.-P. Ng, Q. Yuan, L. Yan, and J. Sun. An adaptive weighted tensor completion method for the recovery of remote sensing images with missing data. *IEEE Transactions on Geoscience and Remote Sensing*, 55(6):3367–3381, 2017.

[131] Y. Zhang, F. Wen, Z. Gao, and X. Ling. A coarse-to-fine framework for cloud removal in remote sensing image sequence. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5963–5974, 2019.

[132] W. He, N. Yokoya, L. Yuan, and Q. Zhao. Remote sensing image reconstruction using tensor ring completion and total variation. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):8998–9009, 2019.

[133] Y. Chen, W. He, N. Yokoya, and T.-Z. Huang. Blind cloud and cloud shadow removal of multitemporal images based on total variation regularized low-rank sparsity decomposition. *ISPRS Journal of Photogrammetry and Remote Sensing*, 157:93–107, 2019.

[134] X. Li, L. Wang, Q. Cheng, P. Wu, W. Gan, and L. Fang. Cloud removal in remote sensing images using nonnegative matrix factorization and error correction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 148:103–113, 2019.

[135] K. Enomoto, K. Sakurada, W. Wang, H. Fukui, M. Matsuoka, R. Nakamura, and N. Kawaguchi. Filmy cloud removal on satellite imagery with multispectral conditional Generative Adversarial Nets. *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition Workshops, page 1533–1541, Jul 2017. arXiv: 1710.04835. doi:10.1109/CVPRW.2017.197.

[136] M. Qin, F. Xie, W. Li, Z. Shi, and H. Zhang. Dehazing for multispectral remote sensing images based on a convolutional neural network with the residual architecture. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(5):1645–1655, 2018.

[137] L. Wang, X. Xu, Y. Yu, R. Yang, R. Gui, Z. Xu, and F. Pu. SAR-to-optical image translation using supervised cycle-consistent adversarial networks. *IEEE Access*, 7:129136–129149, 2019. doi:10.1109/ACCESS.2019.2939649.

[138] M. Fuentes Reyes, S. Auer, N. Merkle, C. Henry, and M. Schmitt. SAR-to-optical image translation based on conditional Generative Adversarial Networks—optimization, opportunities and limits. *Remote Sensing*, 11(17):2067, 2019.

[139] C. Grohnfeldt, M. Schmitt, and X. Zhu. A conditional Generative Adversarial Network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, pages 1726–1729. IEEE, 2018.

[140] W. Li, Y. Li, and J. C.-W. Chan. Thick cloud removal with optical and SAR imagery via convolutional-mapping-deconvolutional network. *IEEE Transactions on Geoscience and Remote Sensing*, 58(4):2865–2879, 2019.

[141] K. Perlin. Improving noise. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, pages 681–682, 2002.

[142] T. Porter and T. Duff. Compositing digital images. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, pages 253–259, 1984.

[143] E. J. McCartney. Optics of the atmosphere: scattering by molecules and particles. *New York*, 1976.

[144] D. Lin, G. Xu, X. Wang, Y. Wang, X. Sun, and K. Fu. A remote sensing image dataset for cloud removal. *arXiv:1901.00600*, 2019.

[145] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei. Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8):4274–4288, 2018. doi:10.1109/TGRS.2018.2810208.

[146] Y. Chen, L. Tang, X. Yang, R. Fan, M. Bilal, and Q. Li. Thick clouds removal from multitemporal ZY-3 satellite images using deep learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:143–153, 2019.

[147] Y. Chen, Q. Weng, L. Tang, X. Zhang, M. Bilal, and Q. Li. Thick clouds removing from multitemporal Landsat images using spatiotemporal neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2020.

[148] Q. Zhang, Q. Yuan, J. Li, Z. Li, H. Shen, and L. Zhang. Thick cloud and cloud shadow removal in multitemporal imagery using progressively spatio-temporal patch group deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:148–160, 2020.

[149] V. Sarukkai, A. Jain, B. Uzkent, and S. Ermon. Cloud removal from satellite images using spatiotemporal generator networks. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1796–1805, 2020.

[150] P. Helber, B. Bischke, A. Dengel, and D. Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[151] M. Schmitt, L. Hughes, C. Qiu, and X. Zhu. SEN12MS-a curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:153–160, 2019.

[152] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl. BigEarthNet: A large-scale benchmark archive for remote sensing image understanding. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019.

[153] J. Gawlikowski, P. Ebel, M. Schmitt, and X. X. Zhu. Explaining the effects of clouds on remote sensing scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:9976–9986, 2022.

[154] P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu. Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):5866–5878, 2020.

[155] M. Schmitt and Y.-L. Wu. Remote sensing image classification with the SEN12MS dataset. *ISPRS Annals Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-2-2021:101–106, 2021.

[156] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[157] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[158] Z. Gu, P. Ebel, M. Schmitt, and X. X. Zhu. Explicit haze and cloud removal for global land cover classification. *CVPR 2022 Workshop on Multimodal Learning for Earth and Environment*, pages 1–6, 2022.

[159] F. Xu, Y. Shi, P. Ebel, W. Yang, and X. X. Zhu. Multi-modal and multi-resolution data fusion for high-resolution cloud removal: A novel baseline and benchmark. *In Review. https://arxiv.org/abs/2301.03432*, 2023.

[160] R. Cresson, D. Ienco, R. Gaetano, K. Ose, and D. H. T. Minh. Optical image gap filling using deep convolutional autoencoder from optical and radar images. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, pages 218–221. IEEE, 2019.

[161] P. Singh and N. Komodakis. Cloud-GAN: Cloud removal for Sentinel-2 imagery using a Cyclic Consistent Generative Adversarial Networks. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, pages 1772–1775, 2018. doi:10.1109/IGARSS.2018.8519033.

[162] M. Rafique, H. Blanton, and N. Jacobs. Weakly supervised fusion of multiple overhead images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1479–1486. IEEE, 2019.

[163] J. Li, Z. Wu, Z. Hu, Z. Li, Y. Wang, and M. Molinier. Deep learning based thin cloud removal fusing vegetation red edge and short wave infrared spectral information for Sentinel-2a imagery. *Remote Sensing*, 13(1):157, 2021.

[164] N. Yokoya, P. Ghamisi, R. Hansch, and M. Schmitt. Report on the 2020 IEEE GRSS data fusion contest-global land cover mapping with weak supervision [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 8(4):134–137, 2020.

[165] H. Pan. Cloud removal for remote sensing imagery via spatial attention Generative Adversarial Network. *arXiv preprint arXiv:2009.13015*, 2020.

[166] J. Gao, Q. Yuan, J. Li, H. Zhang, and X. Su. Cloud removal with fusion of high resolution optical and SAR images using Generative Adversarial Networks. *Remote Sensing*, 12(1):191, 2020.

[167] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:333–346, 2020.

[168] F. Xu, Y. Shi, P. Ebel, L. Yu, G.-S. Xia, W. Yang, and X. X. Zhu. GLF-CR: SAR-enhanced cloud removal with global–local fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 192:268–278, 2022.

[169] R. Zhang, F. Xie, and J. Chen. Single image thin cloud removal for remote sensing images based on conditional Generative Adversarial Nets. In X. Jiang and J.-N. Hwang, editors, *International Conference on Digital Image Processing*, volume

10806. International Society for Optics and Photonics, SPIE, 2018. doi:10.1117/12.2503148.

[170] J. Zheng, X.-Y. Liu, and X. Wang. Single image cloud removal using U-Net and Generative Adversarial Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8):6371–6385, 2021. doi:10.1109/TGRS.2020.3027819.

[171] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

[172] T. J. Kozubowski and K. Podgórski. A multivariate and asymmetric generalization of laplace distribution. *Computational Statistics*, 15:531–540, 2000.

[173] K. Yu and R. A. Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447, 2001.

[174] A. Zupanc. Improving cloud detection with machine learning. Sentinel-Hub. https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13, 2017. Accessed: 2019-10-10.

[175] D. S. Candra, S. Phinn, and P. Scarth. Cloud and cloud shadow removal of Landsat 8 images using multitemporal cloud removal method. In *International Conference on Agro-Geoinformatics*, pages 1–5. IEEE, 2017.

[176] M. Schmitt, L. Hughes, C. Qiu, and X. X. Zhu. Aggregating cloud-free Sentinel-2 images with Google Earth Engine. *PIA19: Photogrammetric Image Analysis*, pages 145–152.

[177] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017. doi:10.1109/TCI.2016.2644865.

[178] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for Generative Adversarial Networks. In *International Conference on Learning Representations*, 2018.

[179] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1833–1844, 2021.

[180] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

[181] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

[182] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[183] V. S. F. Garnot and L. Landrieu. Lightweight temporal self-attention for classifying satellite images time series. In *International Workshop on Advanced Analytics and Learning on Temporal Data*, pages 171–181. Springer, 2020.

[184] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[185] M. R. Arefin, V. Michalski, P.-L. St-Charles, A. Kalaitzis, S. Kim, S. E. Kahou, and Y. Bengio. Multi-image super-resolution for remote sensing using deep recurrent networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 206–207, 2020.

[186] J. Cornebise, I. Orsolic, and F. Kalaitzis. Open high-resolution satellite imagery: The worldstrat dataset–with application to super-resolution. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[187] P. Ebel, M. Schmitt, and X. X. Zhu. Internal learning for sequence-to-sequence cloud removal via synthetic aperture radar prior information. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, pages 2691–2694. IEEE, 2021.

[188] Y. Zeng, J. Fu, and H. Chao. Learning joint spatial-temporal transformations for video inpainting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 528–543. Springer, 2020.

[189] M. Swartwout. The first one hundred cubesats: A statistical look. *Journal of small Satellites*, 2(2):213–233, 2013.

[190] T. Villela, C. A. Costa, A. M. Brandão, F. T. Bueno, R. Leonardi, et al. Towards the thousandth cubesat: A statistical overview. *International Journal of Aerospace Engineering*, 2019, 2019.

[191] M. Swartwout. Cubesats/smallsats/nanosats/picosats/rideshare (sats) in 2022: Making sense of the numbers. In *2022 IEEE Aerospace Conference (AERO)*, pages 1–10. IEEE, 2022.

[192] J. Morrison. A simple mental model for understanding the satellite imagery industry. A Closer Look with Joe Morrison. `https://joemorrison.substack.com/p/a-simple-mental-model-for-understanding`, note = Accessed: 2023-04-01, 2022.

[193] M. Rußwurm and M. Körner. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 11–19, 2017.

A Appendix: Publications

A.1 Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion

Reference: A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt. *Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion.* ISPRS Journal of Photogrammetry and Remote Sensing, 166:333–346, 2020

ISPRS Journal of Photogrammetry and Remote Sensing 166 (2020) 333–346

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion

Andrea Meraner[a,1], Patrick Ebel[a], Xiao Xiang Zhu[a,b,*], Michael Schmitt[a,*]

[a] Signal Processing in Earth Observation, Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany
[b] Remote Sensing Technology Institute, German Aerospace Center (DLR), Münchener Straße 20, 82234 Weßling-Oberpfaffenhofen, Germany

ARTICLE INFO

Keywords:
Cloud removal
Optical imagery
SAR-optical
Data fusion
Deep learning
Residual network

ABSTRACT

Optical remote sensing imagery is at the core of many Earth observation activities. The regular, consistent and global-scale nature of the satellite data is exploited in many applications, such as cropland monitoring, climate change assessment, land-cover and land-use classification, and disaster assessment. However, one main problem severely affects the temporal and spatial availability of surface observations, namely cloud cover. The task of removing clouds from optical images has been subject of studies since decades. The advent of the Big Data era in satellite remote sensing opens new possibilities for tackling the problem using powerful data-driven deep learning methods.

In this paper, a deep residual neural network architecture is designed to remove clouds from multispectral Sentinel-2 imagery. SAR-optical data fusion is used to exploit the synergistic properties of the two imaging systems to guide the image reconstruction. Additionally, a novel cloud-adaptive loss is proposed to maximize the retainment of original information. The network is trained and tested on a globally sampled dataset comprising real cloudy and cloud-free images. The proposed setup allows to remove even optically thick clouds by reconstructing an optical representation of the underlying land surface structure.

1. Introduction

1.1. Motivation

While the quality and quantity of satellite observations dramatically increased in recent years, one common problem persists for remote sensing in the optical domain since the first observation until today: cloud cover. As thick clouds appear opaque in all optical frequency bands, the presence thereof completely corrupts the reflectance signal and obstructs the view of the surface underneath. This causes considerable data gaps in both the spatial and temporal domains. For applications where consistent time series are needed, e.g. agricultural monitoring, or where a certain scene must be observed at a specific time, e.g. disaster monitoring, cloud cover represents a serious hindrance.

The problem of cloud cover becomes even more apparent considering the amount of cloud coverage the Earth's surface experiences every day. An analysis over 12 years of observations by the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument aboard the satellites Terra and Aqua showed that 67% of the Earth's surface is covered by clouds on average (King et al., 2013). Over land surfaces, the cloud fraction averages to 55%, featuring distinctive seasonal patterns. Considering the importance of these cloud occlusion percentages, it becomes clear how a successful cloud removal algorithm would greatly increase the availability of useful data. The task of detecting and removing clouds from satellite images has been tackled since the beginning of Earth observation activities, and is still today an area of active research. In this work, we present a deep learning model capable of removing clouds from Sentinel-2 images. The network design and the integration of additional Sentinel-1 SAR data makes it robust to extensive cloud coverage conditions. The model is trained on a large dataset containing scenes acquired globally, ensuring its general applicability on any land cover type.

1.2. Related works

The reconstruction of missing information in remote sensing data is a long-studied problem. In Shen et al. (2015), a comprehensive review

* Corresponding authors at: Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, 82234 Wessling, Germany and Signal Processing in Earth Observation, Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany.
E-mail addresses: andrea.meraner@eumetsat.int (A. Meraner), patrick.ebel@tum.de (P. Ebel), xiaoxiang.zhu@dlr.de (X.X. Zhu), m.schmitt@tum.de (M. Schmitt).
[1] Present address: EUMETSAT, Eumetsat Allee 1, 64295 Darmstadt, Germany.

https://doi.org/10.1016/j.isprsjprs.2020.05.013
Received 14 January 2020; Received in revised form 15 May 2020; Accepted 18 May 2020
Available online 02 July 2020

of traditional techniques in provided. In the last decades, a multitude of approaches have been proposed for the specific task of cloud removal in optical imagery. Methods that follow traditional approaches can be categorized into three major clusters, namely multispectral, multitemporal and inpainting techniques. Many methods are a hybrid combination of these categories. Multispectral approaches are applied in the case of haze and thin cirrus clouds, where optical signals are not completely blocked but experience partial wavelength-dependent absorption and reflection. In such cases, surface information is partly present and can be restored, e.g. using mathematical (Xu et al., 2019; Hu et al., 2015) or physical models (Xu et al., 2016; Lv et al., 2016). Multispectral methods have the advantage of exploiting information from the original scene without requiring additional data, but are limited to filmy, semi-transparent clouds. Multitemporal approaches restore cloudy scenes by integrating information from reference images acquired with clear sky conditions (Lin et al., 2013; Li et al., 2015; Ramoino et al., 2017; Ji et al., 2018). For this, also multitemporal dictionary learning techniques can be used (Li et al., 2014). The multitemporal data may also come from different sensors on different satellites (Li et al., 2019). Multitemporal methods are the most popular as they substitute corrupted pixels with real cloud-free observations. However, problems arise when reconstructing scenes with rapidly changing surface conditions (e.g. due to phenological events) because of the time difference between the scene to be reconstructed and the reference acquisition. Inpainting approaches fill corrupted regions by exploiting surface information from clear parts of the same cloud-affected image (Meng et al., 2017). Such direct inpainting methods do not require additional images, but achieve good results only with small clouds. To mitigate this problem, the process of selecting the most suitable similar pixel to be cloned is often guided by auxiliary data, e.g. multitemporal (Cheng et al., 2014) or SAR images (Eckardt et al., 2013). Such methods deliver good results but have an increased complexity due to the requirement of multitemporal or multisensorial additional data.

In parallel to traditional approaches for cloud removal, data-driven methods using deep learning have been gaining attention recently. Many of the problems arising from traditional algorithms can be potentially solved by the end-to-end learning of deep neural networks (DNN). For example, the detection and segmentation of clouds as a preliminary step is often not required, as it can be learned implicitly by the networks. In the case of multisensor data fusion, the translation between different sensor domains can also be learned. Moreover, DNNs can be trained to cope with any type of cloud and residual atmospheric conditions. A first paper exploiting the potential of DNNs for restoring missing information in remote sensing imagery was published in Zhang et al. (2018). The method uses a spatial–temporal-spectral convolutional neural network (CNN) to restore data gaps in Landsat TM data. In the case of clouds, an additional multitemporal image of the same scene is used to support the reconstruction. Recent papers have been focusing on using a modern CNN architecture called conditional generative adversarial network (cGAN) (Mirza and Osindero, 2014). In Enomoto et al. (2017), a cGAN is trained to remove simulated clouds from Worldview-2 RGB images using NIR images as auxiliary data, while in Grohnfeldt et al. (2018) a cGAN removes simulated clouds from Sentinel-2 imagery using SAR data as additional information. An evolution of the cGAN, called Cycle-GAN, can be used to avoid the need of paired cloudy-cloudfree images for training (Singh and Komodakis, 2018). A different approach for generating cloud-free images is to perform a direct translation from SAR to optical using cGANs (Bermudez et al., 2018; Bermudez et al., 2019; He and Yokoya, 2018; Fuentes Reyes et al., 2019). Besides their powerful generative capabilities, cGANs can suffer from training and prediction instabilities when fed with bad input data (e.g. large cloud coverage), as reported in some of the referenced studies and in Mescheder et al. (2018). Based on these experiences, the work presented in this paper develops a model architecture that is robust to the presence of large and optically thick clouds in the input data.

In addition to the conceptual considerations, the need of large datasets is also a prominent problem in deep learning for cloud removal. The studies cited above achieve promising results, but the used datasets are very limited and the performance is evaluated on non-independent data. An assessment of the generalization capability of the networks, i.e. their ability to remove clouds on previously unseen scenes, is therefore not directly possible. In contrast, we present and use a large dataset that is suited for a deterministic separation of images for training and testing purposes and thus provides a sound idea of how well the network will generalize to unseen Sentinel-2 data.

1.3. Paper structure

This paper is structured as follows. After this introductory section, the characteristics of the used dataset are presented in Section 2. The proposed methodology, including the designed neural network architecture and custom loss, are explained in Section 3. The conducted experiments and obtained results are then presented in Section 4 and further discussed in 5. Finally, a summary and conclusions are given in Section 6.

2. Data

While the data-driven method proposed in this paper is of generic nature and sensor-agnostic, the specific model we train and our experiments focus on satellite imagery provided by the Sentinel satellites of the European Copernicus Earth observation program (Desnos et al., 2014), as these data are globally and freely available in a user-friendly manner.

2.1. Sentinel-1 and Sentinel-2 missions

The cloud removal algorithm developed in this work is applied on optical data from the Copernicus Sentinel-2 mission (Drusch et al., 2012). The mission provides data for risk management, land use/land cover and environmental monitoring, as well as urban and terrestrial mapping for humanitarian and development aid. Imagery is available over all main land areas from -56° to 84° of latitude with a global revisit time of 5 days at the equator. The optical payload is called Multi Spectral Instrument (MSI) and comprises 13 spectral bands. Four 10 m high-resolution bands are placed in the visible and NIR domain for core mapping applications. Six 20 m resolution bands are used for environmental monitoring and high-level products. Three 60 m bands are used for detection and correction of atmospheric effects. The swath width is 290 km.

The SAR data used in this work originates from the Copernicus Sentinel-1 mission (Torres et al., 2012). The C-band radar instrument (5.4 GHz center frequency) on board of the two constellation satellites can operate in various modes depending on the position of the satellite and the scope of the observations. The main operational mode, called Interferometric Wide Swath (IW), is used over land surfaces and features a swath of 250 km and a resolution of 5 m in range and 20 m in azimuth direction. The combined revisit time is of 6 days. The Sentinel-1 mission was designed to provide data in all weather situations for maritime and land monitoring, emergency response, climate change and security.

2.2. SEN12MS-CR Dataset

The dataset presented and used in this work, called SEN12MS-CR, is an evolution of the SEN12MS dataset (Schmitt et al., 2019b). SEN12MS is publicly available and contains triplets of cloud-free Sentinel-2 optical images, Sentinel-1 SAR images and MODIS land cover maps. It was developed for common remote sensing applications, such as scene classification or semantic segmentation for land cover mapping. Using the same procedure as described in the original paper, SEN12MS-CR

was created specifically as a dataset for training deep learning models for cloud removal.

SEN12MS-CR contains 169 non-overlapping regions of interest (ROIs) sampled across all inhabited continents during all meteorological seasons. The scene locations are randomly drawn from two uniform distributions, namely one over all landmasses and one over urban areas only. This introduces a bias towards urban landscapes, that are often in the focus of remote sensing studies and contain more complex patterns. The ROIs have an average size of approx. 5200×4000 px, which corresponds to 52×40 km ground coverage due to the pixels having 10 m ground sampling distance. Each ROI is composed of a triplet of orthorectified, geo-referenced cloudy and cloud-free Sentinel-2 images, as well as the correspondent Sentinel-1 image. All three images were acquired within the same meteorological season to limit surface changes. To assess the cloud coverage of the optical images, the cloud detector described in Schmitt et al. (2019a) was used. The cloud-free Sentinel-2 images have been selected with a threshold of 10% cloud coverage, while cloudy images are within 20% and 70% of cloud coverage.

The Sentinel-2 data is from the Level-1C top-of-atmosphere reflectance product and has values in the range [0, 10,000]. All 13 original bands were included. The Sentinel-1 data is from the Level-1 GRD product acquired in IW mode with two polarization channels (VV and VH). The values are σ_0 backscatter coefficients that have been transformed into dB scale.

To adapt the images for the ingestion into a CNN, the ROIs were cut into small 256×256 px patches with a 128 px stride. The amount of overlap between neighboring patches is therefore 50%. This has been chosen to maximize the number of patches extractable from an image, while still ensuring an acceptable independency. An automated and manual check of the generated patches was performed to eliminate mosaicking artifacts and other corrupted regions. The final quality-controlled SEN12MS-CR dataset contains $157,521$ patches-triplets with a total of 28 layers, amounting to around 620 GB of storage size. Fig. 1 shows examples of patch triplets from the dataset. In the deep-learning based cloud removal algorithms cited in the related works, the networks are trained on datasets with clear limitations. E.g., in Enomoto et al. (2017), Grohnfeldt et al. (2018), Zhang et al. (2018) the networks are trained exclusively on simulated clouds, by using simple Perlin noise or introducing manually gaps into the imagery. In Singh and Komodakis (2018), a dataset of real unpaired cloudy and cloud-free Sentinel-2 images is used, which however is limited to the RGB channels and comprises only 20 cloudy and 13 cloud-free scenes. In Hu et al. (2015), a dataset of ten paired cloudy and cloud-free scenes acquired by Landsat-8 is used. However, the cloud-contaminated images contain only filmy, partly-transparent clouds. To the best of the authors' knowledge, SEN12MS-CR is the first dataset used for training cloud removal networks that comprises a large and representative number of scenes sampled worldwide, with full multispectral information, containing different types of real-life clouds with their characteristic signature in all channels.

2.3. Train, validation and test datasets

To properly assess the generalization capability of a network, a training, validation and test dataset split must be performed. For this, the 169 ROIs of SEN12MS-CR were split into 149 scenes for training, 10 for validation and 10 for testing, following a random global distribution. Fig. 2 shows the spatial distribution of the ROIs. The split according to the ROIs, rather than the patches, ensures that the three datasets are spatially and temporally completely disparate. All three datasets contain acquisitions from all meteorological seasons. A visual and automated analysis confirmed that all three datasets also have a similar distribution of cloud types and coverage amount. When separating the patches according to this split, the training dataset amounts to $134,907$ patches-triplets, the validation to $11,921$ and the test to $10,693$.

3. A ResNet architecture for cloud removal

3.1. ResNet principle

The deep learning model used as backbone for this work is based on the popular ResNet architecture (He et al., 2016a). ResNets make use of shortcut connections, operations that skip some layers to shuttle the information to lower parts of the network, acting as a direct path for information flow. In the original ResNet case, the shortcut connection performs an additive identity mapping, i.e. the input state of a residual block is added to the output of the bypassed layers.

To further understand the residual learning rationale, let $H(\boldsymbol{x})$ be the mapping that the skipped layers are supposed to learn as in a traditional plain network starting from the input $\boldsymbol{x}$. By adding the additive skip connection, we let the layers explicitly learn a residual function $F(\boldsymbol{x})$ instead:

$$F(\boldsymbol{x}) = H(\boldsymbol{x}) - \boldsymbol{x}. \tag{1}$$

This is helpful since it *preconditions* the task: learning a residual correction to the input has proven to be easier for current optimizers than learning the entire input–output mapping from scratch. This is especially true when the optimal mapping for a residual unit is actually close to the identity, i.e. when the network has to just reproduce the input data in the output.

3.2. Residual learning for cloud removal

For the task of cloud removal, the residual skip connections of a ResNet are helpful in several ways:

- Filmy clouds correction: Residual learning offers a clear advantage in the presence of filmy clouds. In this case, the network has to learn only an additive correction that compensates for the thin cloud disturbance in the overcast regions. Through the band concatenation, the network is able to access both the spectral and spatial features; the still partially present ground information acts as a good preconditioning for the restoration process.
- Cloud-free parts reproduction: Due to the large field of view and the comparably small size of clouds, satellite images are typically a mixture of cloudy and cloud-free regions. Over clear-sky regions, the residual connections offer a direct path to transfer unmodified surface information directly to the output.
- Stability of prediction: a ResNet architecture for cloud removal is robust to the presence of large and optically thick clouds in the input data. Even if an input cloudy image is mostly covered by opaque clouds, the network is at least able to reproduce adequately the cloud-free sections. C-GAN based methods (e.g. see Singh and Komodakis, 2018), tend to suffer from prediction instabilities or complete failures with bad input data.
- Optimized learning of deep models: High representational capacity given by a large number of layers and filters in CNNs is required to reconstruct the signal under thick clouds, where complex structures need to be restored. The ResNet architecture allows to optimize large and deep models in a comparably fast way and with good performance (He et al., 2016b).

3.3. DSen2-CR model

The proposed model, called DSen2-CR, is based on the super-resolution Deep Sentinel-2 (DSen-2) ResNet presented in Lanaras et al. (2018), which is itself derived from the state-of-the-art single-image super-resolution EDSR network (Lim et al., 2017). Similarly to super-resolution, cloud removal can be seen as an image reconstruction task, where missing spatial and spectral information has to be integrated into the image to restore the complete information content. To guide the reconstruction process under thick, optically impenetrable clouds

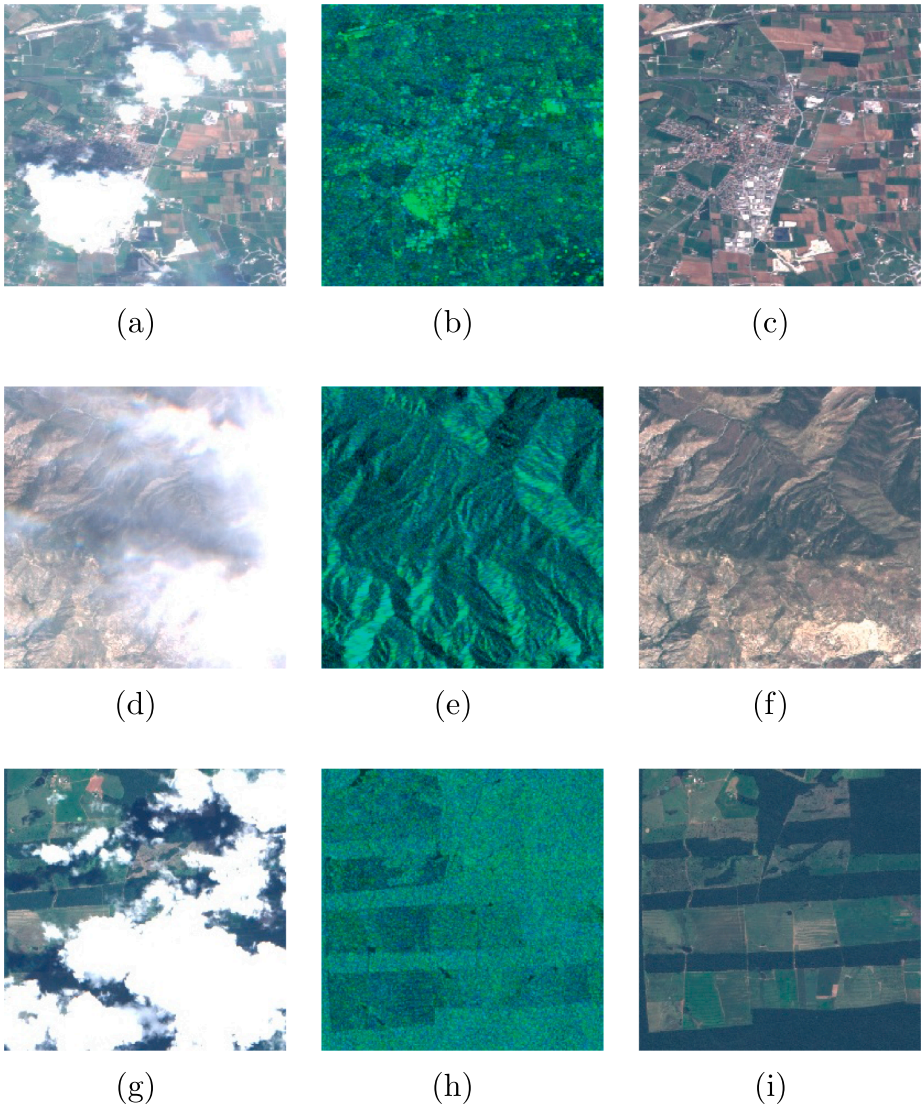


Fig. 1. Example 256 × 256 px patch triplets from the SEN12MS-CR dataset. **(a,d,g)** are the input cloudy optical images, **(b,e,h)** are the input SAR channels, and **(c,f,i)** are the target cloud-free optical images. Throughout the paper, the shown optical images are enhanced true-color RGB composites from the Sentinel-2 10 m resolution B4-B3-B2 bands. The shown SAR images are a composite of the two polarization channels (G = VH, B = VV, R = 0).

where no ground information is available, DSen2-CR leverages a SAR image as a form of prior. For this, a Sentinel-1 image of the same scene is introduced to the network as an additional input. The image's SAR channels are simply concatenated to the other channels of the input optical image. The highly non-linear SAR-to-optical translation, as well as the cloud detection and treatment, are learned and performed implicitly inside the network. The training is done in an end-to-end setup, and a cloud-free image of the same scene is presented to the network as a target for the loss computation. Fig. 3 shows a diagram of the DSen2-CR model and the used residual block design. In the following, further properties and peculiarities of the network are described:

- **Long skip connection:** An additive shortcut shuttles the input cloudy image to an addition layer right before the final output, as originally proposed in Lanaras et al. (2018). This basically means that the entire network is learning to predict a residual map that contains corrections to each pixel of the input cloudy image. In the case of a clear sky input or filmy clouds, the predicted corrections will be minor or non-existent. Conversely, for thick clouds with bright appearance, the corrections will be larger.
- **Residual blocks:** The main part of the network consists of several residual units stacked in sequence. The specific number of units B in the network is a hyperparameter that defines the depth of the network. The residual units each contain four layers and an addition layer for the residual connection. The four skipped layers are a 2D convolution layer with subsequent ReLU activation, a second 2D convolution layer and a final residual scaling layer (see next point). Only one ReLU activation is used after the first convolutional layer but not after the second, since the network is supposed to predict corrections that can be both positive and negative. For both convolutional layers, 3 × 3 kernels are used, following the general community trend to use smaller kernels in deeper models (Lanaras et al., 2018). The output feature dimension F, i.e. the number of different filters, is fixed for all units and is a hyperparameter. A stride of one pixel and zero padding is always used in order to maintain the spatial dimensions of the data throughout the network. Compromising between representational capacity and computational complexity, as well as considering own experiments and the reported experiences in Lanaras et al. (2018), Lim et al. (2017), residual units with $F = 256$ features were selected as a baseline for the DSen2-CR architecture.
- **Residual scaling:** This residual scaling layer is a custom layer that multiplies its inputs with a constant scalar. First proposed in Szegedy et al. (2017), this activations scaling has the effect of

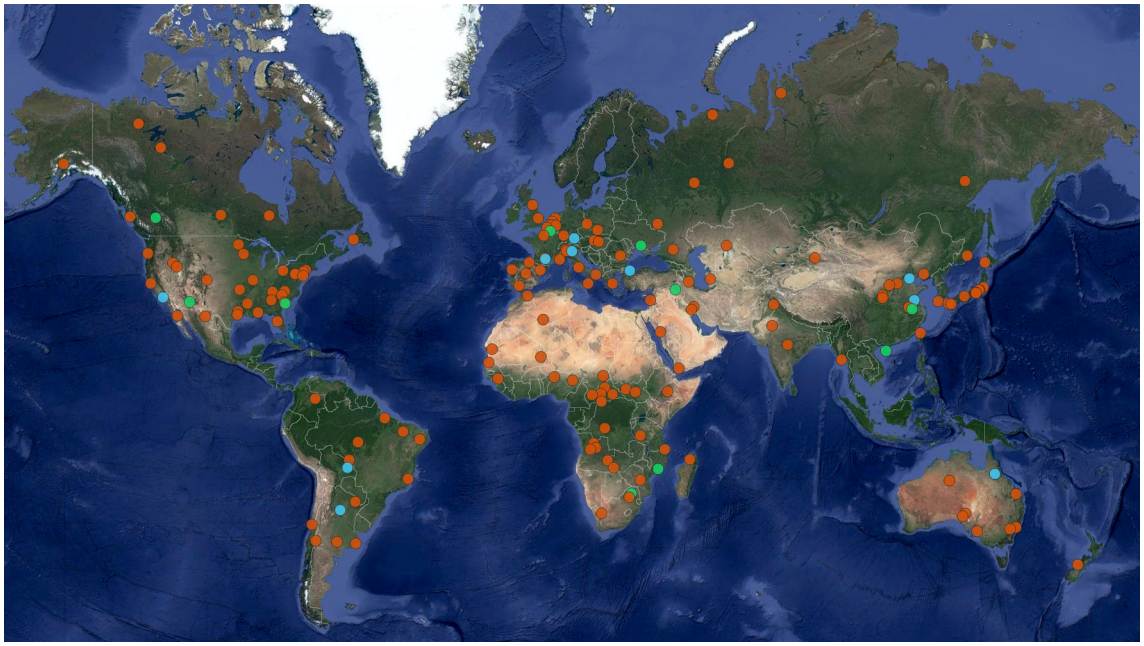


Fig. 2. Global distribution of the 169 ROIs of the SEN12MS-CR dataset. Orange markers denote ROIs selected for training, green for validation and azure for testing. Background image credits: Google Earth/Mapmaker.

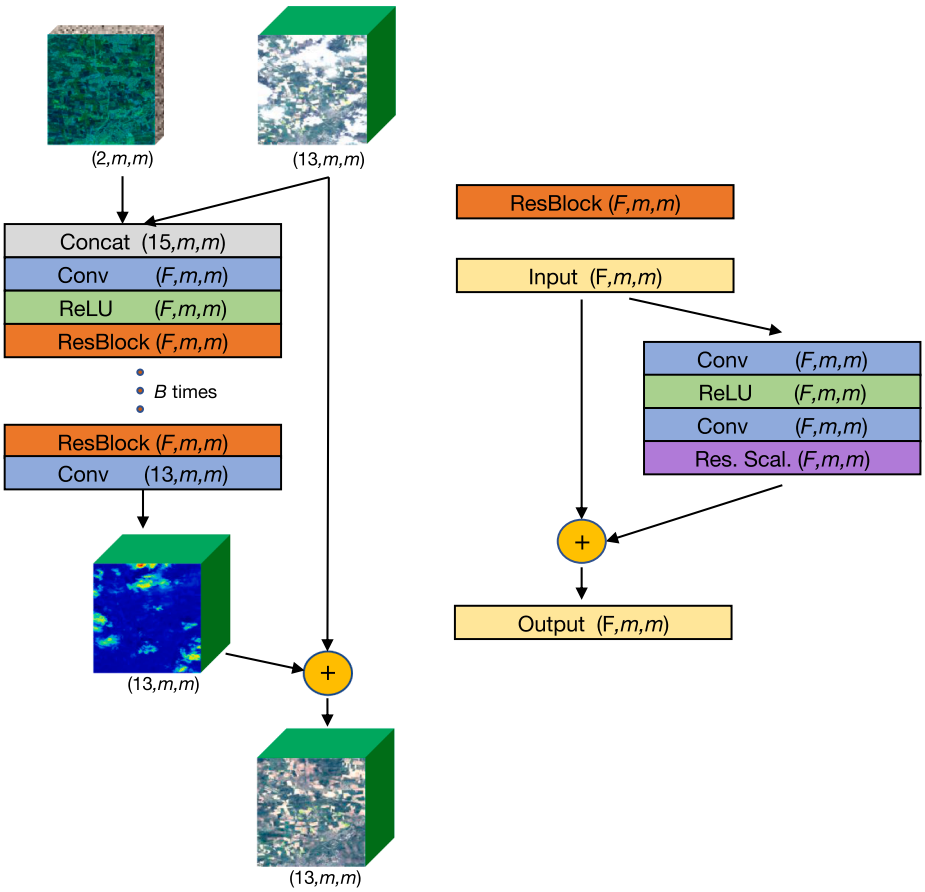


Fig. 3. Left: DSen2-CR model diagram. Right: Residual block design. For each part of the network, the number of layers and the two spatial dimensions are indicated inside parentheses. Since the network is fully convolutional, it can accept input images of arbitrary spatial dimensions m during training and prediction time. F indicates the selected feature dimension and B the selected number of residual blocks included in the network.

stabilizing the training without introducing additional parameters, such as in batch normalization layers. The value of 0.1 is selected for the scaling constant in this work.

- **Additional convolutions:** At the beginning of the network, a concatenation layer stacks vertically the input optical and SAR layers to enable the joint processing. After this, a 3 × 3 convolution layer with ReLU activation is introduced to treat the concatenation before the data is passed through the residual blocks. After the last residual unit, a final 3 × 3 convolution restores the spectral dimensions to match the number of bands of the optical image before reaching the residuals addition layer.

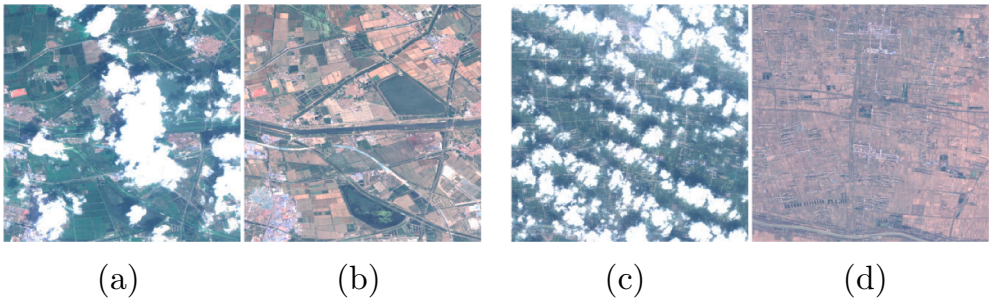


Fig. 4. Example images showing changes in surface conditions between the input cloudy acquisitions **(a,c)** and the target cloud-free images **(b,d)** taken on a different date.

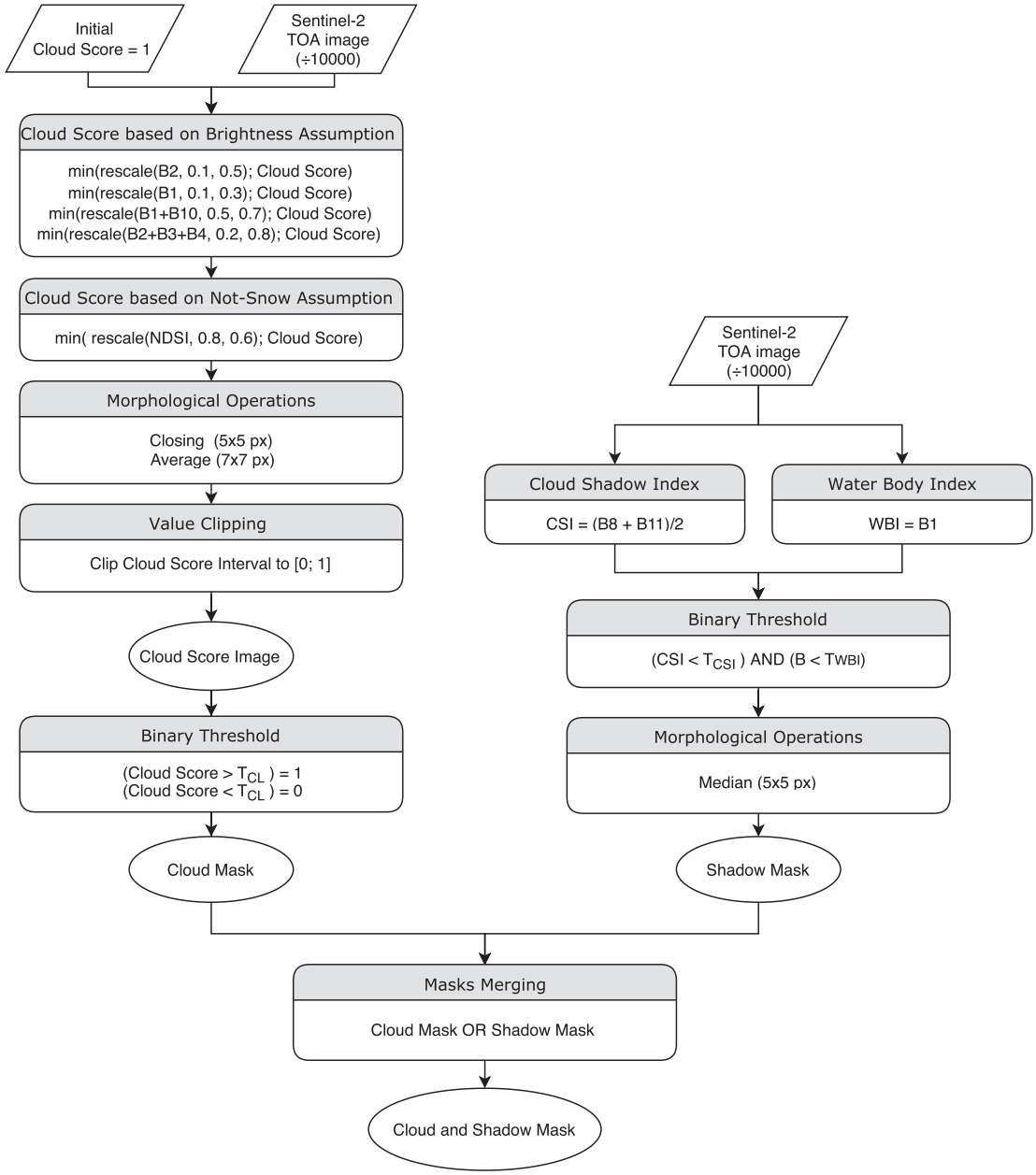


Fig. 5. Flowchart of the cloud (left stream) and shadow (right stream) detectors employed for the mask creation used in the $\mathcal{L}_{\text{CARL}}$ loss.

Several experiments on the network structure and residual block design confirmed the validity and quality of the original DSen-2 architecture. The modifications in DSen2-CR with respect to the original network include the adaptations required to accommodate the two SAR input layers used for guiding the reconstruction, the different number of input and output optical channels, and network depth as described above.

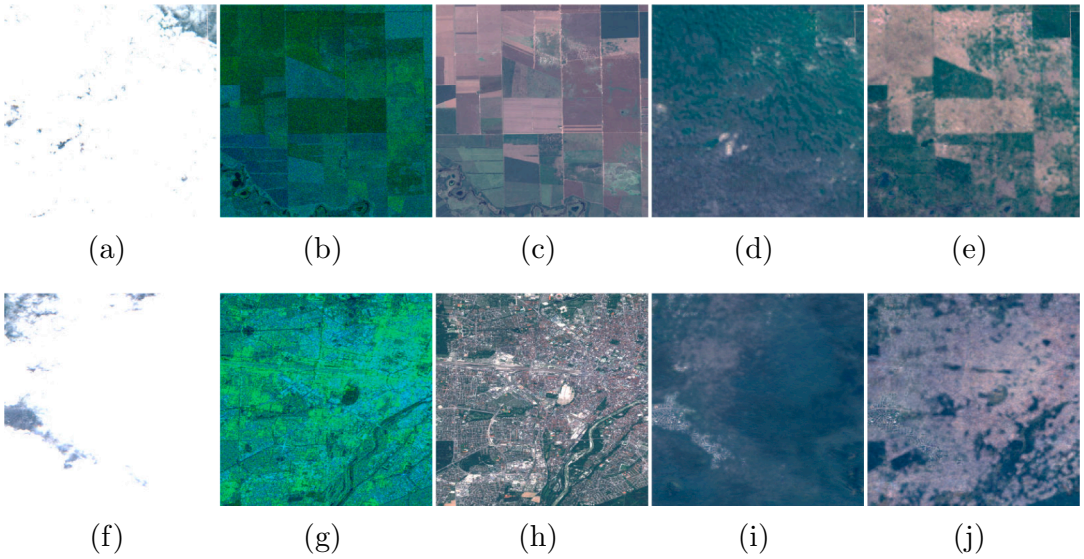


Fig. 6. Example images showing the influence of the SAR input on an agricultural and an urban scene under heavy cloud coverage. **(a,f)** show the cloudy input images, **(b,g)** the input auxiliary SAR images, **(c,h)** the target cloud-free images. **(d,i)** are the model predictions without the SAR input, and **(e,j)** are the predictions of the full DSen2-CR model including the SAR input.

Table 1
Quantitative results computed on the hold-out test dataset. Results are reported for the proposed DSen2-CR network in different configurations: trained on the proposed $\mathcal{L}_{\text{CARL}}$ loss, trained on the plain L1 target loss $\mathcal{L}_{\text{T}}$, and trained on $\mathcal{L}_{\text{CARL}}$ and $\mathcal{L}_{\text{T}}$ but without the SAR input. In the tables, *Target* refers to the error computed between the predicted image and the target cloud-free image. This is the loss as optimized using $\mathcal{L}_{\text{T}}$. *Reprod* denotes the reproduction error, namely the error between the predicted image and the clear parts of the input image. This is part of the $\mathcal{L}_{\text{CARL}}$ loss that is explicitly optimized. *Recon* is the reconstruction error, namely the error between the predicted image and the target image inside the reconstructed clouds and shadow regions.

(a) Test results on pixel-wise metrics

	MAE (ρ_{TOA})			RMSE (ρ_{TOA})	PSNR (dB)
Method	Target	Reprod	Recon	Target	Target
DSen2-CR on $\mathcal{L}_{\text{CARL}}$	0.0290	0.0204	**0.0266**	0.0366	28.7
DSen2-CR on $\mathcal{L}_{\text{T}}$	**0.0270**	0.0398	**0.0266**	**0.0343**	**29.3**
DSen2-CR on $\mathcal{L}_{\text{CARL}}$ w/o SAR	0.0306	**0.0188**	0.0282	0.0387	27.6
DSen2-CR on $\mathcal{L}_{\text{T}}$ w/o SAR	0.0284	0.0389	0.0281	0.0361	28.8
pix2pix	*0.0292*	*0.0210*	*0.0274*	*0.0424*	*28.2*

(b) Test results on spectral and structural fidelity metrics.

	SAM (°)			SSIM
Method	Target	Reprod	Recon	Target
DSen2-CR on $\mathcal{L}_{\text{CARL}}$	8.15	3.94	**8.04**	0.875
DSen2-CR on $\mathcal{L}_{\text{T}}$	**8.07**	6.33	8.13	**0.878**
DSen2-CR on $\mathcal{L}_{\text{CARL}}$ w/o SAR	8.98	**3.86**	8.97	0.870
DSen2-CR on $\mathcal{L}_{\text{T}}$ w/o SAR	8.97	6.17	9.05	0.873
pix2pix	*13.68*	*13.93*	*12.67*	*0.844*

3.4. Cloud-adaptive regularized loss

As described in the dataset section, the input cloudy image and the target cloud-free optical images have been acquired on different days, but within the same meteorological season. Although the time difference is limited, changes in the surface conditions between the images can still often be observed, especially on agricultural landscapes (see Fig. 4). Since the objective of a cloud removal algorithm is to restore ground information below clouds without modifying clear parts, it is of strong importance that the most possible information from the input image is retained in the output. To minimize the influence of ground changes in the target image, a custom training loss was developed in this work.

Following the recommendation of Lanaras et al. (2018), the L_1 metric (mean absolute error) was used as a basic error function due to the robustness to large deviations and the high dynamic range of the Sentinel-2 data. Defining the predicted output image as $\boldsymbol{P}$ and the cloud-free target image as $\boldsymbol{T}$, the classic target loss $\mathcal{L}_{\text{T}}$ based on the simple L_1 distance between prediction and target can be formulated as

$$\mathcal{L}_{\text{T}} = \frac{\|\boldsymbol{P} - \boldsymbol{T}\|_1}{N_{tot}}, \tag{2}$$

with N_{tot} being the total number of pixels in all channels of the optical images. The optimization on this plain L_1 loss is simple and straightforward, but it has a drawback: the network is induced to learn, predict and apply unwanted surface changes, due to being trained on multi-temporal data with changing ground conditions. To reduce these artifacts, a novel loss principle was developed. The idea is to incorporate a binary cloud and cloud-shadow mask (CSM) into the loss computation, and use this information to steer the learning process towards a maximized retainment of input information. This custom loss, which we call Cloud-Adaptive Regularized Loss ($\mathcal{L}_{\text{CARL}}$), is formulated as

$$\mathcal{L}_{\text{CARL}} = \overbrace{\frac{\|\textbf{CSM} \odot (\boldsymbol{P} - \boldsymbol{T}) + (\boldsymbol{1} - \textbf{CSM}) \odot (\boldsymbol{P} - \boldsymbol{I})\|_1}{N_{\text{tot}}}}^{\text{cloud-adaptive part}} + \lambda \overbrace{\frac{\|\boldsymbol{P} - \boldsymbol{T}\|_1}{N_{\text{tot}}}}^{\text{target reg. part}} \tag{3}$$

with $\boldsymbol{P}$, $\boldsymbol{T}$, $\boldsymbol{I}$ denoting respectively the predicted, target, and input optical images. The CSM mask has the same spatial dimensions of the images and pixel values 1 for clouds and shadows pixels or 0 for uncorrupted pixels. $\boldsymbol{1}$ denotes a matrix of ones with the same spatial dimensions as the images and the CSM. The multiplications marked with $\odot$ between the CSM and the image differences are element-wise and applied over all channels. In the cloud-adaptive part, the mean absolute error loss is computed w.r.t. the target image for cloudy or shadowed pixels of the input image, and w.r.t. the input image itself for clear-sky pixels. With this, the network learns that it shall optimize the predictions to match the cloud-free parts of the input, and use the multi-temporal information only when needed, i.e. for the cloud and shadow reconstruction. However, when training with this cloud-adaptive part

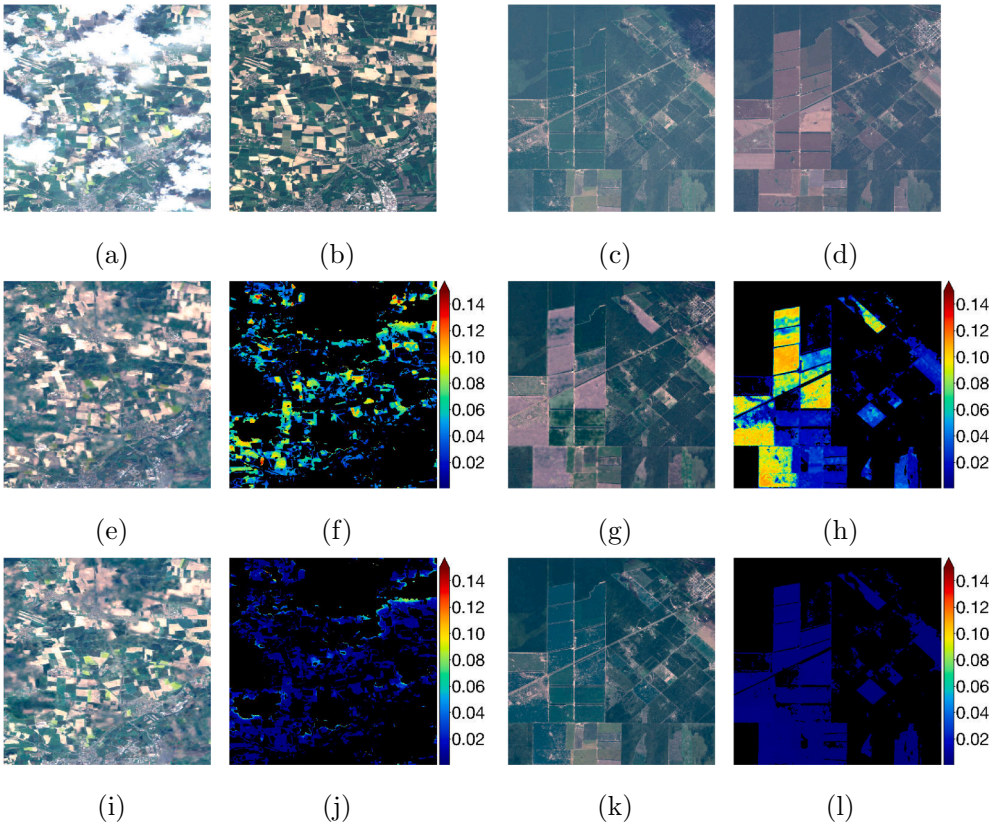


Fig. 7. Example images showing the influence of the $\mathcal{L}_{\mathrm{CARL}}$ loss on two agricultural scenes. **(a,c)** are the input images. **(b,d)** are the target images. **(e,g)** are the predictions obtained by training the DSen2-CR model on the plain $\mathcal{L}_{\mathrm{T}}$, and **(i,k)** are the predictions obtained using $\mathcal{L}_{\mathrm{CARL}}$. **(f,h)** and **(j,l)** are the respective reproduction error maps in units of top-of-atmosphere radiance. The areas within the cloud and cloud-shadow mask (CSM) are depicted in black.

only, it was observed that the network introduced artifacts in the predicted images due to a too precise learning of the mask. To avoid this effect, an additional target regularization term in the form of a classic mean absolute error loss between prediction and target (equivalent to $\mathcal{L}_{\mathrm{T}}$ in Eq. (2)), was added to the loss function. This additional loss induces the network to learn to produce images that still have a natural, smooth appearance similar to the target image. The regularization factor λ, that scales this target regularization term in Eq. (3), is a hyperparameter that effectively balances the input information retainment and the prediction artifacts. After extensive tuning, the value of $\lambda = 1$ was found to provide the best trade-off.

The authors have found that a methodically similar context-aware loss was proposed in Li et al. (2019) in a more generic image processing context. The novelty of the described $\mathcal{L}_{\mathrm{CARL}}$ approach still resides in how a cloud and cloud-shadow mask is created and used in the context of cloud removal, with the specific intent of guiding and improving the reconstruction performance.

For the CSM mask implementation, which is needed during training, a combination of the methods proposed in Schmitt et al. (2019a) (cloud detection) and in Zhai et al. (2018) (cloud-shadow detection) was used. Fig. 5 shows the flowchart of the different processing steps for the mask creation. The threshold $T_{CL} = 0.2$ for the cloud binarization was selected after a visual evaluation. The thresholds for the cloud detection were computed using the parameters $T_{\mathrm{CSI}} = \frac{3}{4}$ and $T_{\mathrm{WBI}} = \frac{5}{6}$. The threshold values were chosen in a conservative manner to reduce false negative detections. We refer to the original papers for further details on the algorithm implementations.

3.5. Preprocessing and training setup

Prior to the ingestion into the network, the images are value-clipped to eliminate small amounts of anomalous pixels. The clipping range for the Sentinel-2 bands is [0, 10,000], for the Sentinel-1 VV and VH polarizations it is [−25,0] and [−32.5,0], respectively. For the Sentinel-2 data, a division by 2000 is further applied to all bands to ensure numerical stability (Lanaras et al., 2018). Similarly, the Sentinel-1 values are shifted into the positive domain and scaled to the range [0, 2] to approximately match the optical data values distribution after scaling. As a data augmentation step, random rotations and flips are applied to the images before the ingestion.

The training framework has been implemented in the Keras open source deep-learning Python library with Tensorflow (Abadi et al., 2016) as backend, basing on the code from (Lanaras et al., 2018). The models were trained on a NVIDIA DGX-1 machine containing 8 P100 GPUs.

The weights of the network have been initialized using a uniform He distribution (He et al., 2015), and the biases were initialized to zero. Several tests with common optimizers showed that the Adam algorithm with integrated Nesterov momentum (Dozat, 2015) delivers the best performance. After a systematic search, the optimal learning rate has been found to be $7{\cdot}10^{-5}$ for a batch size of 16.

4. Experiments & results

For a quantitative evaluation, we report the error metrics obtained by evaluating the results from the entire hold-out test dataset on different network configurations in the following. The used metrics are the mean absolute error (MAE) and the root-mean-square error (RMSE) in units of top-of-atmosphere reflectance ρ_{TOA}, the peak signal-to-noise ratio (PSNR) in decibel units, the spectral angle mapper (SAM) (Kruse et al., 1993) in degrees, and the unitless structural similarity index (SSIM) (Wang et al., 2004). The MAE, RMSE, and PSNR are popular evaluation metrics for pixel-wise reconstruction quality. The SAM gives a measure of the spectral fidelity of the reconstructed images, while the

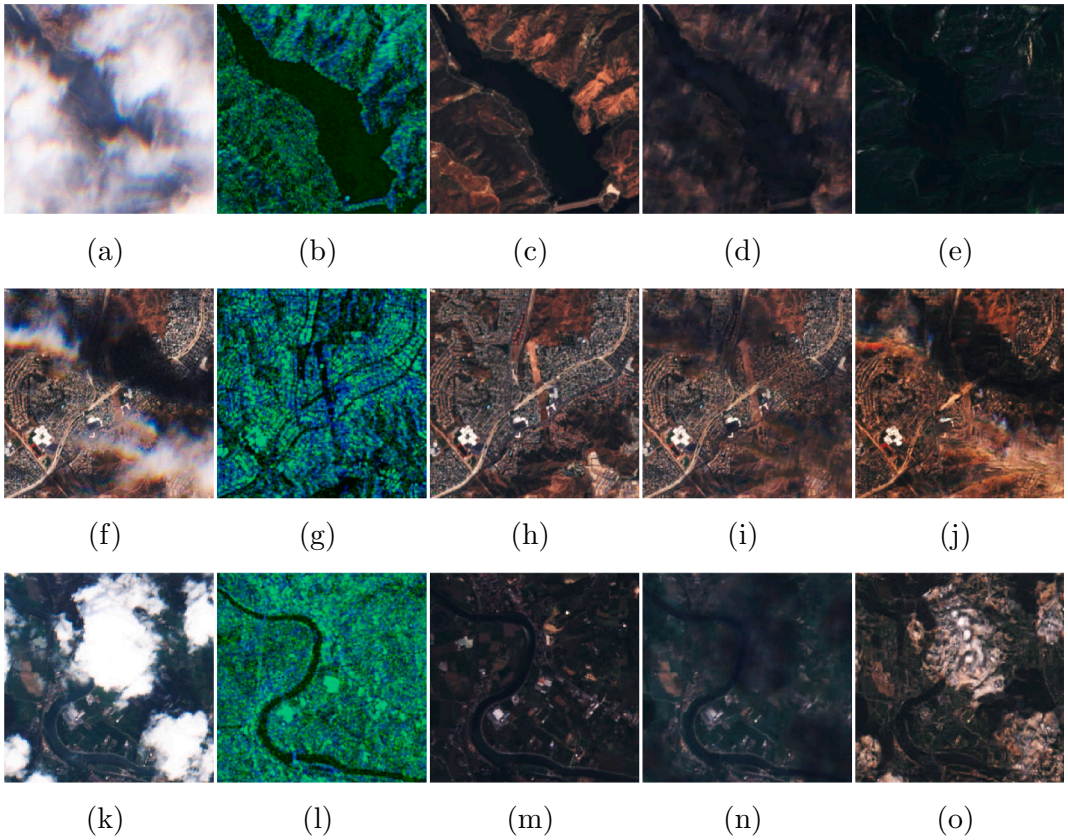


Fig. 8. Example images comparing the cloud removal results of our model with the pix2pix baseline network, both models receiving cloudy optical and SAR data as input. **(a,f,k)** show the cloudy input images, **(b,g,l)** the input auxiliary SAR images, **(c,h,m)** the target cloud-free images. **(d,i,n)** are the predictions of our DSen2-CR model, and **(e,j,o)** are the predictions of the pix2pix baseline. The results show that our model achieves higher-fidelity results, removes cloud shadows better and is less prone to artifacts.

SSIM assesses spatial structure quality based on visual perception principles.

4.1. Influence of SAR-optical data fusion

Several experiments were dedicated to verify the usefulness of the SAR-optical data fusion setup used in DSen2-CR. For this, we performed a full network training with and without including the SAR auxiliary input. In Fig. 6, example results obtained on the hold-out test dataset are visually compared. For better comparability, both networks were trained using the plain L_1 loss $\mathcal{L}_{\mathrm{T}}$. It can clearly be seen that the results which make use of SAR-optical data fusion contain much more structure than the results relying on pure optical-to-optical image translation.

Especially large structures that have regular shapes and a distinctive appearance in the SAR image, e.g. the large fields in the agricultural example scene, are correctly included in the predicted image. Complex objects, e.g. in cityscapes, are harder to integrate due to their more complicated patterns. Here, the model is able to reconstruct the scene only on a coarse scale. For example, the urban example area, with the core town and the river entering from the south, is at least roughly recognizable in the predicted image generated using the SAR information, whereas it is not reconstructed at all if no SAR data is used.

Considerations about the effectiveness of the SAR input can also be made by evaluating the test results reported in Table 1. Here, results from experimental training runs without SAR are provided alongside the full configurations. Comparing the numbers, the network with the SAR input scores better results for most evaluated metrics. Interestingly, however, the networks without SAR achieve lower MAE and SAM *reproduction* errors. This indicates that the network partly integrates SAR information also when reproducing cloud-free regions of the input image. Since such artifacts do not have a correspondence in the original optical image, this leads to a higher reproduction error (for MAE and SAM respectively 2% and 3% using $\mathcal{L}_{\mathrm{T}}$, and 9% and 2% using $\mathcal{L}_{\mathrm{CARL}}$). However, the benefit in terms of *reconstruction* error (approx. 6% for MAE and 11% for SAM for both losses) outbalances this problem, making the SAR-optical data fusion concept beneficial for the overall cloud removal task.

This becomes also clear by a qualitative analysis of the produced images. In Fig. 6, exemplary detail patches under thick cloud cover are presented. By comparing the predicted images with and without SAR prior, the gain in structural content provided by the SAR fusion is clear.

4.2. Influence of the cloud-adaptive regularized loss

One of the main contributions of this work is the design of the so-called cloud-adaptive regularized loss $\mathcal{L}_{\mathrm{CARL}}$. This custom loss is cloud- and shadow-aware and introduces an optimization w.r.t. to the *input* image, in order to retain the most possible amount of information from the uncorrupted input regions. To assess the effectiveness of this proposed loss, we compare the predictions of DSen2-CR models trained on $\mathcal{L}_{\mathrm{CARL}}$ to models trained only on the plain $\mathcal{L}_{\mathrm{T}}$. Fig. 7 shows example images from the test dataset containing two different agricultural landscapes subject to substantial surface changes between the input and the target images. By comparing the RGB composites of the results obtained using $\mathcal{L}_{\mathrm{T}}$ and $\mathcal{L}_{\mathrm{CARL}}$ with the input and the target images, it becomes clear how the network optimized on $\mathcal{L}_{\mathrm{CARL}}$ is able to optimally retain input information and limit the artifact generation in the predicted images. In the left image series, for example, the blooming rapeseed fields captured in the input image are kept in a bright yellow

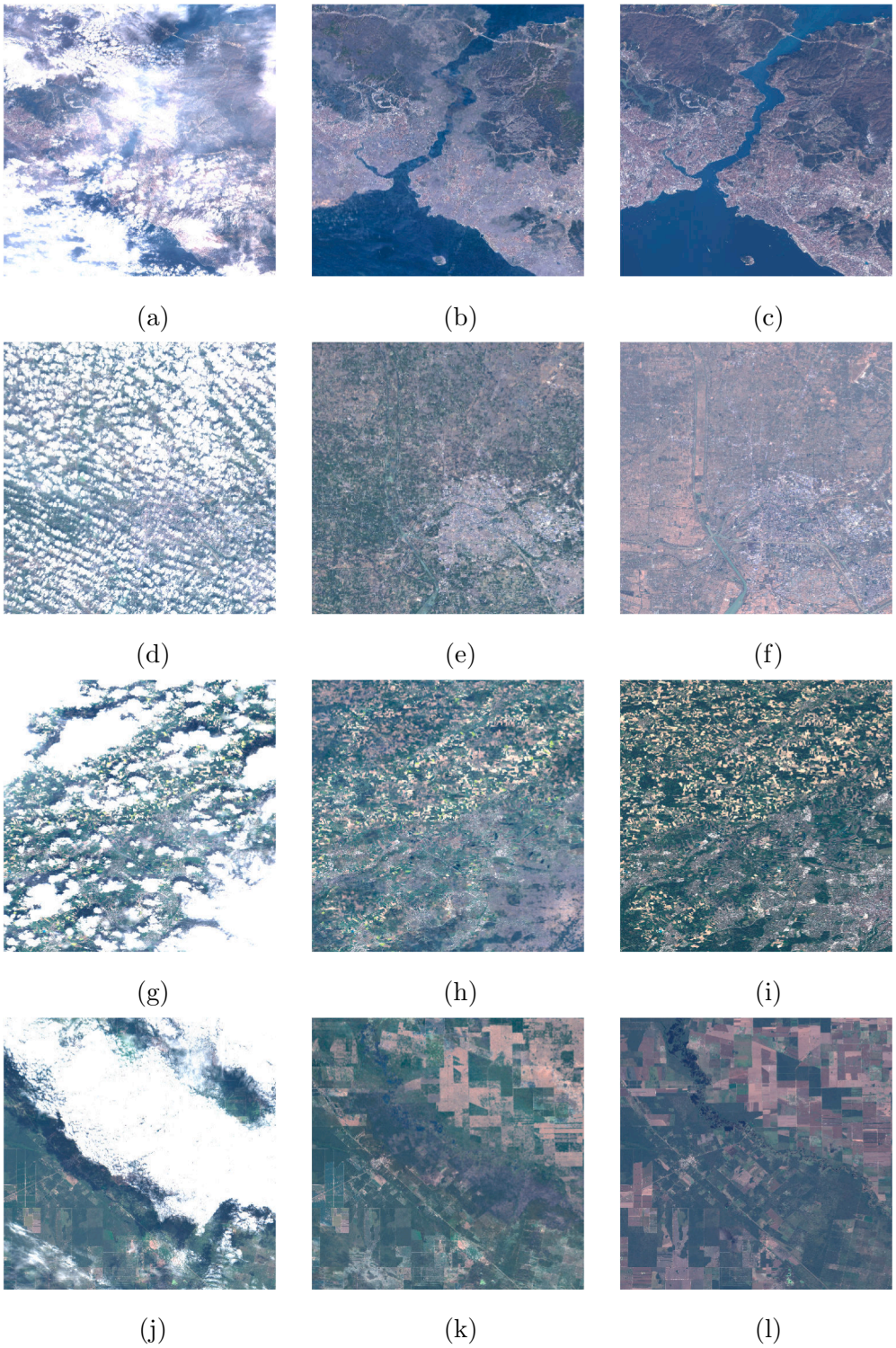


Fig. 9. Example results from the final setup of DSen2-CR using the $\mathcal{L}_{\text{CARL}}$ loss. **(a,d,g,j)** are the input cloudy images, **(b,e,h,k)** the predicted images, and **(c,f,i,l)** the target images.

color by the $\mathcal{L}_{\text{CARL}}$, while being changed to green by $\mathcal{L}_{\text{T}}$.

The shown error maps are the pixel-wise mean absolute error between the predicted image and the cloud-free parts of the input image. In the following, we call this measure *reproduction error*, i.e. the error introduced by the network while reproducing the already cloud-free parts of the input image into the prediction. A low reproduction error indicates an optimal retainment of useful input information. Moreover, it signifies a low artifact generation caused by the training on multi-temporal images with differing ground conditions.

Observing the reproduction error maps shown in the figure, the influence of the adaptive loss is evident, with predictions from $\mathcal{L}_{\text{T}}$ showing much higher reproduction errors in the clear-sky pixels. An evaluation of the final test results in Table 1 shows that model trained on the $\mathcal{L}_{\text{CARL}}$ loss achieves 49% less MAE reproduction error and 38% less SAM reproduction error w.r.t. to the network optimized on $\mathcal{L}_{\text{T}}$. The reconstruction errors between the two models are comparable, showing

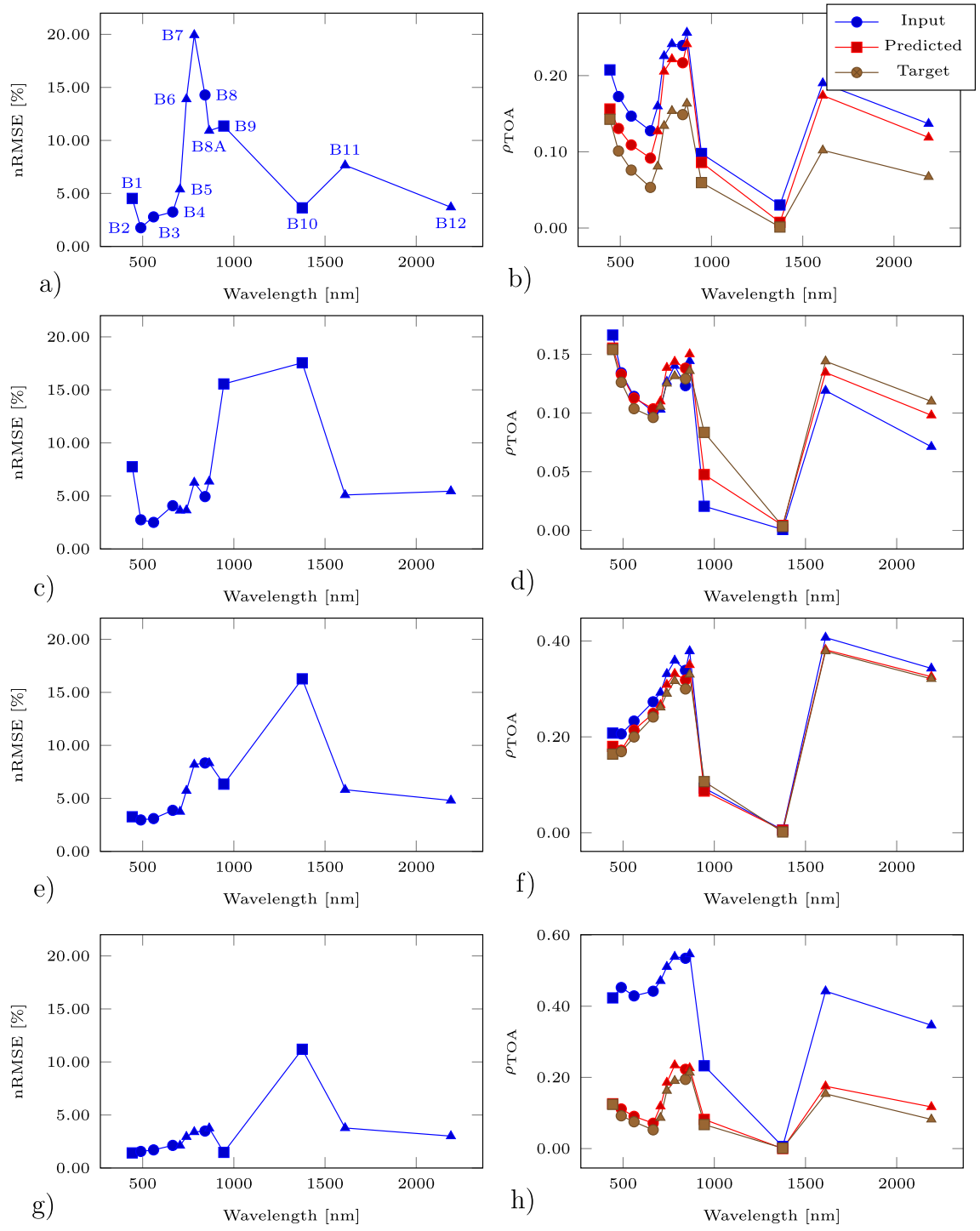


Fig. 10. Left column: channel-wise normalized root-mean-square error (nRMSE) in units of percentage for each image shown in Fig. 9. The normalization was performed using the value range of each band. Right column: Pixel spectra of the central pixel in the respective input, predicted, and target images. The point markers denote the band resolution: circles for 10 m, triangles for 20 m, and squares for 60 m resolution. **(a)** additionally contains labels for each band following the Sentinel-2 bands na.ming convention.

that $\mathcal{L}_{\text{CARL}}$ does not affect negatively the cloud reconstruction performance of the network while optimizing the information retainment capabilities. Considering these observations, we conclude that the usage of $\mathcal{L}_{\text{CARL}}$ in the optimization process is beneficial for the cloud removal task. This is particularly true for agricultural areas, which exhibit phenological changes even within the limited time span lying between the acquisition of the cloud-affected image and the acquisition of the cloud-free target image. It may be noted, however, that using $\mathcal{L}_{\text{T}}$ naturally leads to better results in target-only based metrics (here RMSE, PSNR and SSIM) since the optimization and the evaluation is performed on the same objective. This however does not necessarily signify an improvement in the overall cloud removal performance, due to the artifact generation in cloud-free part as discussed above.

4.3. Comparison against baseline model

In order to compare our model against a standard baseline, we utilized the popular pix2pix architecture (Isola et al., 2017) that was as

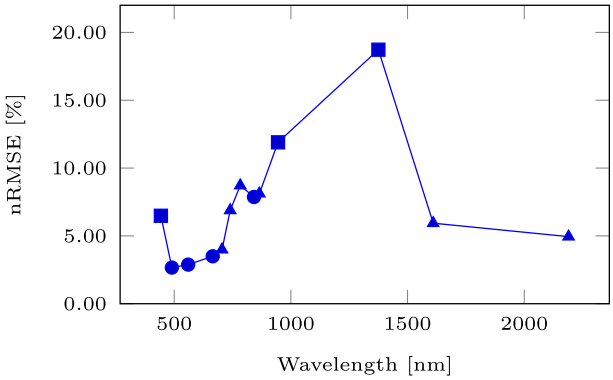


Fig. 11. Average of channel-wise nRMSE over all test images.

well adapted in previous studies on cloud removal (Grohnfeldt et al., 2018; Bermudez et al., 2018). The architecture of our baseline consists of a U-net (Ronneberger et al., 2015) generator and a PatchGAN discriminator (Karacan et al., 2016). The generator takes 13 channel multi-spectral optical and dual-polarimetric SAR patches as input, both of size 256 × 256 pixels. The discriminator takes as input a concatenation of dual-polarimetric SAR patches, the 13-channel multi-spectral cloudy and the real or generated cloud-free patches. SAR patches are clipped to values $[-25, 0]$ and rescaled to range $[-1, 1]$. Optical patches are clipped to values [0, 10,000] and rescaled to range $[-1, 1]$. The network weights are initialized with a Normal initialization and biases are set to zero, The network is trained on the complete training set via ADAM (Karacan et al., 2016) (momentum 0.5) for a total of 10 epochs with the original GAN loss (Isola et al., 2017) and an L_1 loss, weighted with $\lambda_{L_{GAN}}, \lambda_{L_1} = 1, 100$ as in the original study (Goodfellow et al., 2014). Batch normalization (Ioffe and Szegedy, 2015) is applied to the generator. The initial $Niter_{init} = 5$ epochs are trained at a learning rate of $lr_{init} = 2{\cdot}10^{-4}$, followed by $Niter_{decay} = 5$ epochs with lambda learning rate decaying lr_{init} by the multiplicative factor $\lambda_{decay} = 1.0 - max(0, 2 + epoch - Niter_{init})/(Niter_{decay} + 1)$, where *epoch* denotes the number of the current epoch. Both the quantitative results presented in Table 1 and the example images shown in Fig. 8 illustrate the superiority of the our DSen2-CR approach – especially in terms of spectral and structural fidelity.

4.4. Application of the full model on large scenes

For a qualitative evaluation of the operational performance of the full DSen2-CR model trained on the $\mathcal{L}_{\text{CARL}}$ loss including the SAR input, Fig. 9 shows a selection of large reconstructed scenes, i.e. images larger than the 256 × 256-pixel patches the model was trained and validated on. These scenes were concatenated from patches belonging to the hold-out test dataset. To assess the reconstruction performance in all optical channels, Fig. 10 shows the normalized root-mean-square errors (nRMSE) averaged over each optical channel for the pictures shown in Fig. 9. The normalized representation was chosen for better

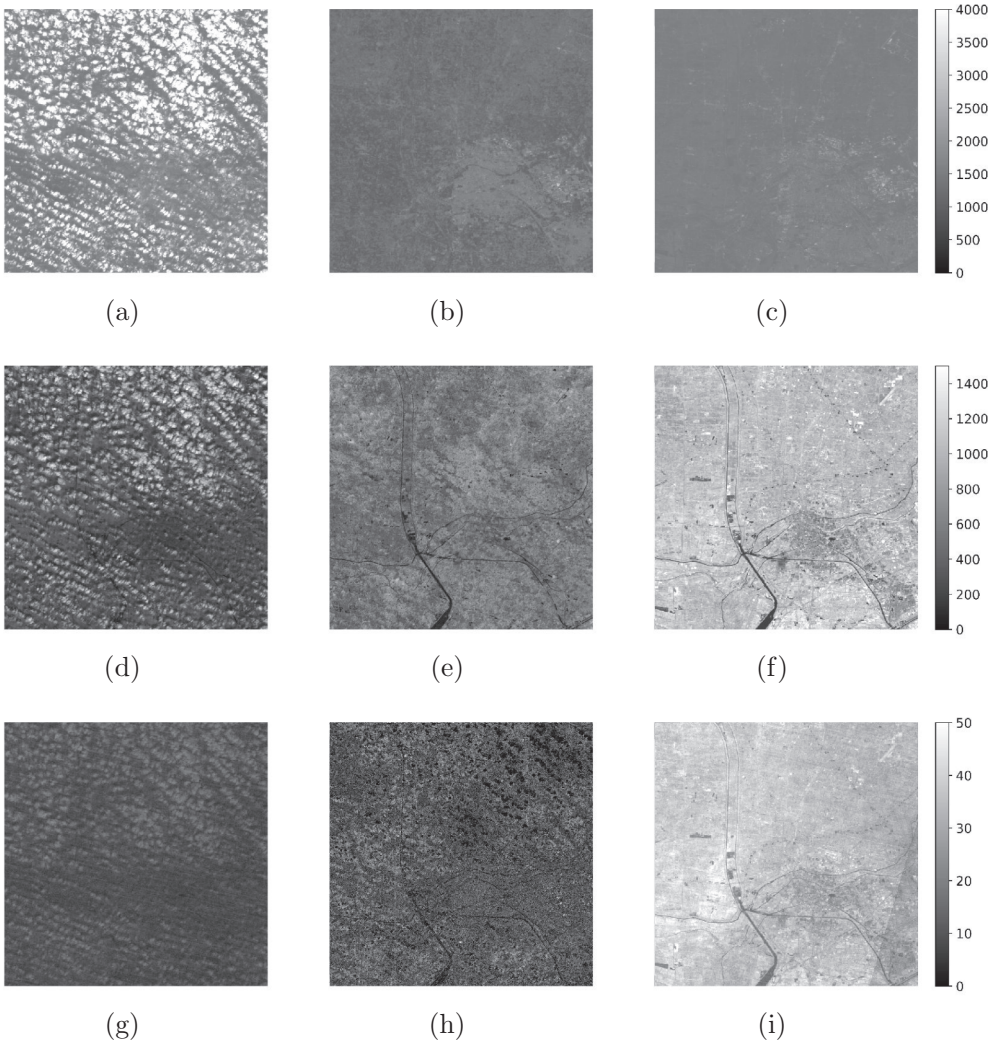


Fig. 12. 60-m resolution channels (B1, B9, B10) for the second image in Fig. 9. Left column: input image. Central column: prediction. Right column: target image.

interpretability, since the absolute RMSE spectra have been observed to correlate with the reflectance spectra. Additionally, in Fig. 10 we also show spectra of the central pixel of each image. To assess the overall band-wise reconstruction quality, averages over all test images of each band-wise normalized RMSE are shown in Fig. 11. It can be seen that the channels, which experience the overall worst reconstruction quality, are B10, followed by B9, and B1 – all of which observe the atmosphere rather than the land surface (see Fig. 12).

Therefore, the performance of the model in reconstructing ground information even below large and thick clouds can still be appreciated on a large scale. The central pixel of the last image (Fig. 10h) is a cloudy pixel, which can be recognized by the high reflectance values of the input image. Here it can be seen how the model successfully reconstructs the entire cloud-free pixel spectrum. For the third image (Fig. 10f) the reconstructed spectrum is also very close to the target, while for the first two images (Figs. 10d and 10b) the reconstruction lies between input and target, either due to prediction inaccuracy or due to the partial retainment of input information induced by the $\mathcal{L}_{\mathrm{CARL}}$ loss.

5. Discussion

As the results summarized in Section 4 show, the DSen2-CR network is generally capable of removing clouds from Sentinel-2 imagery. This is not limited to a purely visual RGB representation of the declouded input image, but includes the reconstruction of the whole pixel spectrum with an average normalized RMSE between 2% and 20%, depending on the band. It should be noted, however, that the worst reconstruction results are achieved for the 60 m-bands, which are not meant to observe the surface of the Earth, but rather the atmosphere: B10, which shows the worst normalized RMSE values, is dedicated to a measurement of Cirrus clouds with a short-wave infrared wavelength; B9 is dedicated to measuring water vapor, and B1 is supposed to deliver information about coastal aerosoles (cf. Fig. 11). Since the SAR auxilary image uses a C-band signal with much longer wavelength, it is not affected by those atmospheric parameters at all and just provides information about the geometrical structure of the Earth surface. This, of course, distorts the reconstruction of the atmosphere-related Sentinel-2 bands, as can be seen in Fig. 11. However, most classical Earth observation tasks, which benefit from a cloud-removal pre-processing step, do not employ those bands anyway and restrict their analyses to the 10 m- and 20 m-bands, which provide actual measurements of the Earth surface. Thus, the inclusion of the SAR auxiliary image can definitely be deemed helpful, which is also confirmed by the numerical results listed in Table 1 and the qualitative examples shown in Fig. 6: The overall best result with respect to pure numbers is achieved when the classic loss $\mathcal{L}_T$ and SAR-optical data fusion are used. The new cloud-adaptive loss $\mathcal{L}_{\mathrm{CARL}}$, however, leads to a much better retainment of the original input and introduces less image translation artifacts, which are usually caused by training on images with a temporal offset. In summary, the combination of SAR-optical data fusion and the cloud-adaptive loss $\mathcal{L}_{\mathrm{CARL}}$ provides the results that generalize best to different situations and also provide reliable cloud-removal for both rather thick clouds and vegetated areas which exhibit phenological changes. In the worst case, i.e. when the scene is comprised of complex patterns and the cloud cover is optically very thick, the network fails to provide a detailed and fully accurate reconstruction (c.f. the urban example in Fig. 6). It has to be stressed again, however, that the dataset used for training of the DSen2-CR model is globally sampled, which means that the network needs to learn a highly complex mapping from SAR to optical imagery for virtually every land cover type existing. By restricting the dataset or fine-tuning the model to a specific region or land cover type, it is expected that the SAR-to-optical translation results would improve significantly.

6. Summary and conclusion

In this paper, we have presented a deep residual neural network for cloud-removal in single-temporal Sentinel-2 satellite imagery. The main features of the proposed approach are threefold: On the one hand, we have incorporated a data fusion strategy to the cloud removal process in order to provide further information about the surface characteristics of the target scene based on Sentinel-1 SAR imagery. On the other hand, we have proposed a cloud-adaptive loss to circumvent the problem that cloud-affected and cloud-free training images can never be acquired at the same time. Finally, we have trained our model on a dataset sampled across the globe and over all meteorological seasons. Based on a deterministic split of training and test data, our experiments confirm the generic applicability of the final cloud-removal model. Both qualitative and quantitative results show that both the SAR-optical data fusion component and the cloud-adaptive training loss help significantly to predict reasonable cloud-free image content. In many cases, the pixel spectra are also improved. Due to the free availability of both Sentinel-2 and Sentinel-1 satellite imagery for all regions of the Earth, it is expected that the presented cloud-removal approach will be beneficial to a more temporally seamless monitoring of our environment.

Declaration of Competing Interest

None.

Acknowledgments

This work was partially supported by the Federal Ministry for Economic Affairs and Energy of Germany in the project “AI4Sentinels – Deep Learning for the Enrichment of Sentinel Satellite Imagery” (FKZ 50EE1910). The work of X. Zhu is jointly supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. [ERC-2016-StG-714087], Acronym: *So2Sat*), Helmholtz Artificial Intelligence Cooperation Unit (HAICU) - Local Unit “Munich Unit @Aeronautics, Space and Transport (MASTr)” and Helmholtz Excellent Professorship “Data Science in Earth Observation - Big Data Fusion for Urban Research”.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, CoRR abs/1603.0.

Bermudez, J.D., Happ, P.N., Oliveira, D.A.B., Feitosa, R.Q., 2018. SAR to optical image synthesis for cloud removal with generative adversarial networks. ISPRS Annals Photogram., Remote Sens. Spatial Inform. Sci., IV-1, 2018, pp. 5–11.

Bermudez, J.D., Happ, P.N., Feitosa, R.Q., Oliveira, D.A.B., 2019. Synthesis of Multispectral Optical Images From SAR/Optical Multitemporal Data Using Conditional Generative Adversarial Networks. IEEE Geosci. Remote Sens. Lett. 16, 1220–1224.

Cheng, Q., Shen, H., Zhang, L., Yuan, Q., Zeng, C., 2014. Cloud removal for remotely sensed images by similar pixel replacement guided with a spatio-temporal MRF model. ISPRS J. Photogram. Remote Sens. 92, 54–68.

Desnos, Y., Borgeaud, M., Doherty, M., Rast, M., Liebig, V., 2014. The European Space Agency’s Earth observation program. IEEE Geosci. Remote Sens. Magaz. 2, 37–46.

Dozat, T., 2015. Incorporating Nesterov momentum into Adam, Technical Report. Stanford University.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services. Remote Sens. Environ. 120, 25–36.

Eckardt, R., Berger, C., Thiel, C., Schmullius, C., 2013. Removal of optically thick clouds from multi-spectral satellite images using multi-frequency SAR data. Remote Sens. 5, 2973–3006.

Enomoto, K., Sakurada, K., Wang, W., Fukui, H., Matsuoka, M., Nakamura, R., Kawaguchi, N., 2017. Filmy cloud removal on satellite imagery with multispectral conditional Generative Adversarial Nets. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), volume 14, IEEE, 2017, pp. 1533–1541.

Fuentes Reyes, M., Auer, S., Merkle, N., Schmitt, M., 2019. SAR-to-optical image translation based on conditional generative adversarial networks – optimization,

opportunities and limits. Remote Sens. 11, 2067.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in neural information processing systems, pp. 2672–2680.
Grohnfeldt, C., Schmitt, M., Zhu, X., 2018. A conditional Generative Adversarial Network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 Images. In: IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 1726–1729.
He, W., Yokoya, N., 2018. Multi-Temporal Sentinel-1 and -2 Data Fusion for Optical Image Simulation. ISPRS Int. J. Geo-Inform. 7, 389.
He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, CoRR abs/1502.0.
He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.
He, K., Zhang, X., Ren, S., Sun, J., 2016b. Identity Mappings in Deep Residual Networks, CoRR abs/1603.0.
Hu, G., Li, X., Liang, D., 2015. Thin cloud removal from remote sensing images using multidirectional dual tree complex wavelet transform and transfer least square support vector regression. J. Appl. Remote Sens. 9, 095053.
Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167.
Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.
Ji, T.-Y., Yokoya, N., Zhu, X.X., Huang, T.-Z., 2018. Nonlocal tensor completion for multitemporal remotely sensed images' inpainting. IEEE Trans. Geosci. Remote Sens. 56, 3047–3061.
Karacan, L., Akata, Z., Erdem, A., Erdem, E., 2016. Learning to generate images of outdoor scenes from attributes and semantic layouts, arXiv preprint arXiv:1612.00215 (2016).
King, M.D., Platnick, S., Menzel, W.P., Ackerman, S.A., Hubanks, P.A., 2013. Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites. IEEE Trans. Geosci. Remote Sens. 51, 3826–3852.
Kruse, F., Lefkoff, A., Boardman, J., Heidebrecht, K., Shapiro, A., Barloon, P., Goetz, A., 1993. The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. Remote Sens. Environ. 44, 145–163.
Lanaras, C., Bioucas-Dias, J., Galliani, S., Baltsavias, E., Schindler, K., 2018. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. ISPRS J. Photogram. Remote Sens. 146, 305–319.
Li, Xinghua, Shen, Huanfeng, Zhang, Liangpei, Zhang, Hongyan, Yuan, Qiangqiang, Yang, Gang, 2014. Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning. IEEE Trans. Geosci. Remote Sens. 52, 7086–7098.
Li, X., Shen, H., Zhang, L., Li, H., 2015. Sparse-based reconstruction of missing information in remote sensing images from spectral/temporal complementary information. ISPRS J. Photogram. Remote Sens. 106, 1–15.
Li, X., Wang, L., Cheng, Q., Wu, P., Gan, W., Fang, L., 2019. Cloud removal in remote sensing images using nonnegative matrix factorization and error correction. ISPRS J. Photogram. Remote Sens. 148, 103–113.
Li, H., Li, G., Lin, L., Yu, H., Yu, Y., 2019. Context-aware semantic inpainting. IEEE Trans. Cybernet. 49, 4398–4411.
Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M., 2017. Enhanced deep residual networks for single image super-resolution. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE. IEEE, pp. 1132–1140.
Lin, C.-H., Tsai, P.-H., Lai, K.-H., Chen, J.-Y., 2013. Cloud removal from multitemporal satellite images using information cloning. IEEE Trans. Geosci. Remote Sens. 51, 232–241.
Lv, H., Wang, Y., Shen, Y., 2016. An empirical and radiative transfer model based algorithm to remove thin clouds in visible bands. Remote Sens. Environ. 179, 183–195.
Meng, F., Yang, X., Zhou, C., Li, Z., 2017. A sparse dictionary learning-based adaptive patch inpainting method for thick clouds removal from high-spatial resolution remote sensing imagery. Sensors 17, 2130.
Mescheder, L., Geiger, A., Nowozin, S., 2018. Which training methods for GANs do actually converge?, CoRR abs/1801.0.
Mirza, M., Osindero, S., 2014. Conditional Generative Adversarial Nets, CoRR abs/1411.1.
Ramoino, F., Tutunaru, F., Pera, F., Arino, O., 2017. Ten-meter Sentinel-2A cloud-free composite—Southern Africa 2016. Remote Sens. 9, 652.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer, pp. 234–241.
Schmitt, M., Hughes, L.H., Qiu, C., Zhu, X.X., 2019a. Aggregating cloud-free Sentinel-2 images with Google Earth Engine. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, volume IV-2/W7, pp. 145–152.
Schmitt, M., Hughes, L.H., Qiu, C., Zhu, X.X., 2019b. SEN12MS – a curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, volume IV-2/W7, pp. 153–160.
Shen, H., Li, X., Cheng, Q., Zeng, C., Yang, G., Li, H., Zhang, L., 2015. Missing information reconstruction of remote sensing data: a technical review. IEEE Geosci. Remote Sens. Magaz. 3, 61–85.
Singh, P., Komodakis, N., 2018. Cloud-Gan: cloud removal for Sentinel-2 imagery using a cyclic consistent Generative Adversarial Network. In: IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium, pp. 1772–1775.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2017. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17) Inception-v4, pp. 4278–4284.
Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Traver, I.N., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L'Abbate, M., Croci, R., Pietropaolo, A., Huchler, M., Rostan, F., 2012. GMES Sentinel-1 mission. Remote Sens. Environ. 120, 9–24.
Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13, 600–612.
Xu, M., Pickering, M., Plaza, A.J., Jia, X., 2016. Thin cloud removal based on signal transmission principles and spectral mixture analysis. IEEE Trans. Geosci. Remote Sens. 54, 1659–1669.
Xu, M., Jia, X., Pickering, M., Jia, S., 2019. Thin cloud removal from optical remote sensing images using the noise-adjusted principal components transform. ISPRS J. Photogram. Remote Sens. 149, 215–225.
Zhai, H., Zhang, H., Zhang, L., Li, P., 2018. Cloud/shadow detection based on spectral indices for multi/hyperspectral optical remote sensing imagery. ISPRS J. Photogram. Remote Sens. 144, 235–253.
Zhang, Q., Yuan, Q., Zeng, C., Li, X., Wei, Y., 2018. Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network. IEEE Trans. Geosci. Remote Sens. 56, 4274–4288.

A.2 Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery

Reference: P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu. *Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery.* IEEE Transactions on Geoscience and Remote Sensing, 59:5866-5878, 2020.

Multisensor Data Fusion for Cloud Removal in Global and All-Season Sentinel-2 Imagery

Patrick Ebel, *Graduate Student Member, IEEE*, Andrea Meraner, Michael Schmitt, *Senior Member, IEEE*, and Xiao Xiang Zhu, *Senior Member, IEEE*

***Abstract*— The majority of optical observations acquired via spaceborne Earth imagery are affected by clouds. While there is numerous prior work on reconstructing cloud-covered information, previous studies are, oftentimes, confined to narrowly defined regions of interest, raising the question of whether an approach can generalize to a diverse set of observations acquired at variable cloud coverage or in different regions and seasons. We target the challenge of generalization by curating a large novel data set for training new cloud removal approaches and evaluate two recently proposed performance metrics of image quality and diversity. Our data set is the first publically available to contain a global sample of coregistered radar and optical observations, cloudy and cloud-free. Based on the observation that cloud coverage varies widely between clear skies and absolute coverage, we propose a novel model that can deal with either extreme and evaluate its performance on our proposed data set. Finally, we demonstrate the superiority of training models on real over synthetic data, underlining the need for a carefully curated data set of real observations. To facilitate future research, our data set is made available online.**

Manuscript received July 9, 2020; revised August 31, 2020; accepted September 14, 2020. Date of publication October 2, 2020; date of current version June 24, 2021. This work was supported by the Federal Ministry for Economic Affairs and Energy of Germany in the project "AI4Sentinels–Deep Learning for the Enrichment of Sentinel Satellite Imagery" under Grant FKZ50EE1910. The work of Xiao Xiang Zhu was supported by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme (Acronym: *So2Sat*) under Grant ERC-2016-StG-714087, in part by the Helmholtz Association through the Framework of Helmholtz Artificial Intelligence Cooperation Unit (HAICU)—Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)," in part by the Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research," and in part by the German Federal Ministry of Education and Research (BMBF) in the framework of the International Future Ai Lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond." *(Corresponding author: Xiao Xiang Zhu.)*

Patrick Ebel is with the Signal Processing in Earth Observation Group, Technical University of Munich, 80333 Munich, Germany (e-mail: patrick.ebel@tum.de).

Andrea Meraner was with the Signal Processing in Earth Observation Group, Technical University of Munich, 80333 Munich, Germany. He is now with the EUMETSAT European Organisation for the Exploitation of Meteorological Satellites, 64295 Darmstadt, Germany (e-mail: andrea.meraner@eumetsat.int).

Michael Schmitt was with the Signal Processing in Earth Observation Group, Technical University of Munich, 80333 Munich, Germany. He is now with the Department of Geoinformatics, Munich University of Applied Sciences, 80335 Munich, Germany (e-mail: michael.schmitt@hm.edu).

Xiao Xiang Zhu is with Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany, and also with the Signal Processing in Earth Observation Group, Technical University of Munich, 80333 Munich, Germany (e-mail: xiaoxiang.zhu@dlr.de).

Digital Object Identifier 10.1109/TGRS.2020.3024744

***Index Terms*— Cloud removal, data fusion, deep learning, generative adversarial network (GAN), optical imagery, synthetic aperture radar (SAR)-optical.**

I. Introduction

ON AVERAGE about 55% of the Earth's land surface is covered by clouds [1], impacting the aim of missions, such as Copernicus, to reliably provide noise-free observations at a high frequency, a prerequisite for applications relying on temporally seamless monitoring of our environment, such as change detection or monitoring [2]–[5]. The need for cloud-free Earth observations, hence, gave rise to a rapidly growing number of cloud removal methods [6]–[12]. While the aforementioned contributions share the common aim of dehazing and declouding optical imagery, the majority of methods are evaluated on narrowly defined and geospatially distinct regions of interest (ROIs). Not only is this specificity posing challenges for a conclusive comparison of methodology but also, furthermore, may cloud-removal performance on a particular ROI poorly indicate performances on other parts of the globe or at different seasons. Moreover, it would be desirable for a cloud removal method to be equally applicable to all regions on Earth, at any season. This generalizability would allow for large-scale Earth observation without the need for costly redesigning or retraining for each individual scene that a cloud removal method is meant to be applied to.

This concern is sustained by previous analysis demonstrating that landcover statistics differ across continents [13] and cloud-coverage is highly variable depending on meteorological seasonality [1]. A major reason for these issues, which is still remaining open nowadays, is the current lack of available large-scale data sets for both training and testing of modern cloud removal approaches. In this work, we address this issue by curating and releasing a novel large-scale data set for cloud removal containing over 100 000 samples from over 100 ROIs distributed over all continents and meteorological seasons of the globe. Especially, we address the challenge of cloud removal in observations from Copernicus mission's Sentinel-2 (S2) satellites. While optical imagery is affected by bad weather conditions and lack of daylight, sensors based on synthetic aperture radar (SAR) as mounted on Sentinel-1 (S1) satellites are not [14] and, thus, provide a valuable source of complementary information. Recent advances in cloud removal combine multimodal data with deep neural networks recovering the affected areas [6], [7], [12], [15].

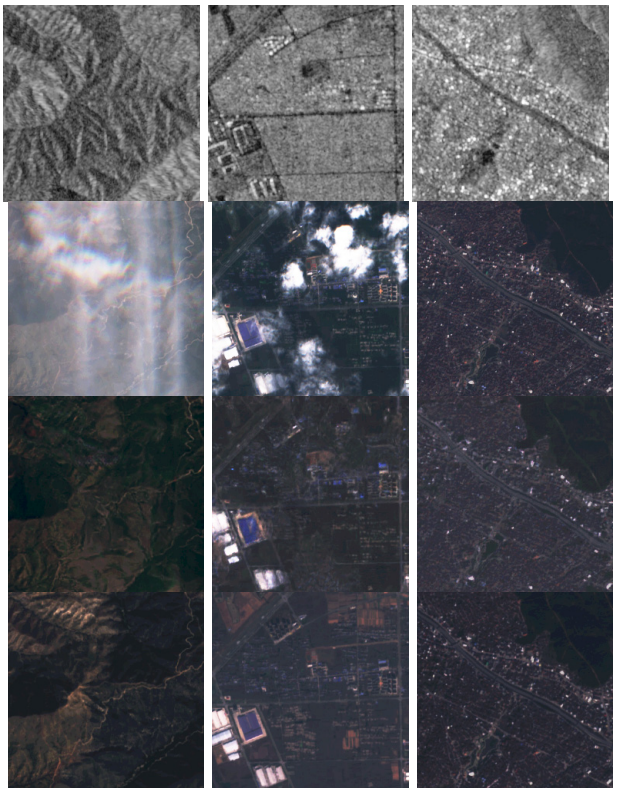


Fig. 1. Exemplary raw data and declouded images. Rows: S1 data (in grayscale), S2 data (in RGB), predicted $\hat{S}2$ data, and cloud-free (target) S2 data. Columns: three different samples. The outcomes show that our model learns to preserve optical data of cloudless areas while replacing cloudy regions by the translation from the SAR domain.

However, many networks are trained on synthetic data or on real data while making strong assumptions on the type and amount of cloud coverage. Moreover, the majority of methods do not explicitly model the amount of cloud coverage and treat each pixel similarly, thereby making unneeded changes to cloud-free areas.

In this work, we address the problem of cloud removal in optical data by means of SAR-optical data fusion, as illustrated in Fig. 1. To redeem the current lack of sufficiently sized and heterogeneous Earth observation data for cloud removal, we release a novel large-scale global data set of coregistered optical cloudy, cloud-free, and SAR observations to train and test the declouding methods. Our data set consists of over 100 000 samples, allowing the training of large models for cloud removal and capturing a diverse range of observations from all continents and meteorological seasons. In addition, we propose a novel generative architecture that reaches competitive performance, as evidenced by two very recently proposed metrics of generated image goodness and diversity. Finally, we show that synthetic data utilized in previous studies are a poor substitute for real cloud coverage data, underpinning the needs for the novel data set proposed in our work.

A. Related Work

The first deep neural architecture to reconstruct cloud-covered images combined near-infrared and red-green-blue (RGB) bandwidth optical imagery by means of a conditional generative adversarial network (GAN) [6], motivated by infrared bandwidth being to a lesser extent impacted by cloud coverage. Subsequent studies replaced the infrared input with SAR observations [7], [15] due to SAR microwaves not being affected by clouds at all [14]. While the early works of [6] and [7] provide a proof-of-concept solely on synthetic data of simulated Perlin noise [16], the networks of [8] and [15] were first to demonstrate performances on real-world data, though focusing primarily on the removal of filmy clouds. Comparable to these studies, we investigate the benefits of SAR-optical data fusion for cloud removal. Unlike the prior work, we address declouding on a carefully curated data set of real imagery sampled over all continents and meteorological seasons, relying neither on synthetic data nor making any strong assumptions about the type and percentage of cloud coverage. Building on the previous studies, the models of [8] and [17] replace the conditional GAN by a cycle-consistent architecture [18], relaxing the preceding models' requirements for pixelwise corresponding training data pairs. While [8] relies solely on cloudy optical input data at inference time, only SAR observations are utilized in [17]. Similar to these two networks, the model that we propose uses a cycle-consistent GAN architecture. We combine cloudy optical with SAR observations and extend on the previous models by incorporating a focus on local reconstruction of cloud-covered areas. This is in line with very recent work [12], [19] that proposed an auxiliary loss term to encourage the model reconstructing information of cloud-covered areas in particular. The network of [12] is noteworthy for two reasons: first, for departing from the previous generative architectures by using a residual network (ResNet) [20] trained supervisedly on a globally sampled data set of paired data; second, for adding a term to the local reconstruction loss that explicitly penalizes the model for modifying off-cloud pixels. Comparable to [12], our network explicitly models cloud coverage and minimizes changes to cloud-free areas. Unlike the model of [12], our architecture follows that of cycle-consistent GAN and has the advantage of not requiring pixelwise correspondences between cloudy and noncloudy optical training data, thereby also allowing for training or fine-tuning on data where such a requirement may not be met. Complementary to the SAR-optical data fusion approach to cloud removal, recent contributions proposed integrating information of repeated observations over time [10], [11]. The work indicates promising results but trades temporal resolution for obtaining a single cloud-free observation, whereas our approach predicts one cloud-free output per cloudy input image and, thus, allows for sequence-to-sequence translation. Moreover, current multitemporal approaches make strong assumptions about the maximum permissible amount of cloud-coverage affecting individual images in the input time series, which is required to be no more than 25% or 50% of cloud coverage for the method of [10] and 10%–30% in the work of [11]. Our curated data sets evidence that such strict requirements on the percentage of cloudiness may, oftentimes, not be met in practice. Consequently, our model makes no assumptions on the maximum amount of tolerable cloud coverage per observation and can gracefully deal with

samples ranging from cloud-free to widely obscured skies due to minimizing changes to cloud-free pixels and using SAR observations unaffected by clouds.

II. Methods

We propose a novel model to recover cloud-occluded information in optical imagery. Our network explicitly processes a continuous-valued mask of cloud coverage computed on the fly, as described in Section II-A, to preserve cloud-free pixels while making data-driven adjustments to cloudy areas. The continues-valued assignment of each pixel in the processed cloud mask can be interpreted as the likelihood of the pixel being cloud-covered according to the cloud detector algorithm of [21]. Our model explicitly processing cloud coverage information is in contrast to previous generative architectures that are agnostic to cloud-coverage [6], [8] and networks that only utilize binary cloud mask information [12] as opposed to more fine-grained continuous-valued masks proposed in this work. A cycle-consistent generative architecture detailed in Section II-B allows for training without the need for coregistered cloudy and noncloudy observations of strict pixelwise one-to-one correspondences compared with earlier approaches that required strict pixelwise alignments [7], [15]. We adapt the architecture to integrate SAR with optical observations and propose a new auxiliary cloud map regression loss that enforces sparse reconstructions to minimize modification on cloud-free areas, as described in Section II-C.

A. Cloud Detection and Mask Computation

To evaluate the cloud coverage statistics of our collected data set and model cloud coverage explicitly while reconstructing cloud-covered information, we compute cloud probability masks m. The masks m are computed online for each cloudy optical image and contain continuous pixel values within [0, 1], indicating, for a given pixel, its probability of being cloud-covered. We compute m via the classifier s2cloudless of [21], which demonstrated cloud detection accuracies on par with the multitemporal classifier MAJA [22], running on single-shot observations. While s2cloudless originally applies classification to compute a sparsified binary cloud mask, we wish to obtain a continuous-valued cloud map. We, therefore, take the intermediate continuous-valued representation of the pipeline of [21], then apply a high-pass filter to only keep values above 0.5 intensity, and, finally, convolve with a Gaussian kernel of width $\sigma = 2$ to get a smoothed cloud map with pixel values in [0, 1]. We note that m may alternatively be computed by a dedicated deep neural network [23], but our solution is lightweight and, thus, perfect to support methods running on very large data sets, at almost no additional computational cost in either memory or run time. Exemplary samples of cloud probability masks are presented in Appendix A.

B. Architecture

The model proposed in this work follows the architecture of cycle-consistent GAN [18], i.e., we use two generative networks $G_{S1\to S2}$ and $G_{S2\to S1}$ that translate images from the source domain of $S1$ to the target domain of $S2$, and vice versa. Distribution $\dot{S1}$ (or $\dot{S2}$) denotes the target when the generator performs a within-domain identity mapping, preserving the input image's sensor characteristics. For each domain, there exists an associated discriminator network, denoted as D_{S1} and D_{S2}, respectively, classifying whether a given image is a sample from the domain's true distribution S1 (or S2) or from the synthesized distribution $\hat{S1}$ (or $\hat{S2}$). An overview of our model ensemble is given in Fig. 2. While we keep the network $G_{S2\to S1}$ as in the original work, we apply spectral normalization [24] to both discriminators and make adjustments as follows: $G_{S1\to S2}$ receives an image from domain $S1$ as input and is additionally conditioned on the corresponding cloudy image from $S2$, as well as the cloud probability mask m. For our cloud-removal network, we keep the encoder–decoder architecture of the generator but add a long-skip connection [20] such that the output is given by

$$\hat{S2} = G_{S1\to S2}(\cdot) = \tanh(S2 + S2_{\text{res}})$$

where $S2_{\text{res}}$ denotes the residual mapping learned by the generator. To demodulate the effects of the output nonlinearity on the long-skipped pixels, the inverse hyperbolic tangent is applied to the cloudy input image from $S2$ before the residual mapping. Furthermore, we insert a regression layer taking the residual maps $S2_{\text{res}}$ as input and returning a prediction $\hat{m}$ of the cloud map m. The purpose of the regressor is to enforce a meaningful relation between the learned $S2_{\text{res}}$ and the conditioning m, making the residual maps sparse. Here, sparseness refers to the residual maps being (close to) zero over noncloudy areas, as opposed to having widespread small values, which would indicate many unneeded changes made to cloud-free pixels. We enforce sparseness of the residual maps by formulating an L1 loss on the cloud mask regression, as defined in Section II-A. The loss term effectively acts as a regularizer on changes made to noncloudy areas, penalizing unnecessary adjustments. The regression layer consists of a [3 × 3] convolutional kernel mapping the generated 3-D image to a single-channel map and, thus, adds little to the overall number of learnable parameters. The architecture of generator $G_{S1\to S2}$ is depicted in Fig. 3, and the details on its parameterization are provided in Table I. Discriminator D_{S2} is well-conditioned on the cloud probability maps m. Importantly, we forward the (unpaired) noncloudy optical images to the discriminator D_{S2}, which learns the noncloudy patchwise statistics and, thus, implicitly forces $G_{S1\to S2}$ to synthesize cloud-free images. In sum, our main contribution with respect to architectural changes is twofold. First, we adjusted the generator predicting cloud-free optical images to learn a residual mapping by introducing a long-skip connection forwarding optical information, removing the previous need to reconstruct (even cloud-free) pixels from scratch. Second, our generator learns to constrain modifications to cloud-covered pixels while keeping clear areas unchanged, which is encouraged by introducing a novel layer regressing the cloud coverage map by the learned residual map.

C. Losses

We adjust the losses such that regions regressed as cloud-free in map m remain untouched, while cloudy areas are

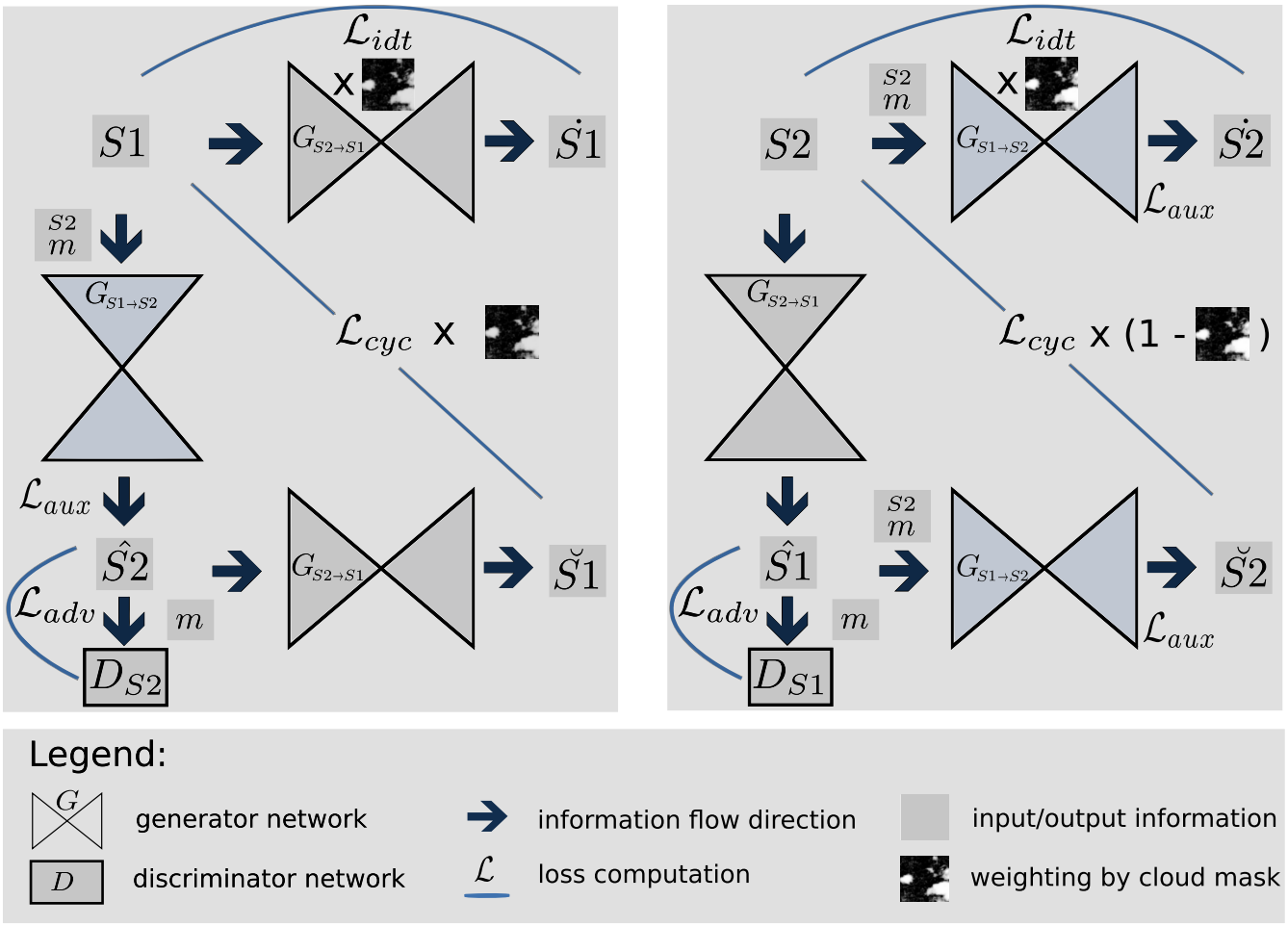


Fig. 2. Overview of our model ensemble based on cycle-consistent GANs [18]. The model consists of two generative networks $G_{S1\to S2}$ and $G_{S2\to S1}$ that translate images from the source domain of $S1$ to the target domain of $S2$, and vice versa. Distribution $\dot{S}1$ (or $\dot{S}2$) denotes the target when the generator performs a within-domain identity mapping, preserving the input image's sensor characteristics. For each domain, there exists an associated discriminator network, denoted as D_{S1} and D_{S2}, respectively, classifying whether a given image is a sample from the domain's true distribution S1 (or S2) or from the synthesized distribution $\hat{S}1$ (or $\hat{S}2$). The network architectures are as in [18]—except for the generator $G_{S1\to S2}$, which is modified as detailed in the main text and in Fig. 3. The losses $\mathcal{L}_{\text{adv}}$, $\mathcal{L}_{\text{cyc}}$, $\mathcal{L}_{\text{idt}}$, and $\mathcal{L}_{\text{aux}}$ are defined in Section II-C.

recovered given the information from domain $S1$. The losses minimized by the generators are

$$\begin{aligned}
\mathcal{L}_{\text{adv}} &= (D_{S1}(\hat{S1}) - 1)^2 + (D_{S2}(\hat{S2}) - 1)^2 \\
\mathcal{L}_{\text{cyc}} &= \|m \cdot (S1 - \breve{S1})\|_1 + \|(1-m) \cdot (S2 - \breve{S2})\|_1 \\
\mathcal{L}_{\text{idt}} &= \|m \cdot (S1 - \dot{S1})\|_1 + \|m \cdot (S2 - \dot{S2})\|_1 \\
\mathcal{L}_{\text{aux}} &= \|(1-m) \cdot (m - \hat{m})\|_1 \\
\mathcal{L}_{\text{all}} &= \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{idt}}\mathcal{L}_{\text{idt}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}
\end{aligned}$$

where $\lambda_{\text{adv}} = 5.0$, $\lambda_{\text{cyc}} = 10.0$, $\lambda_{\text{idt}} = 1.0$, and $\lambda_{\text{aux}} = 10.0$ are the hyperparameters to linearly combine the individual losses within $\mathcal{L}_{\text{all}}$. The loss weightings are set similar to those in [18], with minor adjustments made manually. $\mathcal{L}_{\text{adv}}$ is the adversarial loss originally proposed in LSGAN [25], implementing a least-squares error function on the classifications of the discriminators D_{S1} and D_{S2}. $\mathcal{L}_{\text{cyc}}$ and $\mathcal{L}_{\text{idt}}$ are introduced in [18] but weighted pixelwise with the cloud map m. The purpose of the cycle-consistent loss $\mathcal{L}_{\text{cyc}}$ is to regularizing the mapping $S1 \to S2$ by requiring $S2 \to S1$ being able to reconstruct the original input again (likewise for the direction $S2 \to S1 \to S2$), constraining the potential mappings between both domains. The idea behind $\mathcal{L}_{\text{idt}}$ is to motivate generators to perform an identity mapping and limit unneeded changes in case the provided input is a sample of the target domain. $\mathcal{L}_{\text{aux}}$ is the loss associated with the cloud map regression in $G_{S1\to S2}$, introduced to enforce sparseness of the learned residual feature maps $S2_{\text{res}}$ such that the noncloudy pixels of $S2$ experience little to no adjustments. Our modified generator architecture, the usage of probabilistic cloud maps, and the adjusted losses are showcased in context of a cycle-consistent GAN ensemble, but we remark that they may as well be used within alternative models, such as conditional GAN [26] or ResNet architectures [20].

III. Experiments and Analysis

A. Data

To conduct our experiments, we gather a novel large-scale data set called SEN12MS-CR for cloud removal. For this purpose, we build upon the openly available SEN12MS data set [27] of globally sampled coregistered $S1$ plus cloud-free $S2$ patches and complement the data set with coregistered

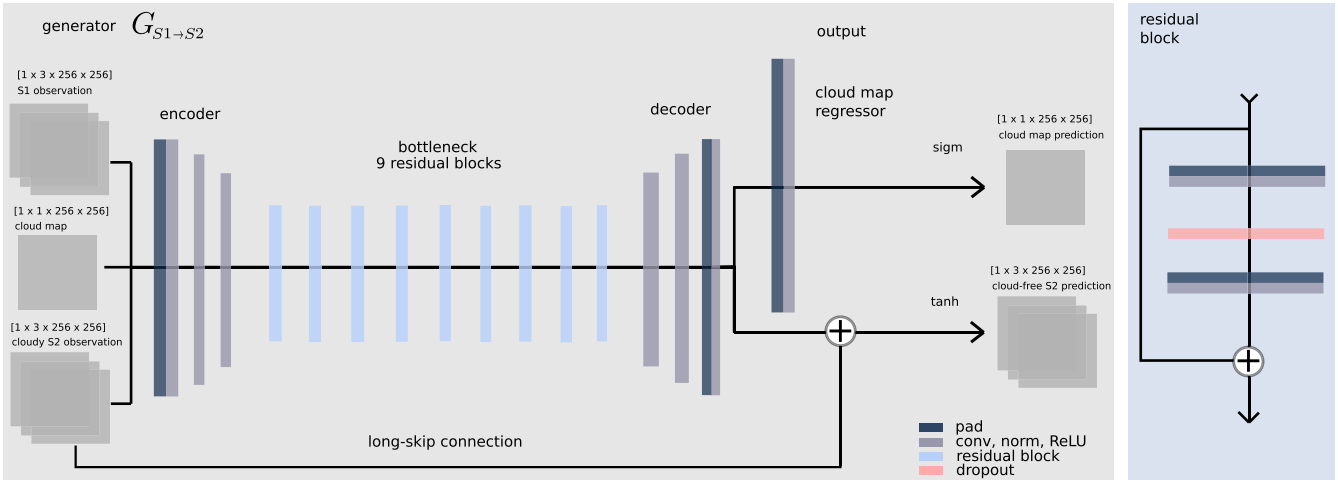


Fig. 3. Detailed architecture of the generator $G_{S1\to S2}$ of Fig. 2. The generator receives $S1$, m, and $S2$ as input, the latter of which is long-skip forwarded and modified by the learned residual map $S2_{\text{res}}$. The result is passed via a nonlinearity as input to the next network, or treated as output. In parallel, $S2_{\text{res}}$ is regressing m to enforce sparseness of the residual map.

TABLE I

ARCHITECTURE OF OUR GENERATOR $G_{S1\to S2}$. THE ARCHITECTURE IS DIVIDED INTO FOUR COMPONENTS, AS ILLUSTRATED IN FIG. 3, AND INFORMATION FLOW IS FROM LEFT TO RIGHT ACROSS COMPONENTS AND TOP TO BOTTOM WITHIN COMPONENTS. SYMBOLS: R (RELU), N (INSTANCE NORMALIZATION), C (CONVOLUTION), AND T (TRANSPOSED CONVOLUTION). FOR (TRANSPOSED) CONVOLUTION, THE PARAMETERIZATION IS (KERNEL HEIGHT × KERNEL WIDTH, NUMBER OF FILTERS, STRIDE, AND PADDING SIZE). THE ARCHITECTURE OF GENERATOR $G_{S2\to S1}$ IS SIMILAR TO THE 9-RESNET BLOCK GENERATOR IN [18], AND THE TWO DISCRIMINATORS ARE KEPT AS THE PATCHGAN DISCRIMINATORS IN [18]

encoder	bottleneck	decoder	output
R(N(C(3 × 3, 64, 1, 1)))	R(N(C(3 × 3, 256, 1, 1)))	R(N(T(3 × 3, 256, 2, 1)))	sigmoid(C(3 × 3, 1, 1, 1))
R(N(C(3 × 3, 128, 2, 1)))	9 × dropout(0.5)	R(N(T(3 × 3, 128, 2, 1)))	tanh(C(3 × 3, 3, 1, 1))
R(N(C(3 × 3, 256, 2, 1)))	R(N(C(3 × 3, 256, 1, 1)))		

cloudy images close in time to the original observations. SEN12MS-CR consists of 169 nonoverlapping ROIs evenly distributed over all continents and meteorological seasons. The ROI has an average size of approximately 52 × 40 km^2 ground coverage, corresponding to complete-scene images of about 5200 × 4000 pixels. Each complete-scene image is checked manually to ensure freedom of noise and artifacts. The cloud-free optical images of four exemplary ROI observed in four different meteorological seasons are depicted in Fig. 4 to highlight the heterogeneity of landcover captured by SEN12MS-CR. Each scene in the data set is subsequently translated into Universal Transverse Mercator coordinate system and then partitioned into patches of size 256 × 256 pixels with a spatial overlap of 50% between neighboring patches, yielding an average of over 700 patches per ROI. Each patch consists of a triplet of orthorectified, georeferenced cloudy, and cloud-free 13-band multispectral Sentinel-2 images, as well as the correspondent Sentinel-1 image (see Fig. 1 for the examples of SAR, cloud-free, and cloudy optical patch triplets). Paired images of the three modalities were acquired within the same meteorological season to limit surface changes. The Sentinel-2 data are from the Level-1C top-of-atmosphere reflectance product. Finally, each patch triples is automatically controlled for potential imaging artifacts, and exclusively, artifact-free patches are preserved to constitute the final cleaned-up version of SEN12MS-CR.

Evaluating the cloudiness of each patch with the algorithm of [21], as described in Section II-A, yields a mean cloud coverage of circa 47.93% ± 36.08%, i.e., about half of all the optical images' information is affected by clouds and the amount of coverage varies considerably. This amount of coverage is notably close to the approximately 55% of global cloud fraction over land that has previously been observed empirically [1]. The distribution of cloud coverage is shown in Fig. 5 and is relatively uniform over the entire domain, with slightly more samples showing (almost) no clouds or being entirely cloud-covered. Note that the computed cloud probability masks are not used to filter any observations or actively guide the data set creation in any manner, and they are solely used *post hoc* to quantify the distribution of cloudiness. For the sake of comparability across models in our experiments and for further studies, we define a train split and a split of hold-out data, which is reserved for the purpose of testing. The train split consists of 114 325 patches sampled uniformly across all continents and seasons and is open to be entirely used for training or in parts for training and validating. The test split consists of 7893 geospatially separated images sampled from ten different ROI distributed across all continents and meteorological seasons, capturing a heterogeneous subset of data.

B. Experiments and Results

A total of three experiments are conducted. First, we train our network and extend it by adding supervised losses for the model to benefit of paired noncloudy and cloudy optical observations in our data set at training time. We systematically

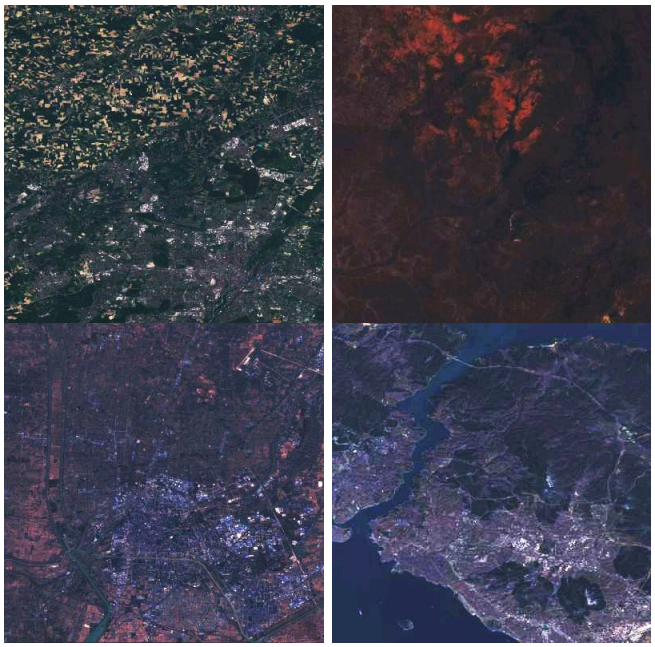


Fig. 4. Cloudless S2 imagery of four exemplary ROI, illustrating the diversity of SEN12MS-CR. The four different scenes are of four different meteorological seasons from the test split of the data set. On average, an ROI is split into over 700 patch samples, each observation of size 256 × 256 pixels.

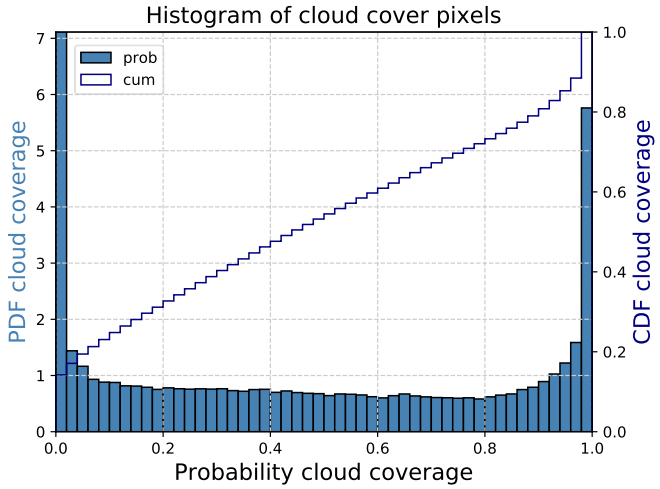


Fig. 5. Statistics of cloud coverage of SEN12MS-CR. On average, approximately 50% of occlusion is observed. The empirical distribution of cloud coverage is relatively uniform and ranges from cloud-free views to total occlusion.

vary the amount of available supervision to investigate its effects on model performance. Second, we evaluate it against a set of baseline models. Third, we retrain the architectures from the previous experiment on synthetic data of generated cloudy observations and evaluate them on real data in order to quantify to which extent models trained on simulated data are capable to generalize to real-world scenarios. To the best of our knowledge, neither of these experiments has previously been conducted in depth. All experiments were conducted on a machine of 8 Intel Core i7-8700 CPU @ 3.20-GHz processors, 16 GB of DIMM DDR4 Synchronous 2667-MHz RAM, and an NVIDIA GeForce RTX 2080, running Ubuntu 18.04. Computation clock time for the training procedure may vary according to the overall task load but is estimated to be about seven days for model ours-0 and about 10–12 days for model ours-100.

1) Metrics to Quantify the Goodness of Cloud Removal: In order to evaluate model performances quantitatively, we utilize the recently developed metrics of improved precision and recall [28], as proposed in the context of generative modeling and improving on previous metrics, such as Inception score or Frechét Inception distance [29], [30]. Improved precision and recall are measures of goodness quantifying similarities between two sets of images in a high-dimensional feature embedding space. Precision is a metric of sample quality, assessing the fraction of generated images that are plausible in the context of the target data distribution. In our context, a generated image is plausible if its high-dimensional feature embedding is sufficiently close to the high-dimensional feature embedding of a cloud-free target image. The distance between both embeddings is sufficiently small if there is no fixed number of neighbors closer to the target embedding than the query embedding. For the formalities behind this metric and motivation of the chosen parameterization, please see Appendix B. Recall measures the diversity in generated images and the extent to which the distribution of target data is covered. Analogous to the metric of precision, a target image is recalled if its high-dimensional feature embedding is sufficiently close to the high-dimensional feature embedding of a generated cloud-free image. Note that this allows interpreting recall as a measure of generated image diversity as the metric can score high only if the generated samples are spread out in the feature embedding's space and provide sufficient coverage of the distribution of target images, capturing the heterogeneity of the target images. To summarize, in the context of our data set of Section III-A, precision specifies the closeness of cloud-recovered information to its cloud-free counterpart, whereas recall captures how well the declouded images capture the heterogeneity of the test data (e.g., its diversity in land-cover and seasonality).

While we emphasize the benefit of both measures to disentangle image quality and image heterogeneity, we also define the F1 score as

$$\mathrm{F1}(X, Y) = 2 \cdot \frac{\mathrm{PR}(X, Y) \cdot \mathrm{RC}(X, Y)}{\mathrm{PR}(X, Y) + \mathrm{RC}(X, Y)}$$

where X and Y are sets of images to be compared, and PR and RC denote the functions of precision and recall, respectively. In contrast to the first two experiments, the generation of synthetic data in the third experiment guarantees a one-to-one pixelwise correspondence between cloudy and ground-truth cloud-free images (i.e., perfect coregistration, no atmospheric disturbances other than the simulated noise, control for no landcover, and daylight changes between both observations), ensuring that pixelwise metrics are well-defined. Therefore, complementary to the previous measures of goodness, we additionally assess performances on synthetic data in the third experiment by means of mean absolute error

(MAE), root-mean-square error (RMSE), peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [31], and spectral angle mapper (SAM) [32], as given by

$$\mathrm{MAE}(x, y) = \frac{1}{C \cdot H \cdot W} \sum_{c=h=w=1}^{C,H,W} |x_{c,h,w} - y_{c,h,w}|$$

$$\mathrm{RMSE}(x, y) = \sqrt{\frac{1}{C \cdot H \cdot W} \sum_{c=h=w=1}^{C,H,W} (x_{c,h,w} - y_{c,h,w})^2}$$

$$\mathrm{PSNR}(x, y) = 20 \cdot \log_{10} \left(\frac{1}{\mathrm{RMSE}(x, y)} \right)$$

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x \mu_y + \epsilon_1)(2\sigma_{xy} + \epsilon_2)}{(\mu_x + \mu_y + \epsilon_1)(\sigma_x + \sigma_y + \epsilon_2)}$$

$$\mathrm{SAM}(x, y) = \cos^{-1} \times \left(\frac{\sum_{c=h=w=1}^{C,H,W} x_{c,h,w} \cdot y_{c,h,w}}{\sqrt{\sum_{c=h=w=1}^{C,H,W} x_{c,h,w}^2 \cdot \sum_{c=h=w=1}^{C,H,W} y_{c,h,w}^2}} \right)$$

where x and y are images to be compared with pixel-values $x_{c,h,w}, y_{c,h,w} \in [0, 1]$, dimensions $C = 3$, $H = W = 256$, means μ_x, μ_y, standard deviations σ_x, σ_y, covariance σ_{xy}, and small numbers ϵ_1 and ϵ_2 to stabilize the computation. MAE and RMSE both are pixel-level metrics quantifying the mean deviation between target and predicted images in absolute terms and units of the measure of interest, respectively. PSNR is an imagewise metric to measure how good of a reconstruction in terms of signal-to-noise ratio a recovered image is to a clear target image. SSIM is a second imagewise metric, quantifying the structural differences between the target and predicted images. It is designed to capture perceived change in structural information between two given images, as well as differences in luminance and contrast [31]. The SAM metric is an imagewise measure, quantifying the spectral angle between two images, measuring their similarity in terms of rotations in the space of spectral bands [32]. Further technical information with respect to the metrics utilized in our experiments to quantify goodness of predictions is provided in Appendix B.

2) Quantifying the Benefits of Paired Data: First, we train the architecture described in Section II without using any pixelwise correspondences, as in a manner conventional for cycle-consistent GAN. For our generative model, we consider the VV and VH channels of images from the S1 domain and add a third mean (VV and VH) channel to satisfy the dimension-preservation requirement of cycle-consistent architectures. For images from the S2 domain, all multispectral information is used when computing cloud probability maps, while the S1–S2 mapping uses exclusively the three RGB channels. All images are value-clipped and rescaled to contain values within $[-1, 1]$, while the cloud probability map values are within $[0, 1]$. Value-clipping is within ranges $[-25; 0]$ and $[0; 10\,000]$ for S1 and S2, respectively. Notably, before training, we perform an imagewise shuffling of the optical data of paired cloudy and cloud-free observations to remove the pixelwise correspondences satisfied when cloudy and cloud-free patches would be available as sorted tuples. That is, the optical cloudy and noncloudy patches presented at one training step may be no longer strictly aligned or could reflect differences in landcover and atmosphere, reflecting practical challenges commonly encountered when gathering data for remote sensing applications. We train our network on a 10 000 images multiregion subset of the training split introduced in Section III-A. Network weights w are initialized by sampling from a Gaussian distribution $w \sim \mathcal{N}(\mu = 0, \sigma^2 = 0.02)$. The optimizer and the hyperparameters for the optimizer and loss weightings are set as in [18]: We use ADAM with an initial learning rate $\epsilon_{\mathrm{lr}} = 0.0002$, momentum parameters $\beta = (0.5, 0.999)$ for computing sliding averages of the gradients, and their squares and a small constant of 10^{-8} added to the denominator to ensure numerical stability of the optimizer. Instance normalization [33] is applied to the generators as in the original architecture [18], with adjustments detailed in Fig. 3 and Table I. Spectral normalization [24] is applied to the discriminators as in [34] in order to prevent mode collapse during training [35]. The networks are trained for $n_{\mathrm{iter}} = 50$ epochs at the initial learning rate of ϵ_{lr} and then for another $n_{\mathrm{decay}} = 25$ epochs with a multiplicative learning rate decay given by $\mathrm{lr}_{\mathrm{decay}}(n_{\mathrm{current}}) = 1.0 - \max(0, 1 + n_{\mathrm{current}} - n_{\mathrm{iter}})/(n_{\mathrm{decay}} + 1)$, where n_{current} denotes the current epoch number. The gentle learning rate decay over a long period of epochs serves to ensure a well-behaved optimization process during training [18], [35]. All our generator networks are trained on center-cropped 200×200 px^2 patches but tested on full-sized 256×256 pixels patches of the hold-out split, as the generator architecture is fully convolutional. As proposed in [36] and implemented in [18], we maintain two pools of the last 50 generated images to update the discriminators with a random sample from the respective image buffers such that oscillations during training are reduced [18], [35]. Representative qualitative outcomes are depicted in Fig. 1. The results highlight that our model can reconstruct cloud-covered areas while preserving information that is not obscured. A quantitative evaluation of the described model (ours-0) is given in Table II.

TABLE II
EFFECT OF PERCENTAGE OF PAIRED TRAINED DATA ON PERFORMANCE OF CLOUD REMOVAL MODEL. THE MORE THE PAIRED TRAINING DATA IS AVAILABLE, THE BETTER THE RESULTING PERFORMANCES

% paired	precision	recall	F1 score
0 (ours-0)	0.560	0.491	0.523
10	0.559	0.499	0.527
20	0.560	0.506	0.532
50	0.562	0.528	0.544
100 (ours-100)	**0.564**	**0.551**	**0.557**

Second, we retrain the model, as described earlier, but on paired cloudy–cloudless optical observations in order to assess the benefits of paired training data, as provided by our data set. To let the cycle-consistent architecture described in Section II benefit of paired training data, we combine the losses defined in Section II-C with cost functions defined on paired images: first, a pixelwise L1 loss penalizing prediction errors between generated and paired target images as in [37]; second, perceptual losses for features and style [38], as evaluated on the features extracted at ReLU layers 11, 20, and 29 of an auxiliary pretrained VGG16 network [39]. We retrain our

network with these losses and systematically vary the percent of paired cloudy and cloud-free optical data available. The paired patches are equally spaced across the training split at the beginning of the training procedure, and patch pairings are fixed across epochs. During training, the presentation of paired and unpaired samples occurs in random order. Table II shows the different models' performances. The base model trained on unpaired data (ours-0) performs worst, while the model fully trained on paired samples (ours-100) achieves the best performances. In general, the more paired samples are available the better the model performs.

3) Model Ablation Experiment: To put the results of the previous experiment into perspective and further evaluate the factors benefiting the robust reconstruction of cloud-covered information, we conduct an ablation study. Especially, we investigate the effectiveness of the novel cloud detection mechanism explained in Section II-A and the local cloud-sensitive loss introduced in Section II-C. For this purpose, we retrain the model ours-0, as described in Section II, but omit the cloud-sensitive terms by fixating the values of all pixels in the cloud probability masks m to 1.0. The effect of this is that the ablated model is no longer encouraged to minimize the changes to areas free of cloud coverage, thus potentially resulting in unneeded changes. As additional baselines, we evaluate the goodness of simply using the S1 observations (VV- or VH-polarized), as well as cloud-covered S2 images as predictions and comparing against their cloud-free counterparts. Table III reports the declouding performance of baseline models and our models (0% and 100% paired data from Table II). Our network of 100% paired data performs best in terms of precision and F1 score. The raw S1 and S2 observations perform relatively poorly, except for the cloudy optical images scoring high on image diversity due to random cloud coverage. While it may be useful to consider the raw data as baselines, it is necessary to keep in mind that modalities, such as SAR, maybe at a disadvantage when directly comparing against the cloud-free optical target images.

4) Assessing the Goodness of Synthetic Data: To compensate for the lack of any large-scale data set for cloud removal, previous works simulated the artificial data [6], [7], [10], [40], [41] of synthetic cloudy optical images. This raises the question of the goodness of the simulated observations, i.e., how good of an approximation such simulations are to any real data. In this experiment, we consider the two architectures ours-0 and ours-100 from Table III and retrain them on synthetic data to subsequently evaluate the retrained models on the real test data and assess if performance generalizes to real-world scenarios. Two approaches to generating synthetic data are evaluated.

1) *Perlin:* We generate cloudy imagery via Perlin noise [16] and alpha-blending as in the preceding studies of [6], [7], and [40]. This approach has the limitation of adding Perlin noise to all of the multispectral bandwidths evenly, due to lack of a better physical model of multispectral cloud noise. Since cloud detectors trained on real observations are expected to fail in such a case, we substitute the cloud map of Section II-A by the synthesized alpha-weighted Perlin noise.

TABLE III
CLOUD-REMOVAL PERFORMANCE OF BASELINE METHODS AND OUR MODELS ON TEST SPLIT OF SEN12MS-CR. ROWS S1 VV AND VH REFER TO THE RAW S1 IMAGE, CHANNELS VV AND VH, RESPECTIVELY, COMPARED WITH THE GRAY-SCALE CLOUD-FREE S2 IMAGE. S2 CLOUDY REFERS TO THE RAW CLOUDY S2 IMAGE COMPARED WITH THE RGB CLOUD-FREE S2 IMAGE. ALL MODELS' METRICS BEAT THE LOWER-BOUND PERFORMANCES ESTABLISHED BY THE RAW DATA, EXCEPT ON THE RECALL METRIC. THE FULL MODELS PERFORM BETTER THAN THE ABLATION MODELS WITHOUT THE CLOUD-SENSITIVE LOSS AND CLOUD PROBABILITY MASKS. MODEL OURS-100 PERFORMS BEST IN TERMS OF PRECISION AND F1 SCORE. NOTE THAT THE RESULTS DEPICT A PRONOUNCED TRADEOFF BETWEEN PRECISION AND RECALL, AS ANALYZED, IN DETAIL, IN [28]

model		precision	recall	F1 score
S1	VV	0.000	0.001	0.001
	VH	0.012	0.017	0.014
S2 cloudy		0.161	**0.705**	0.267
ours-0 (no m)		0.181	0.572	0.279
ours-100 (no m)		0.232	0.535	0.323
ours-0		0.560	0.491	0.523
ours-100		**0.564**	0.551	**0.557**

2) *Copy:* We generate cloudy imagery by taking the ground-truth cloud-free optical observations and combine them via alpha-blending with clouded observations as in the approach of [10]. Different from [10], we benefit from our curated data set and alpha-blend paired cloudy–cloudless observations, whereas the prior study mixed the two unrelated images. Moreover, we alpha-blend weighted by the cloud map of Section II-A, whereas the original study alpha-blended via sampled Perlin-noise. We believe that these modifications better preserve the spectral properties of real observations and keep cloud distribution statistics closer to that of real data, as shown in Figs. 6 and 7.

Furthermore, this allows for synthesizing coverage ranging from semitransparent to fully occluded clouds, which would be less straightforward on unpaired observations. Exemplary observations generated by both simulation approaches and empirical observations are presented in Fig. 6.

The outcomes of this experiment are presented in Table IV. For all data simulation approaches, training a network on generated data and, subsequently, evaluating it on synthetic test data are overestimating the performances on the corresponding real test data. This observation holds for both models evaluated in the experiment. The models display a drop in performance when moving from synthetic to real testing data. The drop being considerably smaller in the case of copy–paste data than for Perlin noise data may be due to the copy-pasted data closer resembling the real data and its underlying sta-

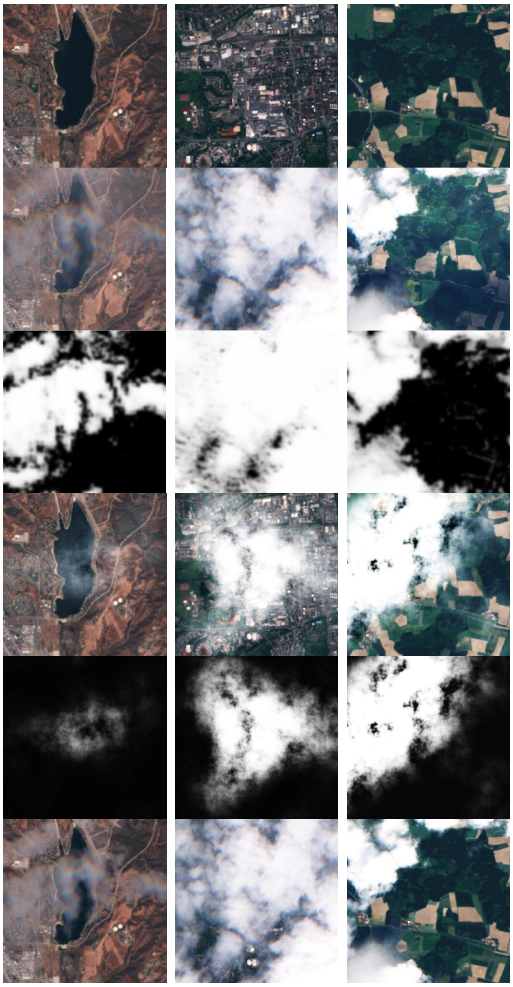


Fig. 6. Exemplary cloud-free, real cloudy, and generated cloudy optical observations. Rows: cloud-free S2 data (plotted in RGB), real cloudy S2 data, real cloud coverage maps (same for copy–paste), Perlin-noise simulated cloudy S2 data, Perlin-noise cloud coverage maps, and copy–paste simulated cloudy S2 data. Columns: three different samples.

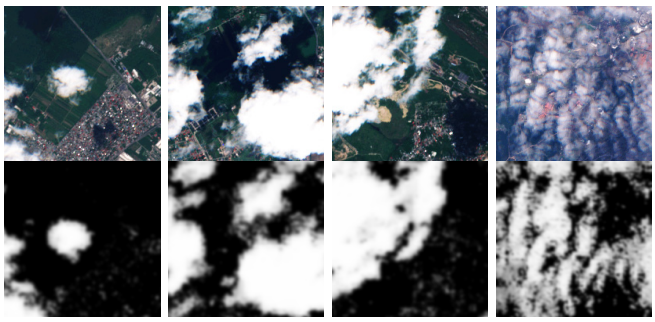


Fig. 7. Exemplary cloudy optical observations and cloud maps. Rows: cloudy S2 data and cloud probability masks. Columns: four different samples.

tistics of cloud coverage and spectral distributions. In this context, it is instructive to investigate spectral distortions by means of SAM, which indicates that models trained and tested on synthetic data are considerably poorer to predict spectral distributions on Perlin-simulated data compared with the copy-pasted observations, which is arguable more alike

TABLE IV
CLOUD-REMOVAL PERFORMANCE OF MODELS OURS-0 AND OURS-100 FROM TABLE III, RETRAINED ON SYNTHETIC CLOUD DATA (EITHER PERLIN-SIMULATED OR COPY-PASTED) AND TESTED ON SYNTHETIC AND REAL DATA. BOTH MODELS, WHEN TRAINED ON SYNTHETIC DATA, PERFORM MUCH BETTER ON SYNTHETIC TEST DATA THAN ON REAL TEST DATA. IMPORTANTLY, THE TEST PERFORMANCE OF MODELS TRAINED ON SYNTHETIC AND TESTED ON REAL DATA IS CONSIDERABLY POORER THAN THAT OF THE SAME ARCHITECTURES TRAINED ON REAL DATA (REPORTED IN TABLE III)

model		ours-0		ours-100	
metric		Perlin	copy	Perlin	copy
MAE		0.045	0.023	0.041	0.017
RMSE		0.067	0.031	0.059	0.023
PSNR		24.75	34.034	25.775	35.802
SSIM		0.803	0.882	0.824	0.904
SAM		27.527	10.626	26.013	9.936
precision	synth	0.155	0.693	0.239	0.692
	real	0.115	0.425	0.168	0.458
recall	synth	0.781	0.851	0.800	0.856
	real	0.624	0.611	0.592	0.586
F1	synth	0.258	0.764	0.368	0.766
	real	0.194	0.501	0.262	0.514

to real data in terms of its spectral properties. The findings in this experiment underline the need for synthetic data to closely capture the properties of real data, yet even when real and synthetic observations may be hardly distinguishable by eye (as the examples shown in Fig. 6), there persist important discrepancies unaccounted for, which hinders the models trained on synthetic sampled to perform equally on real data.

IV. DISCUSSION

The contribution of our work is in providing a large-scale and global data set for cloud removal and developing a new model for recovering cloud-covered information to highlight the data sets benefits. With over 55% of the Earth's land surface covered by clouds [1], the ability to penetrate cloud coverage is of great interest to the remote community in order to obtain continuous and seamless monitoring of our planet. While the focus in this work is on providing the first globally sampled multimodal data set for general-purpose cloud removal, future research should also address the benefits of cloud removal approaches for particular applications common in remote sensing. An example application is in semantic segmentation, which necessitates clear-view observations for accurate land-cover classification. Another, in the context of having consecutive observations over time, would be change or anomaly detection where cloud removal methods may be beneficial particularly for the purpose of early stage detection, which could, otherwise, be delayed in the presence of clouds. A limitation of our proposed cloud removal model is its restriction to work on a subset of the optical observation's spectral bands. While this constraint is required due to the choice of the network architecture as necessitated by our experiments conducted, we are certain that it will be beneficial

for future research to consider the full spectral information. To allow for this, our curated global data set is released with all available information for both modalities, including the full spectrum of bands for the optical observations.[1]

V. Conclusion

We demonstrated the declouding of optical imagery by fusing multisensory data, proposed a novel model, and released the, to the best of our knowledge, first global data set combining over a 100 000 paired cloudy, cloud-free, and coregistered SAR sample triplets. Statistical analysis of our data set shows a relatively uniform distribution of cloud coverage, with clear images occurring just as probable as wide and densely occluded ones—indicating the need for flexible cloud removal approaches to potentially handle either case. Our proposed network explicitly models cloud coverage and, thus, learns to retain cloud-free information while as well being able to recover information of areas covered by wide or dense clouds. We evaluated our model on a globally sampled test set and measure the goodness of predictions with recently proposed metrics that capture both prediction quality and coverage of the target distribution. Moreover, we showed that our model benefits from supervised learning on paired training data as provided by our large-scale data set. Finally, we evaluated the goodness of synthetically generated data of cloudy–cloudless image pairs and show that great performance on synthetic data may not necessarily translate to equal performance on real data. Importantly, when testing on real data, the networks trained on real observations consistently outperform models trained on synthetic observations, indicating the existence of properties of the real observations not modeled sufficiently well by the simulated data. This underlines the need for a set of real observations numerous enough to train large models, as provided by the data set released in this work. In further studies, we will address the fusion of multitemporal and multisensory data, combining and comparing across both currently segregated approaches. To support future research and make contributions comparable, we share our global data set of paired cloudy, cloud-free, and coregistered SAR imagery and provide our test data split for benchmarking purposes.

Appendix A
Cloud Detection

We present exemplary cloudy optical observations and cloud maps in Fig. 7. The cloud masks are as predicted by our cloud detection pipeline detailed in Section II-A. The illustrated examples show that our proposed method can reliably detect clouds and provide continuous-valued cloud masks.

Appendix B
Improved Precision and Recall

We provide a definition of improved precision and recall in line with the definitions in [28]. For further details, the interested reader is referred to the original publication.

Definition (Improved Precision and Recall [28]): Let $X_r \sim P_r$ and $X_g \sim P_g$ denote paired samples drawn from the real and generated distributions of cloud-free images, where P_g is the distribution learned by the generator network whose quality is to be assessed. Each sample is mapped via an auxiliary pretrained network M^2 in a high-dimensional feature space to obtain latent representations $\phi_r = M(X_r)$ and $\phi_g = M(X_g)$ such that the two sets of samples are mapped into two feature sets Φ_r and Φ_g. A distribution $P \in \{P_r, P_g\}$ is approximated by computing pairwise distances between feature embeddings of the observed samples $\Phi \in \{\Phi_r, \Phi_g\}$ and, centered at each feature $\phi \in \Phi$, forming a hypersphere with a radius corresponding to the distance to its kth nearest neighbor embedding $N_k(\phi)$. Hence, whether an embedded sample ϕ falls on manifold Φ or not is given via

$$f(\phi, \Phi) = \begin{cases} 1, & \text{if } \exists \phi' \in \Phi : \|\phi - \phi'\| \leq \|\phi' - N_k(\phi')\|_2 \\ 0, & \text{else.} \end{cases}$$

The fraction of samples that fall on the paired distribution's manifold are then defined in [28] as

$$\text{precision}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_g|} \sum_{\phi_g \in \Phi_g} f(\phi_g, \Phi_r)$$

$$\text{recall}(\Phi_r, \Phi_g) = \frac{1}{|\Phi_r|} \sum_{\phi_r \in \Phi_r} f(\phi_r, \Phi_g).$$

We set parameters $|\Phi| = 7893$ corresponding to the size of the test split of SEN12MS-CR and $k = 10$ because every sample has up to 50% overlap with its neighboring samples. This setting removes the paired target itself plus its eight overlapping samples when computing $N_k(\phi)$.

Appendix C
Cloud Coverage Statistics on Test Split

In addition to the cloud coverage statistics on the entire data set, as reported in Section III-A, Fig. 8 provides the empirically observed distribution of cloud coverage on the data sets test split. Even though the histogram of the test split is less smooth than that of the complete data set due to the test split being much smaller, both distributions are considerably alike.

Appendix D
Exemplary Problematic Cases

For the sake of completeness, we discuss cases that we consider challenging for cloud removal approaches, specifically our method, and present exemplary data and predictions of such cases in Fig. 9. We consider the following challenges.

1) Changes in landcover, atmosphere, day time acquisition, or seasonality that may occur between (visible parts of) the cloudy reference image and the cloud-free target optical image. While our data set is curated to minimize such cases by selecting observations that are close in

[1]The SEN12MS-CR data set is shared under the CC-BY 4.0 open access license and available for download provided by the library of the Technical University of Munich (TUM): https://mediatum.ub.tum.de/1554803. This article must be cited when the data set is used for research purposes.

[2]Here, VGG16 [39], with features extracted at the second fully connected layer, as argued for in [42]. Metric evaluation on alternative pretrained networks has shown to provide virtually identical results [28].

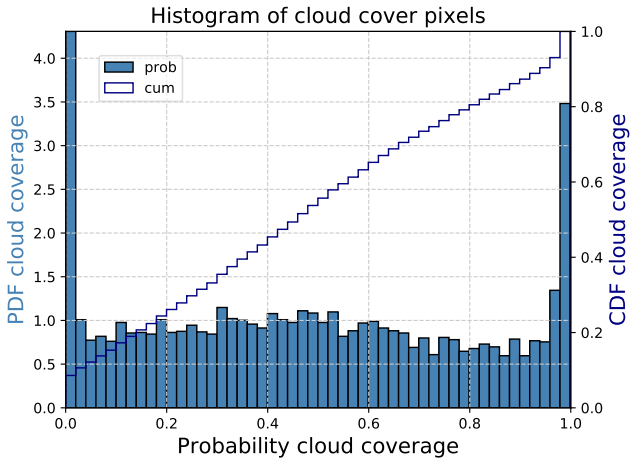


Fig. 8. Statistics of cloud coverage of test split of SEN12MS-CR. As for the statistics on the complete data set, an average of circa 50% of occlusion is observed.

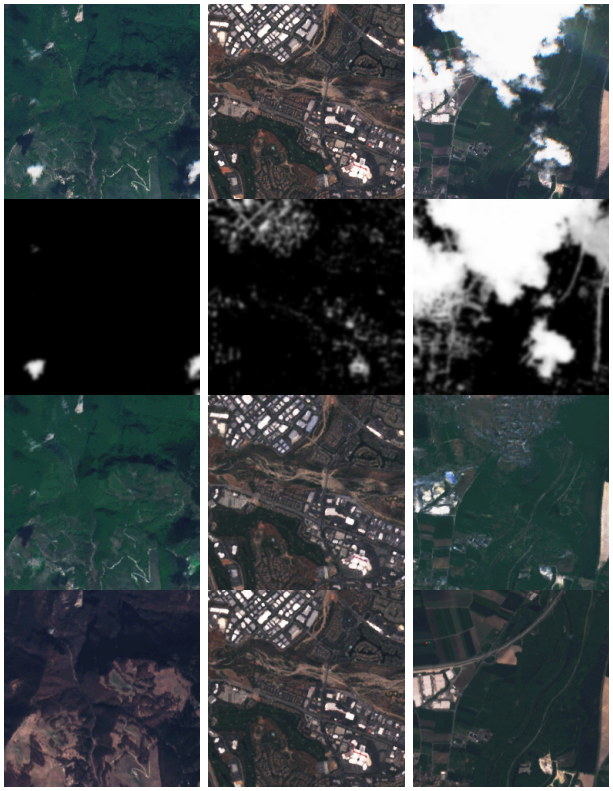


Fig. 9. Exemplary cases posing challenges to our cloud-removal approach. Rows: S2 data (in RGB), predicted cloud map m, predicted $\hat{S}2$ data, and cloud-free (target) S2 data. Columns: three different samples. Reconstructing optical information obscured by clouds is a hard problem. Among the challenges faced by cloud removal approaches may be: 1) overtime changes in landcover, atmosphere, day time acquisition, or seasonality; 2) precise detection of clouds with few misses and false alarms; and 3) correct reconstruction of information fully covered by large and dense clouds.

time, strict ground-truth correspondence is challenging to establish and may only be guaranteed by simulating synthetic data as in experiment III-B4.

2) Precise detection of clouds and accurate cloud masks that minimizes false alarms and misses. With respect to our cloud detection algorithm, there exist cloud masks where, even for completely cloud-free images, pixels are assigned a nonzero (albeit rather low) probability of being cloudy.
3) Correct reconstruction of cloud-covered information. In particular, for the case of complete coverage by large and dense clouds, this is a very challenging problem. We observed the cases where the information reconstructed by our model did not match the target images; for instance, urban-like landcover was predicted in place of agricultural areas.

ACKNOWLEDGMENT

The authors would like to thank ESA and the Copernicus Program for making the Sentinel observations accessed for this submission publicly available. They would also like to thank Lloyd Hughes for having shared his artifact detection preprocessing code with them.

REFERENCES

[1] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 3826–3852, Jul. 2013.
[2] A. Singh, "Digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, 1989.
[3] J. R. Jensen *et al.*, *Introductory Digital Image Processing: A Remote Sensing Perspective*, no. 2. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.
[4] P. Coppin, E. Lambin, I. Jonckheere, and B. Muys, "Digital change detection methods in natural ecosystem monitoring: A review," in *Analysis of Multi-Temporal Remote Sensing Images*. Singapore: World Scientific, 2002, pp. 3–36.
[5] C. E. Woodcock, T. R. Loveland, M. Herold, and M. E. Bauer, "Transitioning from change detection to monitoring with remote sensing: A paradigm shift," *Remote Sens. Environ.*, vol. 238, Mar. 2020, Art. no. 111558.
[6] K. Enomoto *et al.*, "Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1533–1541. [Online]. Available: https://arxiv.org/abs/1710.04835
[7] C. Grohnfeldt, M. Schmitt, and X. Zhu, "A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 1726–1729. [Online]. Available: https://ieeexplore.ieee.org/document/8519215/
[8] P. Singh and N. Komodakis, "Cloud-GAN: Cloud removal for Sentinel-2 imagery using a cyclic consistent generative adversarial networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 1772–1775. [Online]. Available: https://ieeexplore.ieee.org/document/8519033/
[9] J. D. Bermudez, P. N. Happ, R. Q. Feitosa, and D. A. B. Oliveira, "Synthesis of multispectral optical images from SAR/optical multitemporal data using conditional generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1220–1224, Aug. 2019.
[10] M. U. Rafique, H. Blanton, and N. Jacobs, "Weakly supervised fusion of multiple overhead images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1479–1486.
[11] V. Sarukkai, A. Jain, B. Uzkent, and S. Ermon, "Cloud removal in satellite images using spatiotemporal generative networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1796–1805.
[12] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 333–346, Aug. 2020.
[13] M. A. Friedl *et al.*, "Global land cover mapping from MODIS: Algorithms and early results," *Remote Sens. Environ.*, vol. 83, nos. 1–2, pp. 287–302, Nov. 2002.
[14] R. Bamler, "Principles of synthetic aperture radar," *Surv. Geophys.*, vol. 21, nos. 2–3, pp. 147–157, 2000.

[15] J. D. Bermudez, P. N. Happ, D. A. B. Oliveira, and R. Q. Feitosa, "SAR to optical image synthesis for cloud removal with generative adversarial networks," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 4, pp. 5–11, Sep. 2018.

[16] K. Perlin, "Improving noise," in *Proc. 29th Annu. Conf. Comput. Graph. Interact. Techn.*, 2002, pp. 681–682.

[17] M. F. Reyes, S. Auer, N. Merkle, C. Henry, and M. Schmitt, "SAR-to-optical image translation based on conditional generative adversarial network—Optimization, opportunities and limits," *Remote Sens.*, vol. 11, no. 17, p. 2067, Sep. 2019.

[18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251. [Online]. Available: http://ieeexplore.ieee.org/document/8237506/

[19] J. Gao, Q. Yuan, J. Li, H. Zhang, and X. Su, "Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks," *Remote Sens.*, vol. 12, no. 1, p. 191, Jan. 2020.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[21] A. Zupanc. (2017). *Improving Cloud Detection With Machine Learning*. Accessed: Oct. 10, 2019. [Online]. Available: https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13

[22] V. Lonjou *et al.*, "MACCS-ATCOR joint algorithm (MAJA)," *Proc. SPIE*, vol. 10001, Oct. 2016, Art. no. 1000107.

[23] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote Sens. Environ.*, vol. 229, pp. 247–259, Aug. 2019.

[24] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–26. [Online]. Available: https://openreview.net/forum?id=B1QRgziT-

[25] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

[26] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[27] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS—A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2/W7. 2019, pp. 153–160, doi: 10.5194/isprs-annals-IV-2-W7-153-2019.

[28] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 3929–3938.

[29] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[30] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[32] F. A. Kruse *et al.*, "The spectral image processing system (SIPS)-interactive visualization and analysis of imaging spectrometer data," *AIP Conf.*, vol. 283, no. 1, pp. 192–201, 1993

[33] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," Nov. 2016, *arXiv:1607.08022*. [Online]. Available: http://arxiv.org/abs/1607.08022

[34] S. Mo, M. Cho, and J. Shin, "InstaGAN: Instance-aware image-to-image translation," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–26.

[35] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," 2017, *arXiv:1701.00160*. [Online]. Available: http://arxiv.org/abs/1701.00160

[36] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2107–2116.

[37] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[38] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, vol. 9906. Cham, Switzerland: Springer, 2016, p. 694–711.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[40] W. Sintarasirikulchai, T. Kasetkasem, T. Isshiki, T. Chanwimaluang, and P. Rakwatin, "A multi-temporal convolutional autoencoder neural network for cloud removal in remote sensing images," in *Proc. 15th Int. Conf. Electr. Eng./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON)*, Jul. 2018, pp. 360–363.

[41] D. Tedlek, S. Khoomboon, T. Kasetkasem, T. Chanwimaluang, and I. Kumazawa, "A cloud-contamination removal algorithm by combining image segmentation and level-set-based approaches for remote sensing images," in *Proc. Int. Conf. Embedded Syst. Intell. Technol., Int. Conf. Inf. Commun. Technol. Embedded Syst. (ICESIT-ICICTES)*, May 2018, pp. 1–5.

[42] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–35.

Patrick Ebel (Graduate Student Member, IEEE) received the B.Sc. degree in cognitive science from the University of Osnabrück, Osnabrück, Germany, in 2015, and the M.Sc. degree in cognitive neuroscience and the M.Sc. degree in artificial intelligence from Radboud University Nijmegen, Nijmegen, The Netherlands, in 2018. He is pursuing the Ph.D. degree with the SiPEO Laboratory, Department of Aerospace and Geodesy, Technical University of Munich, Munich, Germany.

His research interests include deep learning and its applications in computer vision and to remote sensing data.

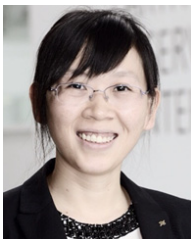
Andrea Meraner received the B.Sc. degree in physics and the M.Sc. degree (Hons.) in Earth oriented space science and technology (ESPACE) from the Technical University of Munich (TUM), Munich, Germany, in 2016 and 2019, respectively.

In 2017, he was a Research Assistant with Atmospheric Modeling Group, German Geodetic Research Institute (DGFI-TUM). In 2018, he was with the German Aerospace Center (DLR)–German Remote Sensing Data Center, Weßling, Germany, in the International Ground Segment department. After spending one semester at IIT Mandi, Suran, India, in 2019, he was a Research Assistant with the Signal Processing in Earth Observation (SiPEO) Group, TUM, and the Remote Sensing Technology Institute, DLR, working on deep learning-based cloud removal algorithms for optical satellite imagery. Since October 2019, he has been a Junior Remote Sensing Scientist for optical imagery at EUMETSAT European Organization for the Exploitation of Meteorological Satellites, Darmstadt, Germany, where he is developing algorithms to process and analyze data from current and future geostationary satellite missions.

Michael Schmitt (Senior Member, IEEE) received the Dipl.Ing. (Univ.) degree in geodesy and geoinformation, the Dr.-Ing. degree in remote sensing, and the Habilitation degree in data fusion from the Technical University of Munich (TUM), Munich, Germany, in 2009, 2014, and 2018, respectively.

Since 2020, he has been a Full Professor of applied geodesy and remote sensing with the Department of Geoinformatics, Munich University of Applied Sciences, Munich. From 2015 to 2020, he was a Senior Researcher and the Deputy Head at the Professorship for Signal Processing in Earth Observation, TUM. In 2019, he was additionally appointed as an Adjunct Teaching Professor at the Department of Aerospace and Geodesy, TUM. In 2016, he was a Guest Scientist with the University of Massachusetts at Amherst, Amherst, MA, USA. His research focuses on image analysis and machine learning applied to the extraction of information from multimodal remote sensing observations. In particular, he is interested in remote sensing data fusion with a focus on synthetic aperture radar (SAR) and optical data.

Dr. Schmitt is the Co-Chair of the Working Group "SAR and Microwave Sensing" of the International Society for Photogrammetry and Remote Sensing and the Working Group "Benchmarking" of the IEEE-GRSS Image Analysis and Data Fusion Technical Committee. He frequently serves as a reviewer for a number of renowned international journals and conferences and has received several best reviewer awards. He is an Associate Editor of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.



Xiao Xiang Zhu (Senior Member, IEEE) received the M.Sc., Dr.Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, The University of Tokyo, Tokyo, Japan, and the University of California at Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. Since 2019, she has been a co-coordinator of the Munich Data Science Research School. Since 2019, she also heads the Helmholtz Artificial Intelligence Cooperation Unit (HAICU)—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the Director of the International Future AI laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. She is a Professor with the Signal Processing in Earth Observation, TUM, and the Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. Her main research interests are remote sensing and Earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of the Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is also an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

A.3 GLF-CR: SAR-enhanced cloud removal with global–local fusion

Reference: F. Xu, Y. Shi, P. Ebel, L. Yu, G.-S. Xia, W. Yang, and X. X. Zhu. *GLF-CR: SAR-enhanced cloud removal with global–local fusion.* ISPRS Journal of Photogrammetry and Remote Sensing, 192:268–278, 2022

ISPRS Journal of Photogrammetry and Remote Sensing 192 (2022) 268–278

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

GLF-CR: SAR-enhanced cloud removal with global–local fusion

Fang Xu [a,b], Yilei Shi [c], Patrick Ebel [b], Lei Yu [a], Gui-Song Xia [d], Wen Yang [a,*], Xiao Xiang Zhu [b,*]

[a] School of Electronic Information, Wuhan University, Wuhan, 430072, China
[b] Data Science in Earth Observation, Technical University of Munich, Munich, 80333, Germany
[c] Remote Sensing Technology, Technical University of Munich, Munich, 80333, Germany
[d] School of Computer Science, Wuhan University, Wuhan, 430072, China

ARTICLE INFO

Keywords:
Cloud removal
Data fusion
SAR
Transformer

ABSTRACT

The challenge of the cloud removal task can be alleviated with the aid of Synthetic Aperture Radar (SAR) images that can penetrate cloud cover. However, the large domain gap between optical and SAR images as well as the severe speckle noise of SAR images may cause significant interference in SAR-based cloud removal, resulting in performance degeneration. In this paper, we propose a novel global–local fusion based cloud removal (GLF-CR) algorithm to leverage the complementary information embedded in SAR images. Exploiting the power of SAR information to promote cloud removal entails two aspects. The first, global fusion, guides the relationship among all local optical windows to maintain the structure of the recovered region consistent with the remaining cloud-free regions. The second, local fusion, transfers complementary information embedded in the SAR image that corresponds to cloudy areas to generate reliable texture details of the missing regions, and uses dynamic filtering to alleviate the performance degradation caused by speckle noise. Extensive evaluation demonstrates that the proposed algorithm can yield high quality cloud-free images and outperform state-of-the-art cloud removal algorithms with a gain about 1.7 dB in terms of PSNR on SEN12MS-CR dataset.

1. Introduction

Earth observation through satellites plays a vital role in understanding the world, and has attracted attention from a wide range of communities (Xia et al., 2018; Requena-Mesa et al., 2021; Girard et al., 2021). However, optical satellite images are often contaminated by clouds, which obstruct the view of the surface underneath, as shown in Fig. 1(a). A study conducted by the MODIS instrument shows that the overall global cloudiness is roughly 67% and the cloud fraction over land is about 55% (King et al., 2013). Thus, cloud removal becomes an indispensable pre-processing step for applications relying on data streams of continuous monitoring (Ebel et al., 2021). Due to the erasure of textures in cloud-covered regions, the task of cloud removal is severely ill-posed. Benefiting from Synthetic Aperture Radar (SAR) (Bamler, 2000) (as shown in Fig. 1(b)), which is not affected by clouds due to its advantage of strong penetrability and measures the backscatter, the challenge of cloud removal can be essentially alleviated. However, the recovery of high-quality cloud-free images with the aid of SAR images is nevertheless a challenging problem due to the following issues:

- **Domain Gap**. SAR and optical images reveal different characteristics of observed objects due to their different imaging mechanisms, and thus a large domain gap exists between them (Schmitt et al., 2017; Liu and Lei, 2018). Transferring the complementary information from a SAR image to compensate for the missing information in cloudy regions is non-trivial.
- **Speckle Noise**. SAR images exhibit bright and dark pixels, i.e., speckle noise, which is uneven, even for homogeneous regions (Yu et al., 2018; Zhu et al., 2021). Moreover, the speckle noise usually exists in the same wave front as the surface information of the target. This undesirable effect leads to performance degradation on reconstruction (Fuentes Reyes et al., 2019; Liu et al., 2021b).

A few SAR-based cloud removal methods to learn to transfer the concatenation of multi-modal images to cloud-free images have been proposed (Gao et al., 2020; Meraner et al., 2020; Ebel et al., 2021). However, the pixel-to-pixel translation does not take into account the long-range varying contextual information of the cloud-free regions, leading to texture and structure discrepancies. Furthermore, this concatenation method only partially explores the interactions or correlations between optical and SAR data, in which complementary

* Corresponding authors.
E-mail addresses: yangwen@whu.edu.cn (W. Yang), xiaoxiang.zhu@tum.de (X.X. Zhu).

https://doi.org/10.1016/j.isprsjprs.2022.08.002
Received 26 April 2022; Received in revised form 4 July 2022; Accepted 6 August 2022
Available online 1 September 2022

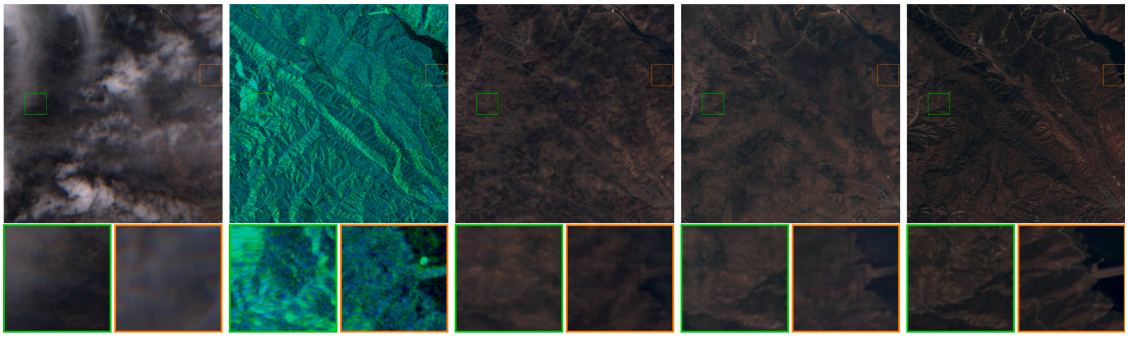


Fig. 1. Illustrative example of SAR-based cloud removal on a large scale cloudy image. (a) Cloudy optical image; (b) SAR image; (c) result of DSen2-CR (Meraner et al., 2020); (d) result of our proposed GLF-CR; (e) cloud-free image. The GLF-CR can restore images with more details and fewer artifacts. The size of each image is 1000×1000 pixels.

information cannot be effectively transferred. Moreover, simply stacking multi-modal images is susceptible to speckle noise, which hinders the cloud removal performance.

To tackle the issues and limitations above, we propose a novel global–local fusion-based cloud removal (GLF-CR) algorithm by exploring the full potential of SAR image. It has been shown that SAR images help to recover texture details by compensating for the missing information in cloudy regions (Meraner et al., 2020). In addition, since a SAR image is not obscured by clouds, it contains reliable global contextual information that can provide valuable guidance for capturing global interactions between contexts to maintain global consistency with the remaining cloud-free regions. Specifically, GLF-CR contains two parallel backbones developed for optical and SAR image representation learning, where SAR features are used in a hierarchical manner to compensate for the loss of information. Inspired by Transformer architectures (Vaswani et al., 2017) that can capture global interactions between contexts, we propose a SAR-guided global context interaction (SGCI) block in which SAR features are used to guide the interactions of global optical feature. Furthermore, a SAR-based local feature compensation (SLFC) block is proposed to transfer complementary information from the corresponding regions in the SAR features to the optical features, where dynamic filtering is used to handle speckle noise. Consequently, the proposed algorithm can generate knowledgeable features with comprehensive information, thereby yielding high-quality cloud-free images.

To sum up, the contributions of this work are three-fold:

- We propose a novel SAR-based cloud removal algorithm, *GLF-CR*. It incorporates the contribution of SAR to restoring reliable texture details and maintaining global consistency, thus enabling the region occluded by cloud cover to be effectively reconstructed.
- We propose a SAR-guided global context interaction (SGCI) block, in which the SAR feature is used to guide the global interactions between contexts in order to ensure that the structure of the recovered cloud-free region is consistent with the remaining cloud-free regions.
- We propose a SAR-guided local feature compensation (SLFC) block to enhance the transference of complementary information embedded in the SAR image while avoiding the influence of speckle noise, and thus generate more reliable texture details.

2. Related work

Cloud Removal. Cloud removal aims to reconstruct the missing information caused by clouds in optical satellite imagery. Early attempts address this problem by assuming the corrupted regions and the remaining regions share the same statistical and geometrical structures. They view cloud removal as an inpainting task and use the information around the corrupted regions to predict the missing data (Chan and Shen, 2001; Maalouf et al., 2009). Many recent studies learn the mapping between cloudy and cloud-free images by benefiting from the remarkable generative capabilities of Generative Adversarial Networks (GANs) (Singh and Komodakis, 2018; Wen et al., 2021; Zi et al., 2022). These methods fail to make accurate inferences when the corrupted region occupies a large portion of the image. To mitigate this problem, a series of studies make use of multispectral data to recover the missing information (Shen et al., 2013; Xu et al., 2015; Enomoto et al., 2017). For example, McGANs (Enomoto et al., 2017) and CR-GAN-PM (Li et al., 2020) utilize additional near-infrared (NIR) images, which process higher penetrability through clouds, to improve visibility. However, as the thickness of clouds increases, all the land signals in the optical bands are obstructed. Consequently, multitemporal-based approaches have been proposed to restore the missing information with data from other time periods (Scarpa et al., 2018; Shen et al., 2019; Zhang et al., 2021; Gao et al., 2021; Ebel et al., 2022). However, when encountering continual cloudy days, cloud-free reference data from an adjacent period is largely unavailable.

Synthetic Aperture Radar (SAR) images are cloud-penetrable and thus provide missing information due to optically impenetrable clouds (Bamler, 2000). There is promising potential in SAR-to-optical image translation. Some researchers have tried to generate optical images directly from SAR (Bermudez et al., 2018; Fuentes Reyes et al., 2019). However, since SAR lacks spectrally resolved measurements, there are domain-specific potentials and peculiarities that cannot be compensated. It is challenging to guarantee the quality of the generated optical image translated from a SAR image. Recently, a few studies have explored the means of SAR-optical data fusion, exploiting the synergistic properties of the two imaging systems to guide cloud removal. Meraner et al. (2020) concatenate the SAR image to the input optical image and use a deep residual neural network to predict the target cloud-free optical image. Gao et al. (2020) utilize a two-step approach, first translating the SAR image into a simulated optical image, and then concatenating the simulated optical image, the SAR image, and the optical image corrupted by clouds to reconstruct the corrupted regions using the generative adversarial network (GAN). Experiments have verified the usefulness of SAR-optical data fusion, but its gain is limited because the concatenation approach has limited ability to utilize the complementary information from the SAR image. To boost the gain that comes with the additional SAR information, we propose a novel cloud removal algorithm, GLF-CR, which incorporates the contribution of SAR to restoring reliable texture details and maintaining global consistency to compensate for information loss in cloudy regions.

Image Restoration. Cloud removal is essentially an image restoration task in which a high-quality clean image is reconstructed from a low-quality, degraded counterpart. Recent advances in image restoration follow convolutional neural network (CNN), and numerous CNN-based models have been proposed to improve restoration performance (Zhang

et al., 2018a, 2020; Wang et al., 2021). Global context plays an important role in local pixel-wise recovery. However, convolution is not effective for long-range dependency modeling under the principle of local processing (Liang et al., 2021). To ensure visually consistent restoration results, a series of research focuses on the attention mechanism to obtain global dependency information. Wang et al. (2019) exploit a two-round four-directional IRNN architecture to accumulate global contextual information. Zheng et al. (2019) introduce a short+long term attention layer to ensure appearance consistency in the image domain. Recently, Transformer that employs a self-attention mechanism to capture global interactions between contexts (Liu et al., 2021c) has been proposed and shows promising performance in image restoration (Liang et al., 2021). While the task of SAR-enhanced cloud removal studied in this paper needs to integrate both the information from the degraded image itself and the information from auxiliary SAR image, which is more challenging.

Most existing cloud removal methods are carried out by extending the input channels of the popular CNN architectures. For example, McGAN (Enomoto et al., 2017) extends the input channels of the conditional Generative Adversarial Networks (cGANs) so that they are compatible with multispectral images. DSen2CR (Meraner et al., 2020) is derived from the EDSR network (Lim et al., 2017), and concatenates the SAR's channels and the other channels of the input optical image as input. These architectures are usually designed for tasks like super-resolution and motion deblurring, where the local information from the original low-quality image is only partially lost. For the cloud removal task, all the local information in the area covered by thick clouds is missing because the clouds completely corrupt the reflectance signal. Thus, the cloud removal methods extended from these architectures have limited ability to fully utilize the spatial consistency between the cloudy and the neighboring cloud-free regions. In comparison, the architecture presented in this work is designed to integrate the global context information under the guidance of the SAR image.

Multi-Modal Data Fusion. Commonly used fusion strategies include element-wise multiplication/addition or concatenation between different types of features (Sun et al., 2019; Fu et al., 2020; Xu et al., 2021); this multi-modal data fusion yields limited performance gain (Wu and Han, 2018; Audebert et al., 2018; Liu et al., 2021a). To better exploit the complementary information of the auxiliary data, Hazirbas et al. (2016) propose FuseNet for semantic segmentation. FuseNet contains two branches to extract features from the RGB and depth images, and constantly fuses them via element-wise summation. Liu et al. (2021a) propose an information aggregation distribution module for crowd counting, which consists of two branches for modality-specific representation learning (i.e., RGB and thermal image) and an additional branch for modality-shared representation learning. It dynamically enhances the modality-shared and modality-specific representations with a dual information propagation mechanism. These methods increase the utilization of complementary information of auxiliary data. Nevertheless, little consideration has been given to SAR-optical data fusion for cloud removal, the specific challenges of which are addressed and resolved in this work.

3. Problem statement

Given a cloudy image I defined over $\mathcal{X} \triangleq \mathcal{C} + \mathcal{O}$ with $\mathcal{C}$ and $\mathcal{O}$ respectively denoting the *cloud-covered* and *cloud-free regions*, the task of **cloud removal** aims at restoring the cloud-covered region of the image, i.e., $I_{\mathcal{C}}$. Generally, this task is severely ill-posed due to the missing information caused by clouds in optical satellite observations.

Inpainting. The basic strategy is to infer the cloud-covered region $I_{\mathcal{C}}$ from the cloud-free part $I_{\mathcal{O}}$, and thus it can be considered as *inpainting* task, i.e.,

$$I_{\mathcal{C}} = \mathbf{F}_{\mathrm{INP}}(I_{\mathcal{O}}; S(I)), \tag{1}$$

where $\mathbf{F}_{\mathrm{INP}}$ is an inpainting operator conditioned by latent structures of the whole image, i.e., $S(I)$. Specifically, $S(I)$ represents priors of images, e.g., smoothness, non-local similarities, or learned features embedding from data, with which the task of cloud removal is tractable. However, latent structures of $S(I)$ are not generally holistic or are even unavailable in a cloud removal task when the cloud-covered region is dominant, leading to the failure of reconstruction if only a cloudy image I is utilized.

Translation. The SAR image B is cloud free and can provide a valuable source that compensates for the information missing from the cloudy region. Inspired by the great success in style transfer work achieved by deep learning, existing SAR-based cloud removal methods mainly translate the SAR image to an optical image to remove clouds pixel-by-pixel:

$$I_{\mathcal{C}} = \mathbf{F}_{\mathrm{TRF}}(B_{\mathcal{C}}; R(B, I)), \tag{2}$$

where $\mathbf{F}_{\mathrm{TRF}}$ is a transfer operator conditioned by the inherent relationship between SAR image B and optical image I, i.e., $R(B, I)$. Specifically, $R(B, I)$ represents the cross-modality transferring, which is usually learned from the dataset using the generative adversarial network (GAN) by feeding the stack of multi-modal data channels. However, the pixel-by-pixel translation does not take the spatial consistency between the cloudy and neighboring cloud-free regions into consideration. It consequently leads to the failure to maintain global consistency. Moreover, its method of stacking the channels of SAR and optical images is somewhat straightforward but only partially explores interactions or correlations between multi-modal data. It thus leads to limited performance improvement despite the assistance of the SAR images. And it is further influenced by the speckle noise in the SAR images, leading to reconstruction error.

Thus the main obstacles to boosting cloud removal performance are two-fold.

- The network should effectively transfer the complementary information from SAR image $B_{\mathcal{C}}$ to the optical image while overcoming the influence of its speckle noise to generate reliable texture details.
- The surface information from the cloud-free region $I_{\mathcal{O}}$ should be considered to maintain the structure of the recovered cloud-free region consistent with the remaining cloud-free regions.

Global–Local Fusion. Thus the task of SAR-enhanced cloud removal is to develop an operator $\mathbf{F}_{\mathrm{fusion}}$ conditioned by both the inherent relationship between SAR and optical images and latent structures of the whole image, i.e.,

$$I_{\mathcal{C}} = \mathbf{F}_{\mathrm{fusion}}(I_{\mathcal{O}}, B_{\mathcal{C}}; S(I/B), R(B, I)), \tag{3}$$

where $S(I/B)$ is the non-local context information of the cloudy image learned under the guidance of SAR image. Since the SAR image is not affected by cloud cover, it can provide valuable guidance for capturing global interactions between contexts, so as to maintain the structure of the recovered cloud-free region consistent with the remaining cloud-free regions. $R(B, I)$ in $\mathbf{F}_{\mathrm{fusion}}$ is different from its counterpart in $\mathbf{F}_{\mathrm{TRF}}$, which incorporates the information of the SAR image by stacking its channels to the optical image. We propose instead a more effective fusion strategy to transfer the complementary information from the corresponding region in the SAR image, so as to generate more reliable texture details.

4. Method

4.1. Overview

The overall framework of the proposed GLF-CR algorithm is illustrated in Fig. 2. It is a two-stream network in which the SAR feature is hierarchically fused into the optical feature to compensate for information loss in cloudy regions. Exploiting the power of SAR information

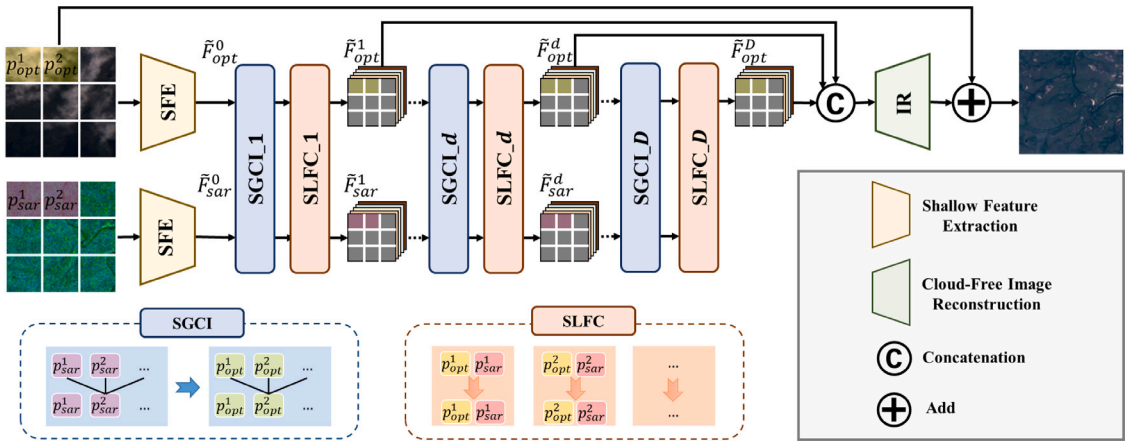


Fig. 2. Overview of the proposed global–local fusion based cloud removal (GLF-CR) algorithm. It is a two-stream network in which the SAR feature is hierarchically fused into the optical feature to compensate for information loss in cloudy areas. Exploiting the power of SAR information to promote cloud removal entails two aspects: global fusion, to guide the relationship among all local optical windows based on the SAR-guided global context interaction (SGCI) block; and local fusion, to transfer the SAR feature corresponding to cloudy areas based on the SAR-based local feature compensation (SLFC) block.

to promote cloud removal entails two aspects: global fusion, to guide the relationship among all local optical windows with the SGCI block; and local fusion, to transfer the SAR feature corresponding to cloudy areas with the SLFC block. Specifically, a cloudy image I and its corresponding SAR image B are first fed into different branches to extract modality-specific features $\hat{F}^0_{opt}$ and $\hat{F}^0_{sar}$ with the shallow feature extraction (SFE) block,

$$\hat{F}^0_{opt} = H_{SFE_{opt}}(I), \hat{F}^0_{sar} = H_{SFE_{sar}}(B), \tag{4}$$

where $H_{SFE_{opt}}(\cdot)$ and $H_{SFE_{SAR}}(\cdot)$ denote the functions to extract the shallow features of the cloudy image and the SAR image, respectively. Then, $\hat{F}^0_{opt}$ and $\hat{F}^0_{sar}$ are fed into D functions composited from the SGCI and SLFC block to obtain knowledgeable features with comprehensive information. More specifically, the intermediate features $\{\hat{F}^1_{opt}, \hat{F}^1_{sar}\}$, $\{\hat{F}^2_{opt}, \hat{F}^2_{sar}\}$, …, $\{\hat{F}^D_{opt}, \hat{F}^D_{sar}\}$ are obtained as

$$\hat{F}^i_{opt}, \hat{F}^i_{sar} = H_{SLFC}(H_{SGCI}(\hat{F}^{i-1}_{opt}, \hat{F}^{i-1}_{sar})), \tag{5}$$

where $H_{SGCI}(\cdot)$ and $H_{SLFC}(\cdot)$ denote the functions of the SGCI block and the SLFC block, respectively. The purpose of the SGCI block is local feature extraction and cross-window feature interaction, where the SAR feature is used to guide the relationship among all local optical windows. Each SGCI block is followed by an SLFC block, which is designed to fuse the complementary information from the corresponding area in a SAR image into the optical feature of a cloudy area. More details about these two blocks will be given in Sections 4.2 and 4.3. Finally, the high-quality cloud-free image I_C is reconstructed by aggregating all the intermediate optical features,

$$I_C = I + H_{IR}([\hat{F}^1_{opt}, \hat{F}^2_{opt}, \ldots, \hat{F}^D_{opt}]), \tag{6}$$

where H_{IR} denotes the function of cloud-free image reconstruction, and $[\hat{F}^1_{opt}, \hat{F}^2_{opt}, \ldots, \hat{F}^D_{opt}]$ refers to the concatenation of the intermediate optical features.

4.2. SAR-guided global context interaction

The SGCI block, whose detail is shown in Fig. 3, has two parallel streams for the input optical and SAR features. Each stream adopts dense connections in an approach similar to the residual dense block (RDB) (Zhang et al., 2018b), which is able to extract abundant local features via dense connected convolutional layers. As previously mentioned, SAR image clearly contributes to compensating for the missing information about cloudy regions, but not for the specific properties of optical images. Nevertheless, the cloud-free regions are conducive

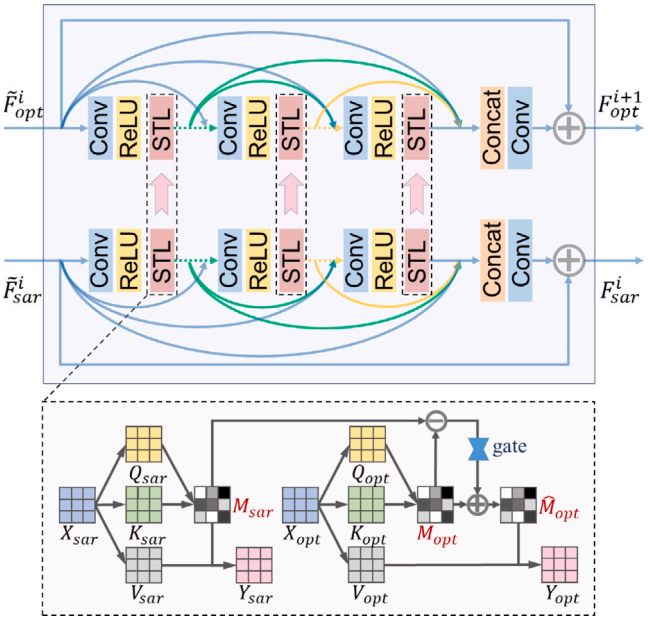


Fig. 3. Detail of the SAR-guided global context interaction (SGCI) block.

to the specific properties. The use of global texture information is necessary for the cloud removal task. Inspired by Transformer's ability to efficiently propagate information across the entire image to accumulate long-range varying contextual information, a Swin Transformer layer (STL) (Liu et al., 2021c) is added after each local convolutional layer for cross-window feature interaction.

The STL first partitions the input feature into non-overlapping $M \times M$ windows, then computes the standard self-attention separately for each window. Specifically, a local window optical/SAR feature $X_{opt}/X_{sar} \in \mathbb{R}^{M^2 \times C}$ is linearly transformed to query $Q_{opt}/Q_{sar} \in \mathbb{R}^{M^2 \times d}$, key $K_{opt}/K_{sar} \in \mathbb{R}^{M^2 \times d}$, and value $V_{opt}/V_{sar} \in \mathbb{R}^{M^2 \times d}$, where d is the dimension of the query or key. The attention weight matrix is computed as follows:

$$M_{opt} = \frac{Q_{opt}K^T_{opt}}{\sqrt{d}} + B, M_{sar} = \frac{Q_{sar}K^T_{sar}}{\sqrt{d}} + B, \tag{7}$$

where B is the learnable relative positional encoding. The essence of this attention matrix is the weight of a particular region that is

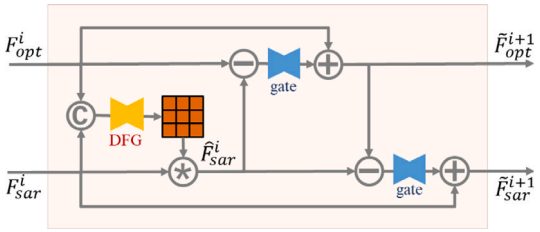


Fig. 4. Detail of the SAR-based local feature compensation (SLFC) block.

absorbing information from other regions. For a cloudy region, due to the information loss, it is difficult to estimate its interactions with cloud-free regions. For the same region in a SAR image, its interactions with other regions can be estimated easily, which provides valuable guidance for the interactions between cloud-free and cloudy regions in the optical image. Thus, we transfer the attention map of the SAR image to refine the attention map of the optical image, i.e., we use M_{sar} to improve M_{opt}. We first obtain the attention map of the optical and SAR features M_{opt} and M_{sar} by Eq. (7). Then we compute the difference between M_{opt} and M_{sar} and obtain M_{res}. Finally, we apply a gating function to adaptively refine M_{opt}:

$$\hat{M}_{opt} = M_{opt} + M_{res} \odot G(M_{res}), \tag{8}$$

where $G(\cdot)$ is the gating function fed with the residual term M_{res} and $\odot$ denotes the element-wise multiplication operation. The optical and SAR output are computed as:

$$Y_{opt} = \text{Softmax}(\hat{M}_{opt})V_{opt}, Y_{sar} = \text{Softmax}(M_{sar})V_{sar}. \tag{9}$$

This module considers the relationship among all local window optical features under the guidance of the SAR feature, denoted in this paper as global fusion.

4.3. SAR-based local feature compensation

The detail of the SLFC block is shown in Fig. 4. Because the SAR image is corrupted by severe speckle noise, we utilize dynamic filtering for SAR features before information transference. Standard convolution filters are shared across all pixels in an image, while the dynamic filters vary from pixel to pixel. Therefore, the dynamic filters can handle the spatial variance issue (Jia et al., 2016; Zhou et al., 2019), thus helping to suppress the spatially unevenly speckle noise. Specifically, a filter is dynamically generated for each position in the SAR feature using the Dynamic Filter Generation (DFG) module. The DFG module takes the concatenation of the optical and SAR features $Concat(F^i_{opt}, F^i_{sar}) \in \mathbb{R}^{H\times W\times 2C}$ as input. The dimension of the generated filter $\mathcal{F}^i$ is $H \times W \times Ck^2$ and is reshaped into a five-dimensional filter. Then, for each position (h, w, c) in the SAR feature $F^i_{sar} \in \mathbb{R}^{H\times W\times C}$, a specific local filter $\mathcal{F}^i(h, w, c) \in \mathbb{R}^{k\times k}$ is applied to the region centered around $F^i_{sar}(h, w, c)$ as

$$\hat{F}^i_{sar}(h, w, c) = \mathcal{F}^i(h, w, c) * F^i_{sar}(h, w, c), \tag{10}$$

where $*$ denotes the convolution operation.

After transforming the extracted SAR feature F^i_{sar} using the dynamic filter to improve tolerance of speckle noise, we propagate the complementary information from the SAR feature to refine the optical feature, in the same way that the attention map is refined. We compute the difference between the optical and SAR features to obtain the residual information $F^i_{s-o} = \hat{F}^i_{sar} - F^i_{opt}$, and apply a gating function to transfer the complementary information,

$$\tilde{F}^i_{opt} = F^i_{opt} + F^i_{s-o} \odot G(F^i_{s-o}). \tag{11}$$

To better exploit interactions among elements of the optical and SAR features for a further performance gain, we adopt a dual information propagation mechanism, i.e., updating the SAR feature as well. We compute the difference between the SAR feature and the updated optical feature $F^i_{o-s} = \tilde{F}^i_{opt} - \hat{F}^i_{sar}$, and also propagate the information through use of a gating function,

$$\tilde{F}^i_{sar} = F^i_{sar} + F^i_{o-s} \odot G(F^i_{o-s}). \tag{12}$$

The enhanced optical and SAR features are then introduced into the next SGCI for further representation learning. This module considers the information transference between local features, denoted as local fusion in this paper.

5. Experiments

5.1. Experimental settings

Dataset and Metrics. The experiments are conducted on the large-scale dataset SEN12MS-CR (Ebel et al., 2021), which is built from freely available data acquired by the Sentinel satellites in the Copernicus program. The dataset contains 1,22,218 samples from 169 non-overlapping regions of interest (ROI) distributed over all inhabited continents during all meteorological seasons. Each sample consists of a triplet of an orthorectified, geo-referenced Sentinel-1 dual-pol SAR image, a Sentinel-2 cloud-free multi-spectral image, and a cloud-covered Sentinel-2 multi-spectral image where the observations of cloud-free and cloud-covered images are close in time. The size of each image is 256 × 256 pixels. The VV and VH polarizations of the SAR images are clipped to values $[-25, 0]$ and $[-32.5, 0]$, and rescaled to the range $[0, 1]$. All bands of the optical images are clipped to values $[0, 10000]$, and rescaled to the range $[0, 1]$ as well. We split the 169 ROIS into 149 ROIs for training, 10 ROIs for validation, and 10 ROIs for test. To avoid overall performance being biased towards specific cloud cover level, we calculate the percentage of cloud cover of each image by utilizing the cloud detection flowchart in Meraner et al. (2020) and randomly select 800 samples from the samples with cloud cover of 0% to 20%, 20% to 40%, 40% to 60%, 60% to 80%, and 80% to 100% as the test set, respectively. Specifically, the training, validation and test set consist of 101,615, 8,623 and 4,000 samples respectively. The results of cloud removal are evaluated with the normalized data based on the peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), spectral angle mapper (SAM), and mean absolute error (MAE).
Implementation Details. The proposed GLF-CR network is implemented using publicly available Pytorch and trained in an end-to-end manner supervised by L1 loss on 4 NVIDIA TITAN V GPUs. We implement the gating functions in Sections 4.2 and 4.3 by employing a convolution layer as well as a Softmax layer, and the dynamic filter generation (DFG) module in Section 4.3 is constituted by a convolution layer followed by two residual blocks. During training, we randomly crop the samples into 128 × 128 patches. In an empirical manner, the batch size is set to 12 and the maximum epoch of training iterations is set to 30. The Adam optimizer is used and the learning rate starts at 10^{-4}, which decays by 50% every five epochs. By trading off the performance and complexity of the model, the number of the SGCI and SLFC blocks D is set to 6; the number of dense connections in each stream of the SGCI block is set to 5; the window size and the attention head number for the STL layer are set to 8 and 8, respectively; and the size of the dynamic filter k is set to 5. The codes, models, and more results are released at: https://github.com/xufangchn/GLF-CR.

5.2. Comparisons with state-of-the-art methods

We compare the proposed GLF-CR networks to state-of-the-art cloud removal methods, including multi-spectral based approaches, SpA GAN (Pan, 2020), the SAR-to-optical image translation approach, SAR2OPT (Bermudez et al., 2018), and SAR-optical data fusion based approaches, SAR-Opt-cGAN (Grohnfeldt et al., 2018), Simulation-Fusion GAN (Gao et al., 2020) and DSen2-CR (Meraner et al., 2020). SpA GAN takes all

Table 1
Quantitative comparisons of proposed GLF-Nets to state-of-the-art methods.

Method	Input		PSNR (dB) ↑	SSIM ↑	SAM (°) ↓	MAE (ρ_{TOA}) ↓
	Optical	SAR				
SpA GAN (Pan, 2020)	✓	✗	24.8688	0.7533	16.0454	0.0444
SAR2OPT (Bermudez et al., 2018)	✗	✓	25.7223	0.7918	14.0501	0.0427
SAR-Opt-cGAN (Grohnfeldt et al., 2018)	✓	✓	25.2948	0.7594	14.4389	0.0441
Simulation-Fusion GAN (Gao et al., 2020)	✓	✓	24.5519	0.6947	15.5929	0.0455
DSen2-CR (Meraner et al., 2020)	✓	✓	27.3780	0.8705	8.5073	0.0319
Concat (Ours)	✓	✓	28.5324	0.8804	8.1088	0.0284
GLF-CR (Ours)	✓	✓	**29.0793**	**0.8855**	**7.6455**	**0.0266**

channels of the input optical image as input. It uses the spatial attention network (SPANet) (Wang et al., 2019) as a generator to model the map from a cloudy image to a cloudless image. SAR2OPT performs SAR-to-optical translation by taking the U-Net as the generator, not relying on any (cloudy) optical satellite information. SAR-Opt-cGAN and DSen2-CR both leverage the SAR image as a form of prior to guide the reconstruction process under thick, optically impenetrable clouds. The SAR's channels are simply concatenated to the other channels of the input optical image to predict the full spectrum of optical bands. SAR-Opt-cGAN is extended from U-Net, while DSen2-CR is extended from the EDSR network (Lim et al., 2017). Simulation-Fusion GAN first translates the SAR image into simulated optical data, then takes the concatenation of the simulated optical image, SAR and the corrupted optical image as input for prediction.

To validate the superiority of GLF-CR in leveraging the power of SAR images, we also refer to the fusion strategy in SAR-Opt-cGAN and DSen2-CR to train the proposed network, by using concatenation, denoted as *Concat*. We concatenate the SAR's channels and optical image's channels as input, and remove the branch for SAR feature learning, the attention map update in the SGCI blocks, and the SLFC blocks. The quantitative results are presented in Table 1. The proposed GLF-CR network brings remarkable improvements compared to state-of-the-art methods. We choose 3 scenes to evaluate qualitative results, as shown in Fig. 5. For each scene, from top-left to bottom-right are respectively the cloudy image, the SAR image, the results from SpA GAN, SAR2OPT, SAR-Opt-cGAN, Simulation-Fusion GAN, DSen2-CR, *Concat* and GLF-CR, and the cloud-free image. We find that the proposed GLF-CR network achieves the best visualization performance. Detailed analyses are presented below.

We first compare the cloud removal performance of SAR-based methods to the conventional method, SpA GAN. As the SAR image encodes rich geometrical information about cloud-covered regions, it facilitates the ground object construction. SpA GAN, which relies solely on cloudy optical images, are less effective than SAR-based cloud removal methods. As shown in Fig. 5, it fails to tackle the thick cloud removal and generates undesirable artifacts, especially for cloud-covered regions.

We next compare the cloud removal performance of the SAR-to-optical image translation approach, SAR2OPT to the SAR-optical data fusion based approaches. SAR2OPT, which relies solely on SAR images, can reconstruct prominent geometric characteristics related to roads, crop fields, etc. But it suffers from content vanishing because the specific potentials and peculiarities of optical images cannot be fully compensated from the SAR images. Moreover, a distinct difference in the color distribution of SAR2OPT's reconstruction results and ground truth can be observed. SAR-Opt-cGAN adopts the same generator architecture as SAR2OPT while taking both the cloudy optical image and the SAR image as input. However, it performs worse than SAR2OPT which only takes the SAR images as input. And as shown in the second scene of Fig. 5, the SAR image clearly emphasizes the surface's physical properties. SAR-Opt-cGAN fails to reconstruct it while SAR2OPT does. It demonstrates the challenge of taking advantage of multi-modal data fusion while avoiding the performance degradation caused by the undesirable effects in each modality. Simulation-Fusion GAN suffers from the performance degradation caused by the undesirable effects in simulated optical image besides the cloudy optical and SAR images, and also has poor color fidelity. To some extent, DSen2-CR alleviates the performance degradation caused by the undesirable effects by utilizing a tailored generator. However, its gain is still limited.

Our methods perform favorably when compared with DSen2-CR, which exploits the inherent advantage of SAR image. Among them, *Concat* adopts the same approach as SAR-Opt-cGAN and DSen2-CR to utilize the complementary information embedded in SAR images. It achieves higher performance than SAR-Opt-cGAN and DSen2-CR, as shown in Table 1. Unlike the approach of SAR-Opt-cGAN and DSen2-CR, *Concat* contains global context interactions, which takes the information embedded in neighboring cloud-free regions into consideration, thus performing better in terms of global consistent structure. But those methods still leave distinct clouds or blur some image textures, which reflects the limitations of the concatenation method. Furthermore, It can be observed that the proposed GLF-CR network outperforms other methods by a large margin. It can restore images with more details and fewer artifacts, as shown in Fig. 5. These significant improvements demonstrate that the proposed method can better use the complementary information embedded in SAR images.

5.3. Analysis on different cloud cover levels

We further compare the proposed GLF-CR networks to state-of-the-art cloud removal methods on different cloud cover levels. We evaluate the performance of cloud removal on the images with cloud cover of 0% to 20%, 20% to 40%, 40% to 60%, 60% to 80%, and 80% to 100%, and show the comparison results in terms of the PSNR, SSIM, SAM, and MAE quality metrics in Fig. 6. The proposed methods perform favorably when compared with state-of-the-art methods on all cloud cover levels.

It is observed that the overall performance of multispectral-based approaches, SpA GAN, is negatively correlated with the cloud cover level. With the higher cloud cover level, they get less prior information and thus perform worse. And the performance of the SAR-to-optical image translation approach, SAR2OPT, is not related to the cloud cover level.

SAR-Opt-cGAN and Simulation-Fusion GAN utilize the prior information from both cloudy images and SAR images. It suffers the performance degradation caused by the undesirable effects in both modalities. When the cloud cover is low, it is not as good as the multispectral-based methods due to the interference from additional SAR image or simulated optical image from SAR image. When the cloud cover is high, it is not as good as the SAR-to-optical image translation approach due to the interference from clouds.

DSen2-CR alleviates the performance degradation to some extent, and thus outperforms the single-modality-based methods. *Concat* adopts the same fusion strategy in DSen2-CR to utilize the complementary information embedded in SAR images while takes the information embedded in neighboring cloud-free regions into consideration, thus its performance is more superior to that of DSen2-CR when more prior information from cloud-free regions is available. And the proposed method is superior in exploiting the power of SAR information in addition to considering the information embedded in neighboring cloud-free regions, and thus steadily outperforms DSen2-CR on all cloud cover levels.

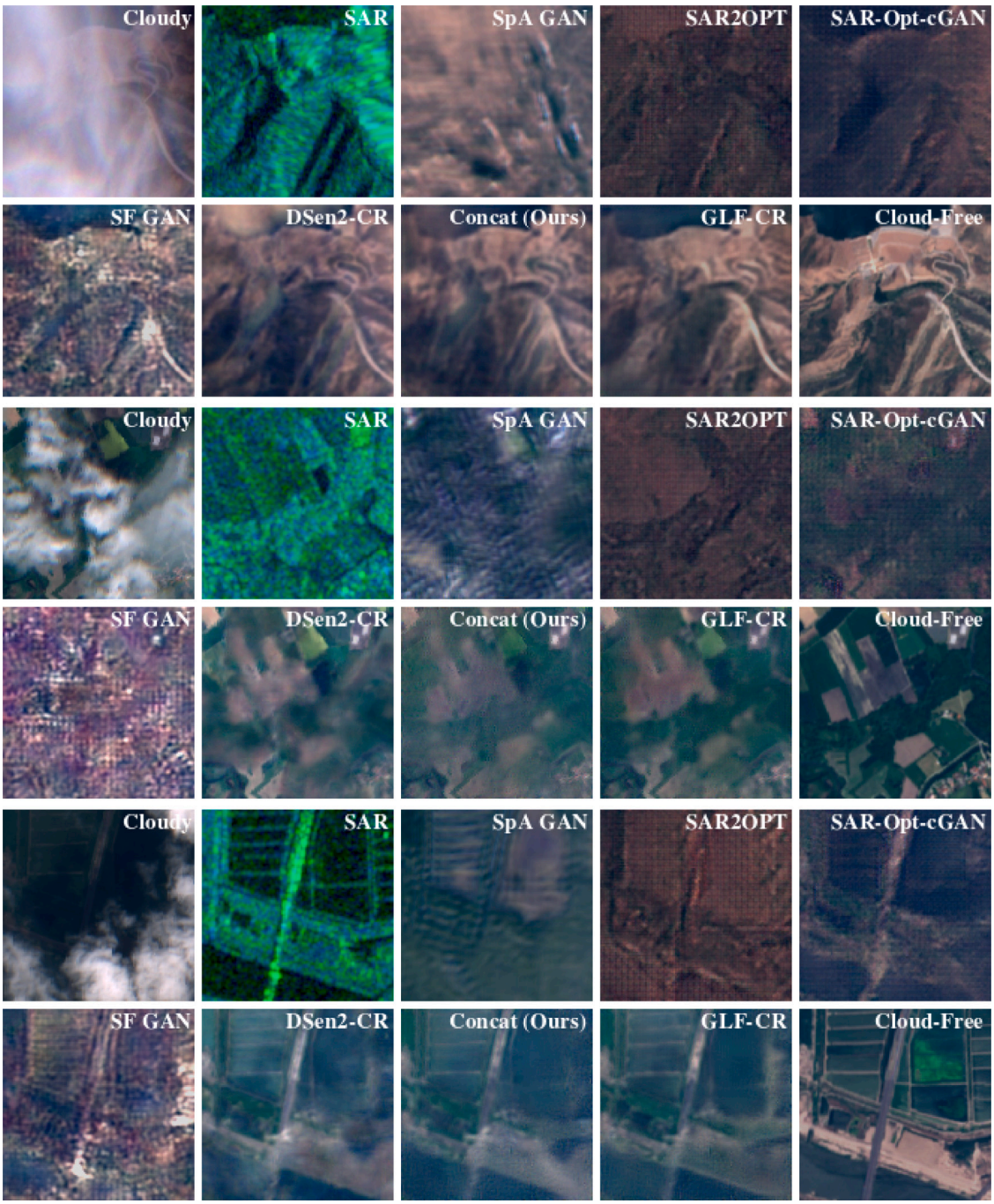


Fig. 5. Qualitative results of cloud removal for 3 different scenes. For each scene, from top-left to bottom-right are respectively the cloudy image, the SAR image, the result from SpA GAN, SAR2OPT, SAR-Opt-cGAN, Simulation-Fusion GAN (SF GAN), DSen2-CR, *Concat*, GLF-CR, and the cloud-free image. The size of each image is 128 × 128 pixels.

5.4. Ablation study

The proposed GLF-CR network improves the performance of SAR-based cloud removal by incorporating global fusion to guide the relationship among all local optical windows with SAR features and local fusion to transfer the SAR feature corresponding to cloudy areas to compensate for the missing information. To determine what contributes to the superior performance of the proposed approach, we analyze the effectiveness of each component by comparing a few variants with and without the use of SAR image (SAR), Concatenation fusion (Concat), STL layer (STL), global fusion (GF), and dynamic filter (DF). The qualitative and qualitative results are shown in Table 2 and Fig. 7, and the results on different cloud cover levels is shown in Fig. 8. From the table and the figure, we can draw the following conclusions:

Importance of SAR Image. We validate the importance of the SAR image by training the GLF-CR network without SAR images, denoted as *w/o SAR*. Since the input is a single source signal, i.e., the cloudy optical image itself, a single-stream network is adopted and no fusion strategy is used. As shown in Fig. 8, it performs comparable to the networks employing SAR images when the cloud cover level is low. However, when the cloud cover level gets higher, the performance gap between the networks with and without SAR images gets larger. And as shown in Fig. 7, *w/o SAR* tends to generate over-smoothed effects for cloud-covered regions, while the networks with SAR images can recover texture details. This demonstrates that the rich complementary

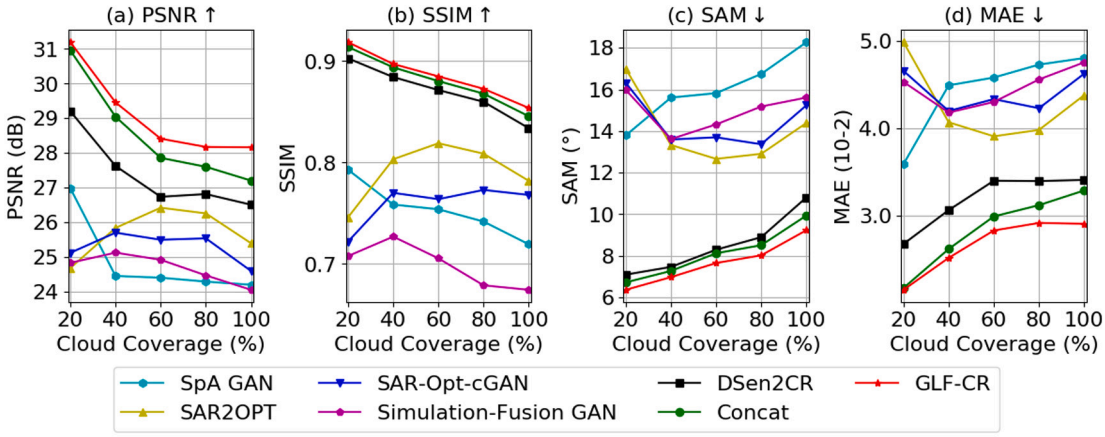


Fig. 6. Quantitative comparisons of proposed GLF-Nets to state-of-the-art methods on different cloud cover levels in terms of the PSNR, SSIM, SAM, and MAE quality metrics.

Table 2
Quantitative ablation study of proposed algorithm with and without use of the SAR image (SAR), concatenation fusion (Concat), STL layer (STL), global fusion (GF), and dynamic filter (DF).

Method	PSNR (dB) ↑	SSIM ↑	SAM (°) ↓	MAE (ρ_{TOA}) ↓
w/o SAR	28.3657	0.8759	8.1783	0.0299
Concat	28.5324	0.8804	8.1088	0.0284
w/o STL	28.5079	0.8825	8.1783	0.0287
w/o GF	28.4983	0.8816	8.0595	0.0287
w/o DF	28.2867	0.8800	7.9853	0.0297
GLF-CR	**29.0793**	**0.8855**	**7.6455**	**0.0266**

information encoded in SAR images can effectively improve the cloud removal performance.

Limitation of Concatenation Fusion. Compared with *w/o SAR*, *Concat* only adds two channels to the input to utilize the SAR image. The gain of utilizing the concatenation fusion is 0.17 dB, while the proposed GLF-CR network obtains a gain of 0.71 dB. As observed from Fig. 8, when the proportion of cloud-free regions is higher, the performance gap between *Concat* and *GLF-CR* is larger, since the proposed GLF-CR network can better exploit the power of SAR information compared with the concatenation fusion. Fig. 7 shows that the proposed GLF-CR network can recover more complete texture structure and obtain better visual effects.

Effectiveness of Global Interactions. Capturing global interactions between contexts plays a vital role in maintaining global consistent structure. We train the GLF-CR network by removing the STL layers in the SGCI blocks, denoted as *w/o STL*. It can be observed that the proposed GLF-CR method which captures the global interactions between contexts can improve cloud removal performance effectively. It recovers clearer and more complete structure for the land in the second and fourth scenes in Fig. 7.

Effectiveness of SAR-Guided Global Interactions. We further validate the effectiveness of guiding the global interactions of optical features with SAR features. We train the GLF-CR network by reserving the STL layer but not using the SAR feature to guide the global optical interactions, denoted as *w/o GF*. Compared with *w/o STL*, it can be observed that *w/o GF* has only a slight performance improvement in terms of SAM, despite using additional STL layers to maintain the spatial consistency, since estimating the interactions from the cloudy optical image itself will introduce some error. As shown in the third scene in Fig. 7, *w/o GF* generates undesirable artifacts. As the SAR image is not affected by cloud cover, it can provide valuable guidance for capturing global interactions between contexts. This point can be validated by comparing the results of *w/o GF* and *GLF-CR*. It can be seen that guiding the global interactions of optical features with SAR features can effectively improve the performance of cloud removal and make the structure of the predicted cloud-free image more consistent with ground truth.

Effectiveness of the Dynamic Filter. The proposed GLF-CR network uses dynamic filtering to handle the speckle noise of SAR images. To validate the effectiveness of the dynamic filter, we train the GLF-CR network by removing the dynamic filter in SLFC blocks, denoted as *w/o DF*. It can be seen from Fig. 8 that the performance of *w/o DF* degrades more severely in terms of PSNR and MAE that measure the quality of reconstructed images than in terms of SSIM and SAM that quantify spectral and structural similarity. And it can be observed that the trends of *w/o DF* and *GLF-CR* relative to the cloud cover level are similar. As both methods adopt the same strategy to utilize the information of the cloud-free regions and SAR images, while the proposed GLF-CR network can alleviate the problem of speckle noise in the SAR image and generate clearer images.

6. Discussion

Performance on Challenging Conditions. Cloud removal is quite challenging when the image to be processed is completely cloudy. To see how the proposed method behaves in the challenging conditions, Fig. 9 shows the results on the images where the ground information is almost obscured by clouds. It can be found that the proposed method can recover the approximate information of ground objects while with poor texture details. Since the images are completely cloudy, no cloud-free part can be accessed and only SAR information is available to reconstruct the cloud-free images. The quality of reconstructed cloud-free images depends entirely on the information embedded in the SAR image. While the SAR image fails to feature the different agricultural landscapes, as seen in the first scene in Fig. 9, the reconstructed cloud-free image loses the corresponding details. And since no spectral information is available, the spectral fidelity of the reconstructed cloud-free image degrades.

Speckle Noise in SAR Data. The SAR data in SEN12MS-CR dataset is from the Level-1 GRD product, which has been multi-looked for reduced speckle. Notwithstanding, the multi-looked data still exhibits a high degree of speckle noise, as seen from Figs. 1, 5 and 7, since speckle noise is multiplicative in nature and difficult to distinguish from the original signal. And, while commonly referred to as “speckle noise”, speckle is not only noise but in some sense has an information content (Argenti et al., 2013). At this point, we do not consider an explicit despeckling preprocessing step, but implicitly handle the spatially varying speckle distribution by the dynamic filter embedded in the network. It is also possible to preprocess the SAR data with a despeckling technique before feeding it to the network. Therefore,

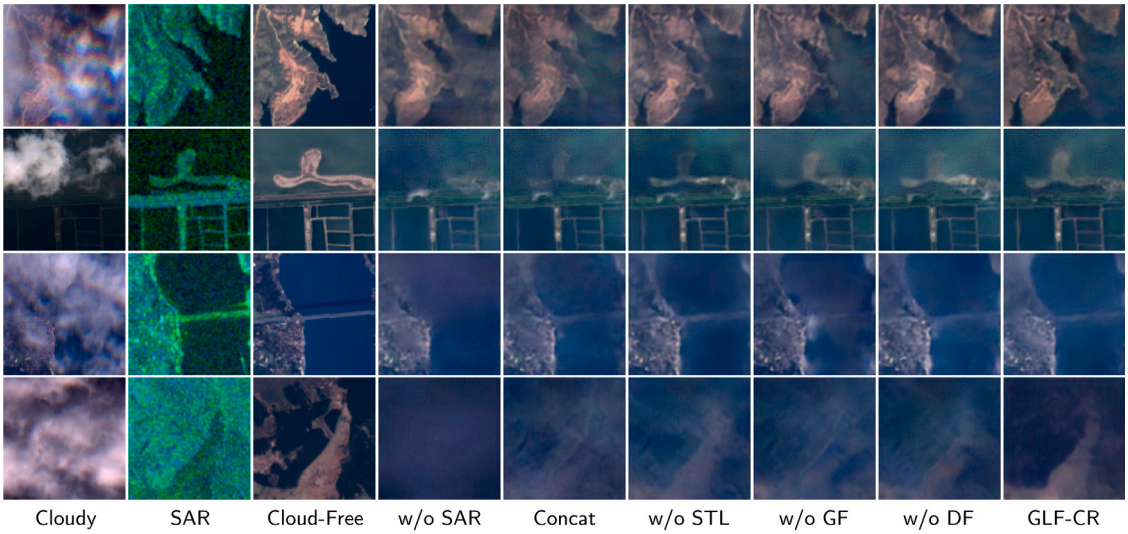


Fig. 7. Qualitative ablation study with 4 scenes by different GLF-CR networks. For each scene, from left to right, are respectively the cloudy image, the SAR image, the cloud-free image, and the result by *w/o SAR*, *Concat*, *w/o STL*, *w/o GF*, *w/o DF*, and GLF-CR. The size of each image is 128 × 128 pixels.

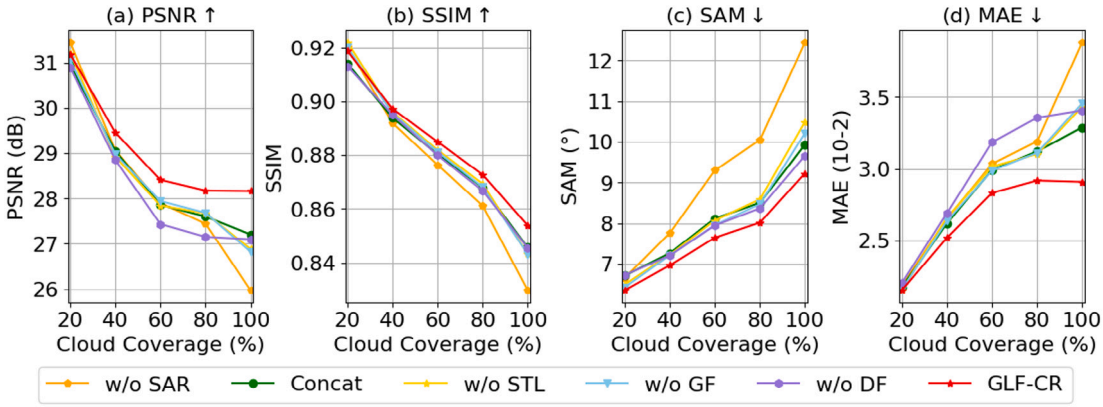


Fig. 8. Quantitative ablation study on different cloud cover levels in terms of the PSNR, SSIM, SAM and MAE quality metrics.

we train the GLF-CR network by removing the dynamic filter in SLFC blocks while feeding the SAR data despeckled with a median filter. As shown in Table 3, we can see that preprocessing the SAR data with a despeckling technique can reduce the influence of speckle noise on cloud removal. While the proposed method implicitly mitigates the influence of speckle noise based on the dynamic filter and can achieve better performance.

Geometric Distortion in SAR Data. It is well-known that there is an inherent geometric distortion in SAR data when the terrain is undulating, due to the sensor's sideways view. It will lead to the inconsistency between the information in the SAR data and the actual state of the ground objects, adversely affecting the cloud removal performance. The experiments in this paper are conducted on the SEN12MS-CR dataset (to our best knowledge, the only open-source cloud removal dataset with SAR data), where the SAR data is provided by the Sentinel-1 satellites. Its resolution is $10m$ and thus does not show excessive distortion. Furthermore, depending on the large scale of the dataset, the proposed powerful model can address this aspect to some extent.

Registration error between the optical and SAR Data. The registration error between the optical image and its corresponding SAR image is expected to affect the learning process. The data instructions given by ESA illustrate that the Sentinel-1 SAR L1 productions and the Sentinel-2 optical L1C productions have a co-registration accuracy of within 2 pixels. We set the size of the dynamic filter in the SLFC blocks to 5 for a larger receptive field, which allows the proposed model to work when tiny deviations exist between the SAR and optical images.

Table 3
Performance of proposed algorithm with use of despeckled SAR data.

Method	PSNR (dB) ↑	SSIM ↑	SAM (°) ↓	MAE (ρ_{TOA}) ↓
w/ despeckled SAR	28.5377	0.8818	8.0719	0.0286
w/o DF	28.2867	0.8800	7.9853	0.0297
GLF-CR	29.0793	0.8855	7.6455	0.0266

Nuisances between Cloudy Reference Image and Cloud-Free Target Image. The cloud removal performance in the paper is assessed on the SEN12MS-CR dataset by comparing the prediction with the cloud-free image temporally close to the cloudy one. There are some inevitable nuisances determined by the sunlight condition, acquisition geometry, humidity, pollution, change of landscape, etc, while the SEN12MS-CR dataset is curated to minimize such cases. However, the inevitable nuisances are negligible for a relatively large-scale test split that is globally and seasonally sampled without any bias to specific sunlight condition, etc. It implies that models biased to specific condition will not have any unfair advantages on the test split. Overall, the influence of nuisances can be averaged out. It poses no concern about the fairness of benchmarking the proposed model on the considered dataset.

In addition, we test the proposed method on images where the interval between the cloud-free and cloudy image is different. The date

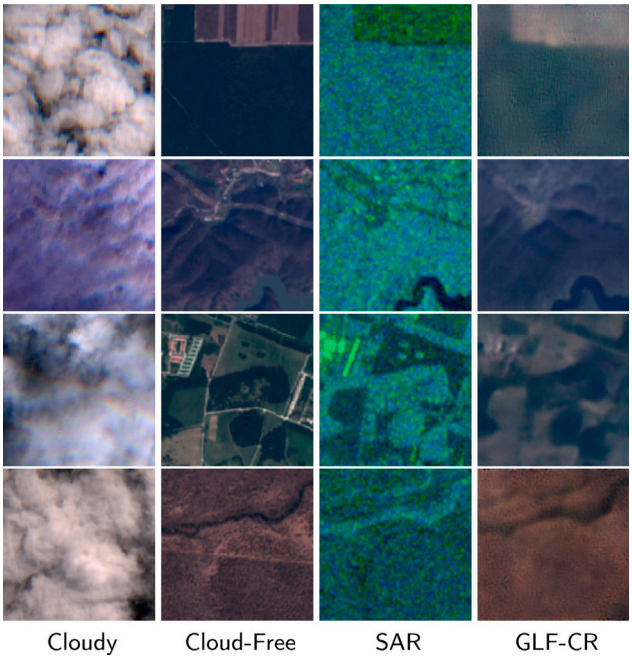


Fig. 9. Example results of GLF-CR on the images completely obscured by clouds.

Table 4
Evaluating cloud removal performance using the cloud-free images with different intervals from cloudy images.

Interval	Method	PSNR (dB) ↑	SSIM ↑	SAM (°) ↓	MAE (ρ_{TOA}) ↓
13 days	DSen2-CR	27.6299	0.8618	6.9426	0.0293
	GLF-CR	28.6470	0.8707	6.9005	0.0260
25 days	DSen2-CR	26.3796	0.8403	8.0728	0.0334
	GLF-CR	26.9173	0.8444	8.6355	0.0317
72 days	DSen2-CR	25.1544	0.8247	10.0843	0.0382
	GLF-CR	25.3110	0.8324	10.6852	0.0378

of input cloudy image is July 17, 2018, and we use the SAR image with the closest interval to cloudy image as auxiliary data, whose date is July 18, 2018. And the date of cloud-free images used for the assessment are July 30, 2018, August 11, 2018 and September 28, 2018, respectively. The results are shown in Table 4. We can observe that the proposed method performs better than the best baseline DSen2-CR overall, which is consistent with the results on the SEN12MS-CR dataset. It shows the feasibility of assessing the performances with temporally close cloud-free images. And we can observe that, when the interval between the reference cloud-free image used to calculate the value of the metrics and the cloudy image is larger, the methods performs worse in terms of the metrics. It indicates that the method is able to restore the surface information of the input cloudy image, and thus the cloud-free image with the larger interval to input cloudy image has less reference value.

Strict ground truth correspondence may only be guaranteed by generating synthetic cloud coverage superimposed on cloud-free observations, as done in Enomoto et al. (2017) and Gao et al. (2020). However, the experimental results in Ebel et al. (2020) has indicated that popular synthetic cloud simulation techniques suffer from severe limitations in approximation to the real data. The great performance on synthetic data may not necessarily translate to equal performance on real data. Hence we follow the approach of using real observations, despite acknowledgeable shortcomings at other ends.

7. Conclusion

In this work, we propose a novel global–local fusion based cloud removal (GLF-CR) algorithm for high quality cloud-free image reconstruction. It boosts cloud removal performance from two aspects, on the one hand, it guides the relationship among all local optical windows with the SAR feature to fully utilize the spatial consistency between the cloudy and the neighboring cloud-free regions, and on the other hand, it enhances the utilization of SAR data to compensate for missing information while alleviating the performance degradation caused by speckle noise. Extensive experiments demonstrate that the power of the information embedded in neighboring cloud-free regions and corresponding SAR data over different cloud cover levels. The proposed method can achieve state-of-the-art performance on all different cloud cover levels.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

F. Xu is supported by the China Scholarship Council (CSC). The work of W. Yang is supported by the National Natural Science Foundation of China (NSFC) under Grant 61771351. The work of X. Zhu is jointly supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. [ERC-2016-StG-714087], Acronym: *So2Sat*), by the Helmholtz Association through the Helmholtz Excellent Professorship "Data Science in Earth Observation - Big Data Fusion for Urban Research" (grant number: W2-W3-100), by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001) and by German Federal Ministry of Economics and Technology in the framework of the "national center of excellence ML4Earth" (grant number: 50EE2201C).

References

Argenti, F., Lapini, A., Bianchi, T., Alparone, L., 2013. A tutorial on speckle reduction in synthetic aperture radar images. IEEE Geosci. Remote Sens. Mag. 1 (3), 6–35.

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. ISPRS J. Photogramm. Remote Sens. 140, 20–32.

Bamler, R., 2000. Principles of synthetic aperture radar. Surv. Geophys. 21 (2), 147–157.

Bermudez, J.D., Happ, P.N., Oliveira, D.A.B., Feitosa, R.Q., 2018. SAR to optical image synthesis for cloud removal with generative adversarial networks. ISPRS Ann. Photogramm., Remote Sens. Spat. Inf. Sci. IV-1, 5–11.

Chan, T.F., Shen, J., 2001. Nontexture inpainting by curvature-driven diffusions. J. Vis. Commun. Image Represent. 12 (4), 436–449.

Ebel, P., Meraner, A., Schmitt, M., Zhu, X.X., 2021. Multisensor data fusion for cloud removal in global and all-season sentinel-2 imagery. IEEE Trans. Geosci. Remote Sens. 59 (7), 5866–5878.

Ebel, P., Schmitt, M., Zhu, X.X., 2020. Cloud removal in unpaired sentinel-2 imagery using cycle-consistent GAN and SAR-optical data fusion. In: IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 2065–2068.

Ebel, P., Xu, Y., Schmitt, M., Zhu, X.X., 2022. SEN12MS-CR-ts: A remote-sensing data set for multimodal multitemporal cloud removal. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Enomoto, K., Sakurada, K., Wang, W., Fukui, H., Matsuoka, M., Nakamura, R., Kawaguchi, N., 2017. Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 48–56.

Fu, K., Fan, D.P., Ji, G.P., Zhao, Q., 2020. JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3052–3062.

Fuentes Reyes, M., Auer, S., Merkle, N., Henry, C., Schmitt, M., 2019. SAR-to-optical image translation based on conditional generative adversarial networks—optimization, opportunities and limits. Remote Sens. 11 (17), 2067.

Gao, J., Yi, Y., Wei, T., Zhang, G., 2021. Sentinel-2 cloud removal considering ground changes by fusing multitemporal SAR and optical images. Remote Sens. 13 (19), 3998.

Gao, J., Yuan, Q., Li, J., Zhang, H., Su, X., 2020. Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks. Remote Sens. 12 (1), 191.

Girard, N., Smirnov, D., Solomon, J., Tarabalka, Y., 2021. Polygonal building extraction by frame field learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5891–5900.

Grohnfeldt, C., Schmitt, M., Zhu, X., 2018. A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from sentinel-2 images. In: IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 1726–1729.

Hazirbas, C., Ma, L., Domokos, C., Cremers, D., 2016. FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In: Asian Conference on Computer Vision. Springer, pp. 213–228.

Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V., 2016. Dynamic filter networks. In: Advances in Neural Information Processing Systems, vol. 29, pp. 667–675.

King, M.D., Platnick, S., Menzel, W.P., Ackerman, S.A., Hubanks, P.A., 2013. Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites. IEEE Trans. Geosci. Remote Sens. 51 (7), 3826–3852.

Li, J., Wu, Z., Hu, Z., Zhang, J., Li, M., Mo, L., Molinier, M., 2020. Thin cloud removal in optical remote sensing images based on generative adversarial networks and physical model of cloud distortion. ISPRS J. Photogramm. Remote Sens. 166, 373–389.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R., 2021. SwinIR: Image restoration using Swin Transformer. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1833–1844.

Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K., 2017. Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 136–144.

Liu, L., Chen, J., Wu, H., Li, G., Li, C., Lin, L., 2021a. Cross-modal collaborative representation learning and a large-scale RGBT benchmark for crowd counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4823–4833.

Liu, L., Lei, B., 2018. Can SAR images and optical images transfer with each other? In: IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 7019–7022.

Liu, S., Lei, Y., Zhang, L., Li, B., Hu, W., Zhang, Y.-D., 2021b. MRDDANet: A multiscale residual dense dual attention network for SAR image denoising. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021c. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10012–10022.

Maalouf, A., Carré, P., Augereau, B., Fernandez-Maloigne, C., 2009. A bandelet-based inpainting technique for clouds removal from remotely sensed images. IEEE Trans. Geosci. Remote Sens. 47 (7), 2363–2371.

Meraner, A., Ebel, P., Zhu, X.X., Schmitt, M., 2020. Cloud removal in sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. ISPRS J. Photogramm. Remote Sens. 166, 333–346.

Pan, H., 2020. Cloud removal for remote sensing imagery via spatial attention generative adversarial network. arXiv:2009.13015.

Requena-Mesa, C., Benson, V., Reichstein, M., Runge, J., Denzler, J., 2021. EarthNet2021: A large-scale dataset and challenge for Earth surface forecasting as a guided video prediction task. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1132–1142.

Scarpa, G., Gargiulo, M., Mazza, A., Gaetano, R., 2018. A CNN-based fusion method for feature extraction from sentinel data. Remote Sens. 10 (2), 236.

Schmitt, M., Tupin, F., Zhu, X.X., 2017. Fusion of SAR and optical remote sensing data—Challenges and recent trends. In: IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 5458–5461.

Shen, H., Li, X., Zhang, L., Tao, D., Zeng, C., 2013. Compressed sensing-based inpainting of aqua moderate resolution imaging spectroradiometer band 6 using adaptive spectrum-weighted sparse Bayesian dictionary learning. IEEE Trans. Geosci. Remote Sens. 52 (2), 894–906.

Shen, H., Wu, J., Cheng, Q., Aihemaiti, M., Zhang, C., Li, Z., 2019. A spatiotemporal fusion based cloud removal method for remote sensing images with land cover changes. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12 (3), 862–874.

Singh, P., Komodakis, N., 2018. Cloud-GAN: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks. In: IEEE International Geoscience and Remote Sensing Symposium. pp. 1772–1775.

Sun, T., Di, Z., Che, P., Liu, C., Wang, Y., 2019. Leveraging crowdsourced GPS data for road extraction from aerial imagery. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7509–7518.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30.

Wang, L., Wang, Y., Lin, Z., Yang, J., An, W., Guo, Y., 2021. Learning a single network for scale-arbitrary super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4801–4810.

Wang, T., Yang, X., Xu, K., Chen, S., Zhang, Q., Lau, R.W., 2019. Spatial attentive single-image deraining with a high quality real rain dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12270–12279.

Wen, X., Pan, Z., Hu, Y., Liu, J., 2021. Generative adversarial learning in YUV color space for thin cloud removal on satellite imagery. Remote Sens. 13 (6), 1079.

Wu, A., Han, Y., 2018. Multi-modal circulant fusion for video-to-language and backward. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 1029–1035.

Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2018. DOTA: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3974–3983.

Xu, M., Pickering, M., Plaza, A.J., Jia, X., 2015. Thin cloud removal based on signal transmission principles and spectral mixture analysis. IEEE Trans. Geosci. Remote Sens. 54 (3), 1659–1669.

Xu, F., Yu, L., Wang, B., Yang, W., Xia, G.-S., Jia, X., Qiao, Z., Liu, J., 2021. Motion deblurring with real events. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2583–2592.

Yu, Z., Wang, W., Li, C., Liu, W., Yang, J., 2018. Speckle noise suppression in SAR images using a three-step algorithm. Sensors 18 (11), 3643.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y., 2018a. Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision. pp. 286–301.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y., 2018b. Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2472–2481.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y., 2020. Residual dense network for image restoration. IEEE Trans. Pattern Anal. Mach. Intell. 43 (7), 2480–2495.

Zhang, Q., Yuan, Q., Li, Z., Sun, F., Zhang, L., 2021. Combined deep prior with low-rank tensor SVD for thick cloud removal in multitemporal images. ISPRS J. Photogramm. Remote Sens. 177, 161–173.

Zheng, C., Cham, T.J., Cai, J., 2019. Pluralistic image completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1438–1447.

Zhou, S., Zhang, J., Pan, J., Xie, H., Zuo, W., Ren, J., 2019. Spatio-temporal filter adaptive network for video deblurring. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2482–2491.

Zhu, X.X., Montazeri, S., Ali, M., Hua, Y., Wang, Y., Mou, L., Shi, Y., Xu, F., Bamler, R., 2021. Deep learning meets SAR: Concepts, models, pitfalls, and perspectives. IEEE Geosci. Remote Sens. Mag. 9 (4), 143–172.

Zi, Y., Xie, F., Song, X., Jiang, Z., Zhang, H., 2022. Thin cloud removal for remote sensing images using a physical-model-based CycleGAN with unpaired data. IEEE Geosci. Remote Sens. Lett. 19, 1–5. http://dx.doi.org/10.1109/LGRS.2021.3140033.

A.4 SEN12MS-CR-TS: A Remote-Sensing Data Set for Multimodal Multitemporal Cloud Removal

Reference: P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu. *SEN12MS-CR-TS: A remote sensing data set for multimodal multitemporal cloud removal.* IEEE Transactions on Geoscience and Remote Sensing, 60:1–14, 2022

SEN12MS-CR-TS: A Remote-Sensing Data Set for Multimodal Multitemporal Cloud Removal

Patrick Ebel, Yajin Xu, Michael Schmitt, *Senior Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— About half of all optical observations collected via spaceborne satellites are affected by haze or clouds. Consequently, cloud coverage affects the remote-sensing practitioner's capabilities of a continuous and seamless monitoring of our planet. This work addresses the challenge of optical satellite image reconstruction and cloud removal by proposing a novel multimodal and multitemporal data set called SEN12MS-CR-TS. We propose two models highlighting the benefits and use cases of SEN12MS-CR-TS: First, a multimodal multitemporal 3-D convolution neural network that predicts a cloud-free image from a sequence of cloudy optical and radar images. Second, a sequence-to-sequence translation model that predicts a cloud-free time series from a cloud-covered time series. Both approaches are evaluated experimentally, with their respective models trained and tested on SEN12MS-CR-TS. The conducted experiments highlight the contribution of our data set to the remote-sensing community as well as the benefits of multimodal and multitemporal information to reconstruct noisy information. Our data set is available at *https://patrickTUM.github.io/cloud_removal*.**

***Index Terms*— Cloud removal, data fusion, image reconstruction, sequence-to-sequence, synthetic aperture radar (SAR)-optical, time series.**

I. Introduction

THE majority of our planet's land surface is covered by haze or clouds [1]. Such atmospheric distortions impede the capability of spaceborne optical satellites to reliably and seamlessly record noise-free data of the earth's surface. The presence of clouds is detrimental to typical remote-sensing applications, for instance, land cover classification [2], semantic segmentation [3], [4], and change detection [5], [6].

Hence, the need for cloud-free earth observation gave rise to a rapidly growing number of haze and cloud removal methods [3], [7]–[14]. Most previous methods focus on a multimodal approach [8], [13]–[15] to reconstruct cloud-covered pixels via information translated from synthetic aperture radar (SAR) or other sensors more robust to atmospheric disturbances [16], yet focus on only a single time point of observations. In comparison, recent models attempt a temporal reconstruction of cloudy observations by means of inference across time series [12], [17], [18], utilizing the circumstance that the extent of cloud coverage over a particular region is variable over time and seasons [1].

The work at hand aims to combine both preceding approaches and thus considers the challenge of cloud removal in optical satellite imagery by integrating information across time and within different modalities. For this purpose, we curate a new data set called SEN12MS-CR-TS, which contains multitemporal and multimodal satellite observations. Specifically, SEN12MS-CR-TS consists of 1-year long time series of coregistered radar Sentinel-1 (S1) as well as multispectral Sentinel-2 observations (S2) acquired in a paired manner, covering regions of interest (ROIs) from all over the world. We highlight the benefits of the proposed data set by training and testing two different models on our data set: First, a multimodal multitemporal 3-D-Convolution Neural Network that predicts a cloud-free image from a sequence of cloudy optical and radar images. Second, a sequence-to-sequence translation model that predicts a cloud-free time series from a cloud-covered time series. Both approaches are evaluated experimentally, with their respective models trained and tested on SEN12MS-CR-TS. Exemplary outcomes are highlighted in Fig. 1. The conducted experiments highlight the contribution of our curated data set to the remote-sensing community as well as the benefits of multimodal and multitemporal information to reconstruct noisy information.

A. Related Work

As the presence of clouds in optical satellite imagery poses a severe hindrance for remote-sensing applications, there has been plenty of preceding research on cloud removal methods [3], [7]–[10], [12]–[14], [20]. The focus of this overview is on data sets for cloud removal methods. Much of the

Manuscript received September 18, 2021; revised January 1, 2022; accepted January 20, 2022. Date of publication January 25, 2022; date of current version March 17, 2022. This work was supported in part by the Federal Ministry for Economic Affairs and Energy of Germany in the Project "AI4Sentinels—Deep Learning for the Enrichment of Sentinel Satellite Imagery" under Grant FKZ50EE1910. The work of Xiao Xiang Zhu was jointly supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme under Grant ERC-2016-StG-714087, Acronym: *So2Sat*, in part by the Helmholtz Association through the Framework of Helmholtz AI under Grant ZT-I-PF-5-01—Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)" and Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research" under Grant W2-W3-100, in part by the German Federal Ministry of Education and Research (BMBF) in the Framework of the International Future AI Laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001, and in part by the German Federal Ministry of Economics and Technology in the Framework of the "National Center of Excellence ML4Earth" under Grant 50EE2201C. *(Corresponding author: Xiao Xiang Zhu.)*

Patrick Ebel and Yajin Xu are with Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: patrick.ebel@tum.de).

Michael Schmitt is with the German Aerospace Center, Remote Sensing Technology Institute, 82234 Wessling, Germany, and also with the Chair of Earth Observation, Bundeswehr University Munich, 85577 Neubiberg, Germany (e-mail: michael.schmitt@unibw.de).

Xiao Xiang Zhu is with Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany, and also with the German Aerospace Center, Remote Sensing Technology Institute, 82234 Wessling, Germany (e-mail: xiaoxiang.zhu@dlr.de).

Digital Object Identifier 10.1109/TGRS.2022.3146246

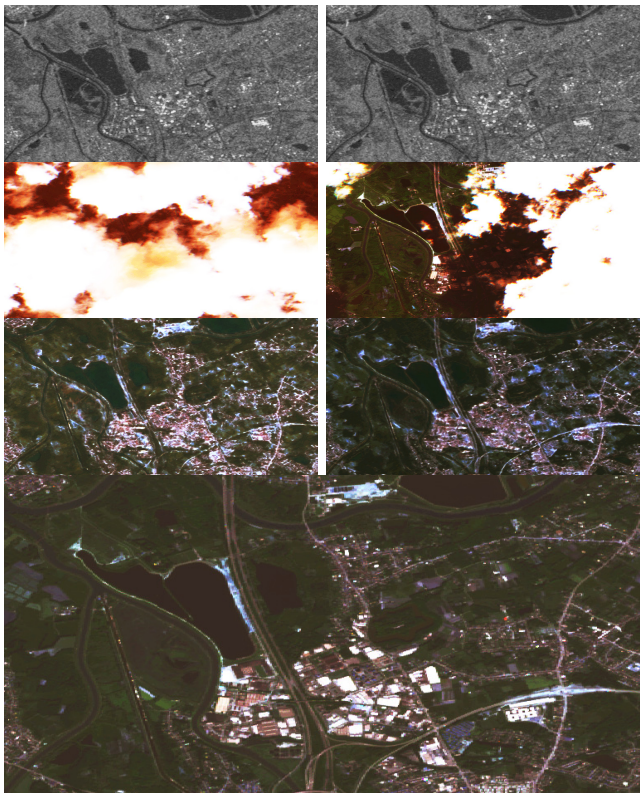


Fig. 1. Example observations and cloud-free predictions. Columns: Samples at two different time points. Rows: S1 data (in grayscale), cloudy S2 data (in RGB), predicted cloud-free $\hat{S}2$ data, and reference cloud-free S2 data of a later point in time. The results highlight that our network is able to integrate multimodal and multitemporal information to predict a clear-view sequence of multispectral observations, even in the presence of heavy cloud coverage.

early work on cloud removal considered data of simulating cloudy observations [3]. Copying cloudy pixel values from one image to another clear-view one [3] captures the spectral properties of naturally cloudy observations more faithfully than synthetic noise (e.g., Perlin noise [21]) [7], [8], [20], but neither precisely reproduce the statistics of satellite images containing natural cloud occurrences [14]. Consequently, our data set contain cloud-free as well as naturally occurring cloud-covered optical satellite recordings. The SEN12MS-CR data set [14] provides a globally distributed collection of coregistered mono-temporal Sentinel-1 as well as cloudy and cloud-free Sentinel-2 observations. Our data set is an extension of SEN12MS-CR in the sense that we collect repeated measures per ROI and therefore provide a time series of coregistered S1 and S2 observations, gathered such that matched observations of both modalities are no more than two weeks apart. In comparison to the preceding data set, ours allows integrating information not solely across different sensors, but also across different points in time distributed throughout the year. Similarly, the work of [12] allows for time-series cloud removal by providing a collection of tri-temporal RGB (NIR)-channel optical data and corresponding models. Our contribution extends this work by providing true multimodal data recorded by two distinct sensors, SAR Sentinel-1 measurements, as well as 13-band multispectral Sentinel-2 observations. Furthermore, the length of each time series is increased considerably, from 3 to 30 samples. Finally, [12] exclude observations with greater than 30% cloud coverage from their data set, which deviates from real conditions. Our approach aims to model the complete spectrum of cloud coverage, including conditions commonly encountered by remote-sensing practitioners. In sum, our work and its main contribution, a large-scale multimodal multitemporal data set for cloud removal in optical satellite imagery, build on a history of research and improve upon the current state of image reconstruction in remote sensing by providing a novel, carefully curated data set.

II. Data

This work introduces SEN12MS-CR-TS, a multimodal and multitemporal data set for training and evaluating global and all-season cloud removal methods. The data set consists of 53 globally distributed ROI, curated as detailed in Section II-A. The ROIs are over 4000×4000 px^2 each, covering about 40×40 km^2 of land such that the total surface area covered by the data set is over 80000 km^2. Of all collected ROI, 40 are defined as a training split and 13 as a hold-out split to evaluate cloud removal approaches on. For every ROI, we collect 30 coregistered and paired S1 and S2 full-scene images evenly spaced in time throughout the year of 2018. Each acquired image is inspected and quality-controlled manually. The spatial distribution of all ROI is depicted in Fig. 2 and highlights the global sampling of our data set. The empirical distribution of the cloud coverage of all optical observations (examples are shown in Fig. 3) is computed as detailed in Section II-C and the statistics are presented in Figs. 4 and 5 for the train and the test splits, respectively. The cloud-free Sentinel-2 (RGB-channel) observations of four example ROI illustrating the diversity of our data set are illustrated in Fig. 6. Importantly, the data set is curated without excluding any interval of cloud coverage such that the collected observations also reflect conditions of high cloud coverage as commonly encountered in practice [1]. The data is made available under *https://patrickTUM.github.io/cloud_removal*. It is about 2 Tb in size and compatible with the SEN12MS-CR data set [14]. That is, no train ROI of SEN12MS-CR is part of our data set's test ROI and vice versa.

A. Data Collection

All curated data are recorded via the SAR Sentinel-1 and multispectral Sentinel-2 (level 1-C top-of-atmosphere reflectance product) instruments of European Space Agency's (ESA's) Copernicus mission. The recorded observations are acquired via Google Earth Engine [22] and a custom semiautomatic processing pipeline. We randomly sample the geospatial locations of 53 ROIs from SEN12MS-CR [14]. To minimize mosaicing, observations of cells covered by a single pass are collected. The samples are referenced within the World Geodetic System 1984 (WGS84) coordinate system. For every ROI, 30 time intervals are evenly spaced throughout the year of 2018. For every time interval, a coregistered, geo-referenced, and full-scene S1 image as well as a paired full-scene S2

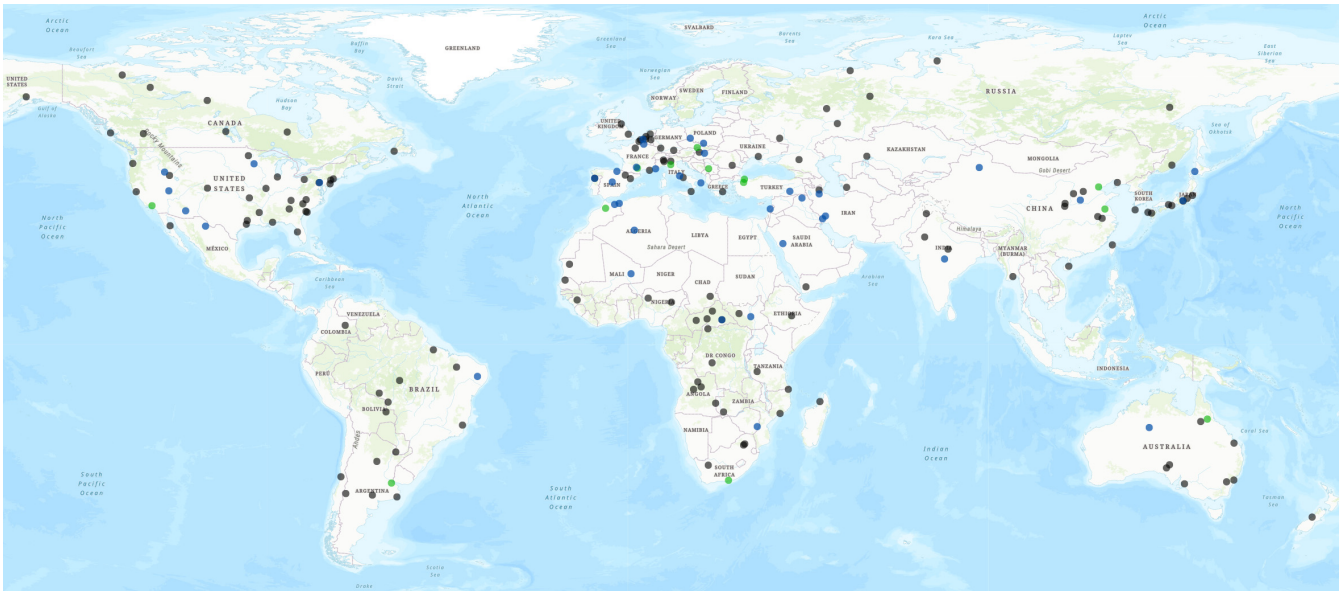


Fig. 2. Spatial distribution of the ROI constituting SEN12MS-CR-TS. Areas belonging to the training split are denoted in blue, and regions of the testing split are colored in green. The ROIs of SEN12MSCR [14], nonoverlapping and compatible with our data set, are depicted in gray. Any graphical overlap of the semitransparently plotted dots is rendered in darker tones so close-by dots can easier be discerned.

image (level 1-C) are collected. The acquisition within the same interval window is such that corresponding multimodal images are no more than two weeks apart. Further statistics regarding the pairing of observations are provided in appendix.

B. Preprocessing

To prepare the collected raw data and translate it into a format that neural networks for cloud removal can handle the following preprocessing steps are taken: Each band of every observation is upsampled to 10-m resolution (i.e., to the native resolution of Sentinel-2's bands 2, 3, 4, and 8). Every full-scene image is sliced into nonoverlapping patches of dimensions 256×256 px^2. The S1 observations are processed via the Sentinel-1 toolbox [23] (including border and thermal noise removal, radiometric calibration, and orthorectification) and decibel-transformed. An example patch-wise tuple of paired S1 and S2 data is illustrated in Fig. 3. Input patches to any ResNet model [24] are preprocessed in line with the pipeline of [13] as follows: the vertical-vertical (VV) and vertical-horizontal (VH) channels of S1 observations are value-clipped in the ranges $[-25; 0]$, $[-32.5; 0]$ and rescaled to the interval $[0; 2]$, while S2 patches are value-clipped to $[0; 10000]$ and normalized to the range $[0; 5]$. For all other networks with a different backbone architecture, preprocessing is done as follows: each patch is value-clipped and then rescaled for every pixel to take normalized values within the unit range of $[0, 1]$. The modalities S1 and S2 are value-clipped within the intervals of $[-25; 0]$ and $[0; 10000]$, respectively. This way, we follow the preprocessing protocol of the preceding work and avoid any unnecessary adjustments, for the sake of simplicity. For evaluation, the pixel values of all input patches, target images, and predictions are remapped to the unit interval $[0, 1]$, where the goodness of predictions is assessed according to the metrics stated in Section IV-A.

C. Cloud Detection and Mask Computation

In order to analyze the statistics of cloud coverage in SEN12MS-CR-TS and to model the spatio-temporal extent of clouds, we compute binary cloud masks m. For each optical image, the masks m are computed on-the-fly via the cloud detector of s2cloudless [19], which provides a binary mask of pixel-wise values in $\{0, 1\}$ that indicate cloud-free and cloud-covered pixels, respectively. The cloud mask accuracy of s2cloudless is reported to be on par with the multitemporal classifier MACCS-ATCOR joint algorithm (MAJA) [25], but the considered detector can be applied on mono-temporal satellite observations. Note that, alternatively to s2cloudless, the masks m may be computed via a dedicated neural network for cloud detection [26], [27]. However, s2cloudless has proved to be lightweight and provides sufficient performance at little extra computational cost in run time or memory, making it an appealing cloud detector to be applied on a large-scale data set such as SEN12MS-CR-TS. Example cloud detections are illustrated in Fig. 3.

III. Methods

We consider two distinctively different methods to highlight the benefits of our curated data set and the diverse tasks it allows to approach. The first method is a neural network reconstructing cloud covered pixels in time series of multimodal data to predict a single target image acquired at a cloud-free time point. The second approach introduces a neural network that performs sequence-to-sequence cloud removal, that is, it predicts a time series of cloud-free observations the same length as the cloudy input sequence.

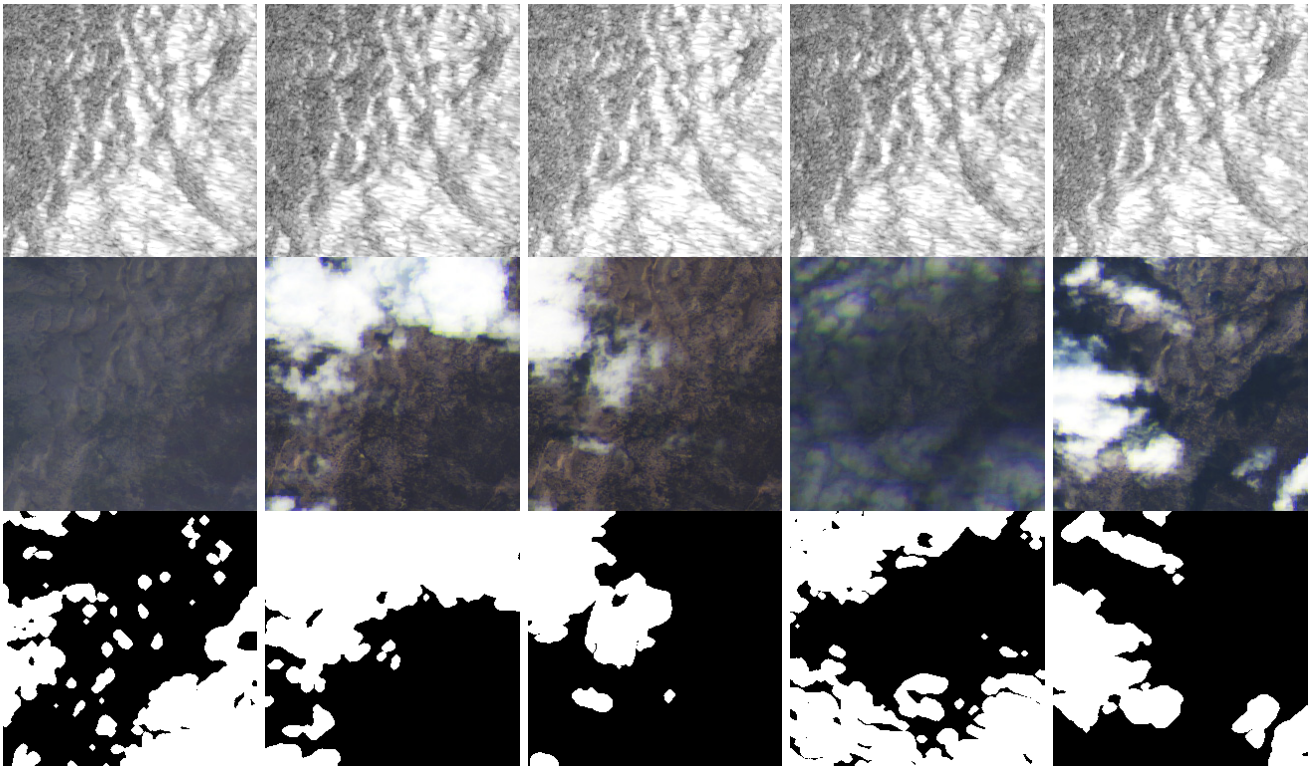


Fig. 3. Example data, preprocessed as stated in Section II-B. Rows: S1 data (in grayscale), S2 data (in RGB), and binary cloud masks (as per s2cloudless [19]). Columns: Samples of five different time points. The illustrations show that the observed region is affected by variable atmospheric disturbances and covered by a dynamic extent of clouds, changing over time. The detected cloud coverage at the individual time points is 36%, 49%, 23%, and 48%, with an average of about 39% across all illustrated samples. While some pixels are clear at least at one point in the series and may thus be reconstructed by integrating across time, whereas others are cloud-covered throughout the sequence and require spatial context or cloud-robust sensor information to be reconstructed.

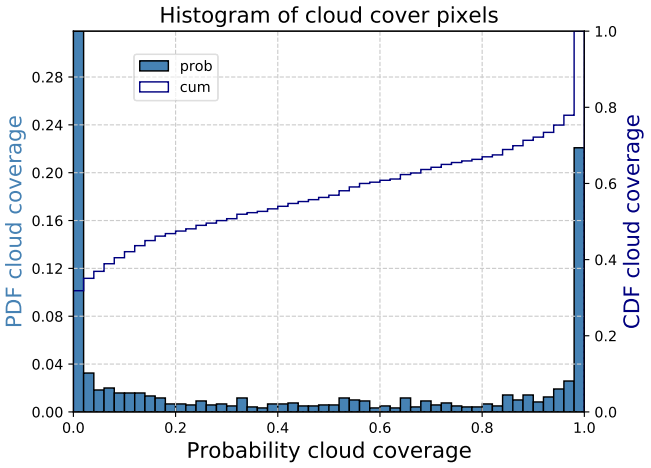


Fig. 4. Statistics of cloud coverage of SEN12MS-CR-TS train split, computed on full-scene images via the detector of [19]. On average, approximately 44% (±42%) of occlusion is observed. The empirical distribution of cloud coverage is bimodal and ranges from cloud-free views to total occlusion.

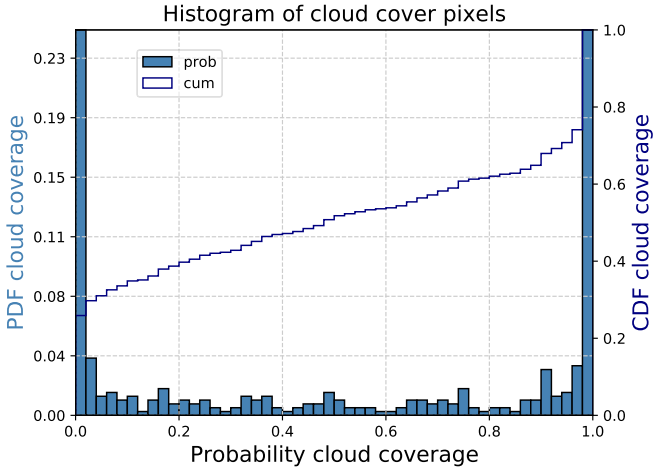


Fig. 5. Statistics of cloud coverage of SEN12MS-CR-TS test split, computed on full-scene images via the detector of [19]. On average, approximately 50% (±42%) of occlusion is observed. The empirical distribution of cloud coverage is bimodal and ranges from cloud-free views to total occlusion.

A. Multitemporal Multimodal Cloud Removal

For multitemporal multimodal cloud removal, we consider a deep neural network that builds on the generator of [12]. Our model receives a sequence of $t = 1, \ldots, n$ input tuples $(\mathrm{S1}, \mathrm{S2})_t$ and predicts a cloud-removed multispectral image $\hat{\mathrm{S2}}$. The architecture of the proposed model uses a ResNet [24] backbone, with Siamese residual branches processing the individual time points until their information gets integrated. That is, we replaced the pairwise concatenation of 2-D feature maps in [12] by stacking features in the

Fig. 6. Four different regions contained in SEN12MS-CR-TS, highlighting the diversity of sampled landcovers. The depicted S2 observations (RGB channels) are cloud-free samples of their respective time series. The average ROI covers about 40×40 km^2 and is split into over 700 patch samples, with each patch of size 256×256 px^2.

temporal domain, followed by 3-D convolutions. Moreover, as the first part of the generator of [12] is effectively a single time-point cloud removal subnetwork (as each time point is processed individually up to this point), we substitute this component by the established ResNet-based [24] cloud removal network of [13]. Subsequently, the feature maps are stacked in the temporal dimension and 3-D convolutions are applied to integrate information across time. The output of the network is a single cloud-free image prediction $\hat{S2}$. A schematic overview of the described architecture is shown in Fig. 7.

B. Internal Learning for Sequence-to-Sequence Cloud Removal

The sequence-to-sequence cloud removal method [28] follows the 3-D encoder–decoder architecture of [29], constituted of an encoder as well as a decoder component. Both components are arranged symmetrically in the style of U-Net [30] and linked via skip connections between paired layers. The input to the network is a sequence of multitemporal S1 samples and its output is a sequence of multitemporal cloud-removed S2 predictions. With regard to its input-to-output mapping, the proposed architecture resembles earlier SAR-to-optical translation method [31], [32]. Similar to these earlier domain translation approaches, our network learns information of the target domain (i.e., the optical imagery) via the supervision signal. Different from these approaches, the internal learning framework described below removes clouds and directly learns to denoise the target image sequence.

The architecture of the network is summarized in Fig. 8. Note that the key difference between the given model and the sequence-to-point method of Section III-A (depicted in Fig. 7) is in the output dimensions: Whereas the sequence-to-point architecture maps a sequence of n cloudy inputs to a single cloud-removed prediction, the sequence-to-sequence approach preserves the temporal information by mapping to a time series of n cloud removed outputs. Moreover, the point estimator receives tuples of S1 and S2 inputs, whereas the network of Fig. 8 is driven solely by S1 data (or Gaussian noise, as proposed in [33] and [29]). Finally, the sequence-to-point

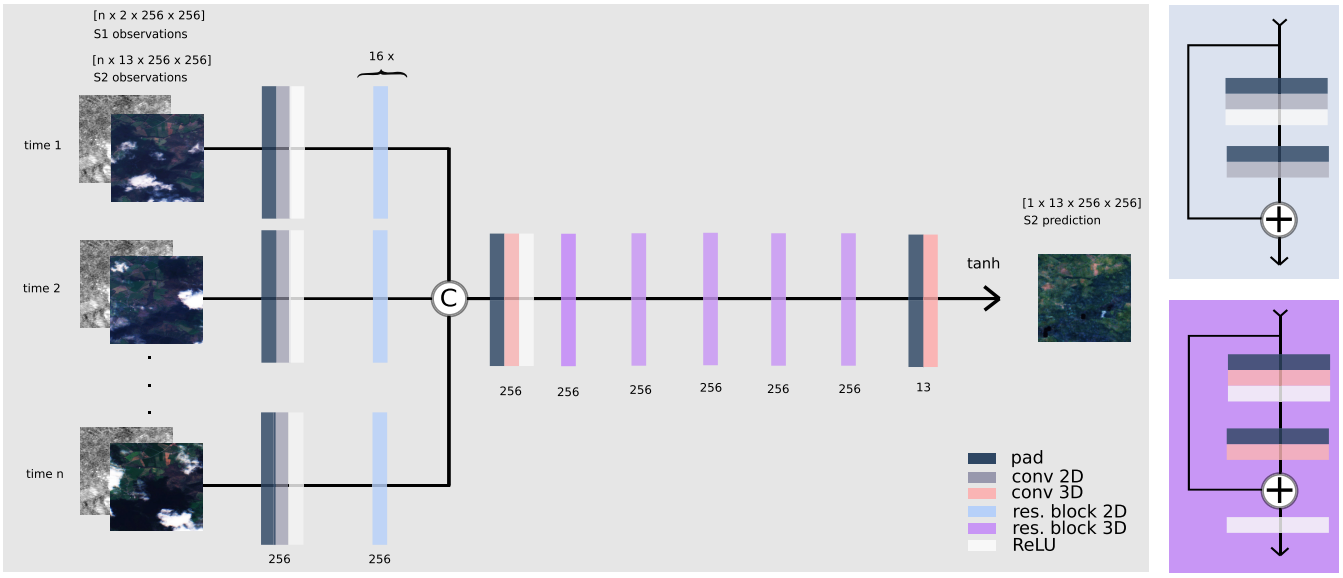


Fig. 7. Conceptual illustration of the sequence-to-point cloud removal architecture $G_{\text{seq2point}}$. The network is based on the architecture of [12] and consists of n Siamese ResNet branches [13] doing single time-point cloud removal on n individual time points. Subsequently, the feature maps are stacked in the temporal dimension and 3-D convolutions are applied to integrate information across time. The output of the network is a single cloud-free image prediction.

Algorithm 1 Internal Learning to Remove Clouds

procedure SEQ2SEQDECLOUDING($S1, S2, iterMax$)
  $G_{S1\to S2}$ = init. new NeuralNetwork()
  iterCount = 0
  while iterCount < iterMax **do**
    $\hat{S2} = G_{S1\to S2}(S1)$
    $G_{S1\to S2}.backpropagate(\mathcal{L}_{all}(S2, \hat{S2}))$
    iterCount = iterCount + 1
  Return $\hat{S2}$

network of Fig. 7 builds on the Siamese architecture of [12] with a ResNet backbone [13] plus 3-D convolutions, whereas the sequence-to-sequence approach of Fig. 8 follows a 3-D convolutional variant of U-Net [30], as proposed in [29].

The training procedure of the sequence-to-sequence network follows that of internal learning for image inpainting [29], [33], which is formalized in Algorithm 1. In this framework, for a given target sequence, a neural network is trained from scratch directly on the target sequence (without any need for additional or cloud-free training data) in order to reconstruct its noisy pixels. The observations exhibit spatio-temporal regularities and patterns (i.e., signal in the data), which is first modeled and learned by the network. The irregularities in the sequence (i.e., noise in the target data) are only internalized after, similar to a conventionally trained network overfitting to noise on training data. The internal learning approach exploits this signal–noise dichotomy and teaches a model to reconstruct cloud-covered pixels in the target sequence of S2 observations, without need for any external or cloud-free training data. In detail, a neural network is initialized and trained from scratch directly on the target sequence. At each iteration, the model receives input driving its activations (e.g., Gaussian noise or S1 recordings) and predicts a sequence $\hat{S2}$. The predictions $\hat{S2}$ are compared against the target sequence S2 (e.g., according to a cost function L_{all} as in 5) and the network learns to reproduce the cloud-free pixels. The training stops before the network overfits to internalizing the cloudy pixels.

With respect to its application and functionality, our sequence-to-sequence neural network resembles classical low-rank and sparse signal decomposition methods [34]–[37]: First, while neural networks are typically trained on a dedicated training data set separated from the test observations, numeral signal decomposition methods can be directly utilized on the data of interest. Similarly, our model can be directly applied on the test data. Second, unmixing of signals is very generic and can be applied to matrices as well as tensors. In comparison, the deep image prior approach applies to single images as well as time series [29], [33], too. Finally, the decomposition itself is into a low-rank part and a sparse component. The low-rank part denotes the data's compact representation and regularities. That is, spatial, spectral, or temporal (auto-)correlations such as the land cover mapped by a satellite. The sparse component consists of the irregular part of the data which has only a few nonzero entries, such as the appearance of clouds. In comparable terms, the internal learning technique allows our network to discover the regularities in the data and generalizing it to cloud-covered samples, before overfitting to the noise.

IV. Experiments and Results

This method details the experimental design and the corresponding results on the considered cloud removal methods as well as their ablation variants. Section IV-A specifies the measures of goodness used to assess the quality of the individual techniques' predictions. Section IV-B introduces the baselines compared against the proposed model of III-A on the sequence-to-point cloud removal task. Sections IV-C and IV-D

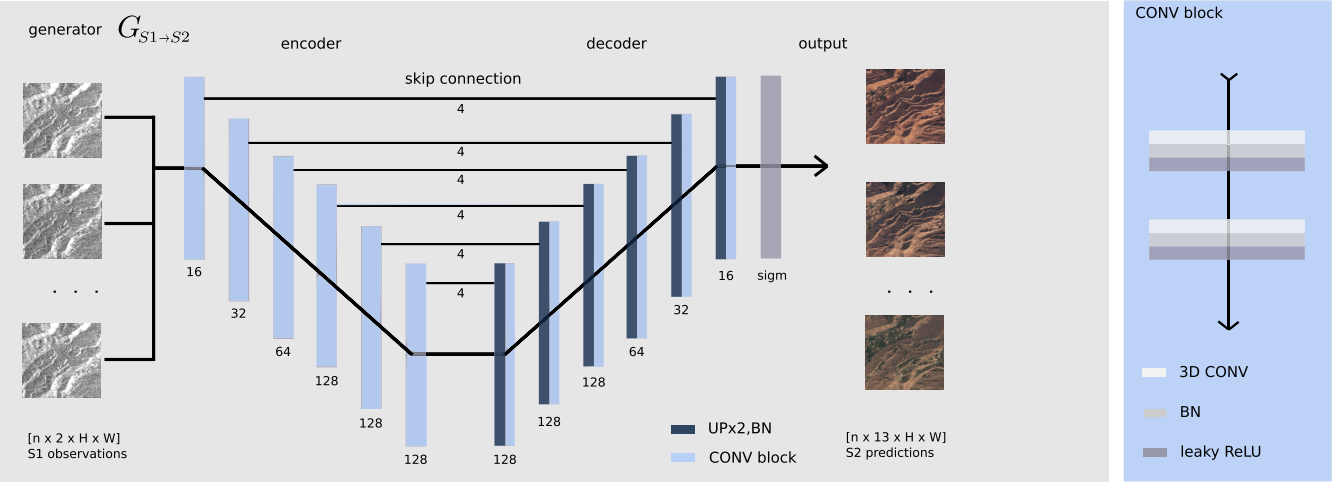


Fig. 8. Conceptual illustration of the 3-D encoder–decoder architecture G_{seq2seq} employed in the sequence-to-sequence cloud removal model [28]. The network is based on the architecture of [29] and consists of encoder and decoder parts arranged symmetrically in the style of U-Net [30], with skip connections between paired layers. Input to the network is a batch of multitemporal S1 observations. The output is a predicted batch of multitemporal multispectral S2 observations. For the ablation model considered in Section IV-D, Gaussian noise is used as an input as in [33] and [29].

detail the experiments and outcomes for the sequence-to-point and sequence-to-sequence cloud removal tasks, respectively.

A. Metrics

We evaluate the quantitative performance in terms of normalized root mean squares error (NRMSE), peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [38], and Spectral Angle Mapper (SAM) [39], defined as

$$\text{NRMSE}(x, y) = \sqrt{\frac{1}{C \cdot H \cdot W} \sum_{c=h=w=1}^{C,H,W} (x_{c,h,w} - y_{c,h,w})^2}$$

$$\text{PSNR}(x, y) = 20 \cdot \log_{10}\left(\frac{1}{\text{NRMSE}(x, y)}\right)$$

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + \epsilon_1)(2\sigma_{xy} + \epsilon_2)}{(\mu_x + \mu_y + \epsilon_1)(\sigma_x + \sigma_y + \epsilon_2)}$$

$$\text{SAM}(x, y) = \cos^{-1} \times \left(\frac{\sum_{c=h=w=1}^{C,H,W} x_{c,h,w} \cdot y_{c,h,w}}{\sqrt{\sum_{c=h=w=1}^{C,H,W} x_{c,h,w}^2 \cdot \sum_{c=h=w=1}^{C,H,W} y_{c,h,w}^2}}\right)$$

with images x, y compared via their respective pixel values $x_{c,h,w}, y_{c,h,w} \in [0, 1]$, dimensions $C = 3$, $H = W = 256$, means μ_x, μ_y, standard deviations σ_x, σ_y, covariance σ_{xy} as well as constants ϵ_1, ϵ_2 to stabilize the computation. NRMSE belongs to the class of pixel-level metrics and quantifies the average discrepancy between the target and the predicted pixels in Units of the measure of interest. PSNR is evaluated on the whole image and quantifies the signal-to-noise ratio of the prediction as a reconstruction of the target image. SSIM is another image-wise measure that builds on PSNR and captures the SSIM of the prediction to the target in terms of perceived change, contrast, and luminance [38]. The SAM measure is a third image-level metric that provides the spectral angle between the bands of two multichannel images [39]. For further analysis, the pixelwise NRMSE is evaluated in three manners: 1) over all pixels of the target image (as per convention), 2) only over cloud-covered pixels (visible in neither of any input optical sample) to measure reconstruction of noisy information, and 3) only over cloud-free pixels (visible in at least one input optical patch) quantifying preservation of information. The pixel-wise masking is performed according to the cloud mask given by the detector of [19].

B. Baseline Methods

To put the performances of our proposed model and ablations into context, we consider the following baseline methods. First ("least cloudy"), taking the least-cloudy input observation and forwarding it without further modification to be compared against the cloud-free target image. This provides a measure of how hard the cloud removal task is with respect to the extent of cloud-coverage present in the data. Second ("mosaicing"), we perform a mosaicing method that averages the values of pixels across cloud-free time points, thereby integrating information across time. That is, for any pixel, if there is a single clear-view time point, then its value is copied; for multiple cloud-free samples, the mean is formed and in case no cloud-free time point exists, then a value of 0.5 is taken as a proxy. This is to avoid any extreme values, such as cloudy pixels of high intensity. The mosaicing technique provides a measure of how much information can be reconstructed across time, from multispectral optical observations exclusively. Third, ResNet refers to a residual neural network as described and trained in Sections III-A and IV-C. The architecture is based on the model of [13] and serves as a relevant baseline because parts of this model are used as Siamese residual branches within our model, as detailed in Section III-A. It provides an estimate of how well a point-to-point cloud removal model can perform as a baseline. Fourth, the baseline spatio-temporal generative adversarial network (STGAN) denotes the "Branched ResNet generator [infra-red (IR)]" architecture of [12]. It is a sequence-to-point cloud removal model, and the architecture of our own

sequence-to-point neural network closely follows its design, as detailed in Section III-A. In sum, the purpose of assessing these baselines is to analyze whether trivial solutions to the multimodal multitemporal sequence-to-point cloud removal problem exist, and how any more sophisticated deep learning approach compares against these methods and our proposed model trained on SEN12MS-CR-TS.

C. Sequence-to-Point Cloud Removal

This section details the training specifics of the sequence-to-point cloud removal architecture introduced in Section III-A. As detailed in Section III-A, up to the temporal concatenation layer, we use a version of the ResNet-based [24] cloud removal network of [13] and pretrained it on SEN12MS-CR [14] according to the training specifics of [13]. All our considered sequence-to-point cloud removal networks and ablation models share this pretrained single-temporal cloud removal network as a starting point for the sake of comparability and in order to reduce the duration of training. The networks are trained for a total of ten epochs on one tuple of patches per location for every ROI in the training split. For training, the input S2 patches are filtered to display within 0%–50% of cloud coverage. The target S2 patch is selected to be the sample showing the minimum cloud coverage over the given time series, that is, it is not necessarily temporally preceding or following the input patches. For the first 25 000 steps in the training procedure, the networks are trained with the initial ResNet Siamese components frozen, exclusively optimizing the subsequent 3-D convolution layers. After the steps with the pretrained weights frozen and once the deeper layers have been calibrated to the initial network's latent feature maps, the full network is trained end-to-end for the remainder of the process. During training, the network minimizes the loss $\mathcal{L}_{\text{all}}$

$$\mathcal{L}_{\text{all}} = \lambda_{\text{L1}}\mathcal{L}_{\text{L1}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} \tag{1}$$

$$\mathcal{L}_{\text{L1}} = ||\text{S2} - \hat{\text{S2}}||_1 \tag{2}$$

$$\mathcal{L}_{\text{perc}} = ||\text{VGG16(S2)}, \text{VGG16}(\hat{\text{S2}})||_2 \tag{3}$$

with $\lambda_{\text{L1}} = 100$ according to [12] and $\lambda_{\text{perc}} = 1$ as hyperparameters weighting the individual pixel-wise loss $\mathcal{L}_{\text{L1}}$ and the perceptual loss $\mathcal{L}_{\text{perc}}$. The perceptual loss is computed by means of an auxiliary Visual Geometry Group 16 (VGG16) network [40] resulting in sharper image reconstructions [41]. In comparison to other VGG16 pretrained on classical computer vision data sets such as ImageNet [42] and thus limited to RGB channel data, we pretrained a VGG16 for landcover classification on the SEN12MS data set [43] according to the training protocol of [2]. The proposed sequence-to-point cloud removal network and its ablation variants are optimized via Adaptive Moment Estimation (ADAM) [44], with a learning rate of 0.0002 and momentum parameters $[0.5, 0.999]$ as in [12]. A batch size of one tuple of samples per iteration is used for training.

To evaluate performances on the test split, samples containing S2 observations from the complete range of cloud coverage (between 0 and 100%) are considered for input. Table I compared the results of our proposed model with the baselines detailed in Section IV-B. The results show that the proposed network outperforms the baselines in the majority of metrics, except for PSNR (where mosaicing comes first) and the NRMSE (clear) preservation metric (where the "least cloudy" approach performs best). This demonstrates that a deep neural network approach can typically outperform trivial solutions to the multimodal multitemporal cloud removal problem. Exemplary outcomes for the considered baselines on four different samples from the test split are presented in Fig. 9. The considered cases are cloud-free, partly cloudy, cloud-covered with no visibility except for a single time point, and cloud-coverage with no visibility at any time point. The results show that the considered models typically outperform the simple heuristics. One exceptional case is least cloudy in the absence of clouds, which manages to accomplish a faithful prediction in such settings. Moreover, the illustrations underline that multitemporal and multimodal data may benefit image reconstruction: While most methods perform well in the cloud-free or partly cloudy cases, multisource integration is needed if individual time points contain dense cloud coverage over wide areas. When all input data is covered by thick clouds, then this poses a severe challenge for all approaches considered. To analyze the benefits of including S1 SAR data, we perform an ablation study and compare a multisensor model against one only utilizing multispectral S2 input. Table II compared the results of the multimodal model with an ablation version not using S1 SAR data. The comparison illustrates the benefits of including SAR data when reconstructing cloud-covered pixels. Next, we conduct an ablation experiment to assess the additional benefits of utilizing the introduced perceptual loss. Table III compared the results of our proposed model with an ablation version not using the perceptual loss (i.e., setting $\lambda_{\text{perc}} = 0$ in eq 1). The outcomes imply that the usage of a perceptual loss results in cloud-removed predictions of a higher quality. Finally, we consider the extension of the proposed model into networks integrating four and five time points of input information. Table IV compared the performance of our model as a function of input time points ($n = 3, 4, 5$). The results indicate that considering longer time series may provide further improvements in terms of reconstructing cloud-covered information. In a final experiment on sequence-to-point cloud removal, Table V reports the performance of our proposed model ($n = 3$, with S1 and perceptual loss) as a function of cloud coverage. That is, for a given interval of cloud

TABLE I

QUANTITATIVE EVALUATION OF THE PROPOSED SEQUENCE-TO-POINT MODEL WITH BASELINE APPROACHES IN TERMS OF NORMALIZED ROOT MEAN SQUARED ERROR (NRSME), PSNR, SSIM [38], AND THE SAM [39] METRIC. OUR MODEL PERFORMS BEST IN THE MAJORITY OF METRICS, DEMONSTRATING THAT A DEEP NEURAL NETWORK APPROACH YIELDS ADDITIONAL BENEFITS OVER TRIVIAL SOLUTIONS TO THE MULTIMODAL MULTITEMPORAL CLOUD REMOVAL PROBLEM

model	NRMSE (all)	NRMSE (cloudy)	NRMSE (clear)	PSNR	SSIM	SAM
least cloudy	0.079	0.082	**0.031**	—	0.815	0.213
mosaicing	0.062	0.064	0.036	**31.68**	0.811	0.250
ResNet	0.060	0.062	0.040	26.04	0.810	0.212
STGAN	0.057	0.059	0.050	25.42	0.818	0.219
ours (n=3)	**0.051**	**0.052**	0.040	26.68	**0.836**	**0.186**

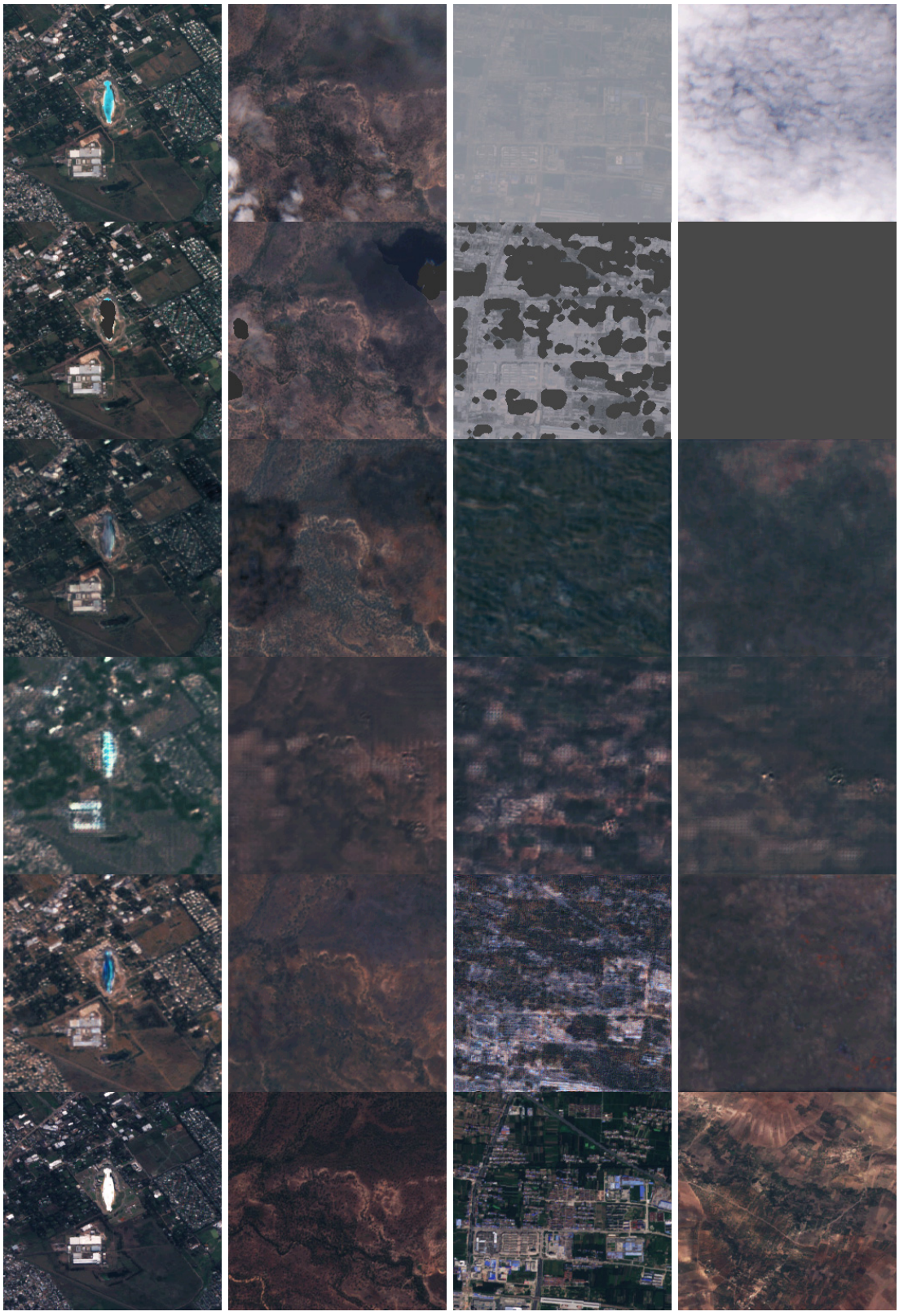


Fig. 9. Exemplary predictions and cloud-free target images for all baselines reported in Table I. Columns: Four different samples from the test split. The considered cases are cloud-free, partly cloudy, cloud-covered with no visibility except for a single time point, and cloud-covered with no visibility in any time point. Rows: Predictions of least cloudy, mosaicing, ResNet, STGAN, ours (n=3), as well as the cloud-free reference image. The results show that the considered models outperform the simple heuristics. Moreover, the illustrations underline that multitemporal and multimodal data may benefit image reconstruction.

TABLE II

COMPARISON OF THE PROPOSED SEQUENCE-TO-POINT MODEL INCLUDING SAR OBSERVATIONS VERSUS AN ABLATION VERSION WITHOUT SAR OBSERVATIONS IN TERMS OF NRSME, PSNR, SSIM [38], AND THE SAM [39] METRIC. THE COMPARISON ILLUSTRATES THE BENEFITS OF INCLUDING SAR DATA WHEN RECONSTRUCTING CLOUD-COVERED PIXELS

model	NRMSE (all)	NRMSE (cloudy)	NRMSE (clear)	PSNR	SSIM	SAM
ours (no S1)	0.054	0.057	0.054	25.35	0.832	0.194
ours (with S1)	**0.051**	**0.052**	**0.040**	**26.68**	**0.836**	**0.186**

TABLE III

COMPARISON OF THE PROPOSED SEQUENCE-TO-POINT MODEL INCLUDING PERCEPTUAL LOSS VERSUS AN ABLATION VERSION WITHOUT PERCEPTUAL LOSS IN TERMS OF NRSME, PSNR, SSIM [38], AND THE SAM [39] METRIC. THE OUTCOMES IMPLY THAT THE USAGE OF A PERCEPTUAL LOSS DURING TRAINING RESULTS IN CLOUD-REMOVED PREDICTIONS OF A HIGHER QUALITY AT TEST TIME

model	NRMSE (all)	NRMSE (cloudy)	NRMSE (clear)	PSNR	SSIM	SAM
ours (no percept.)	0.052	0.053	**0.039**	26.66	0.835	**0.180**
ours (with percept.)	**0.051**	**0.052**	0.040	**26.68**	**0.836**	0.186

TABLE IV

QUANTITATIVE EVALUATION OF THE PROPOSED SEQUENCE-TO-SEQUENCE MODEL WITH VARYING NUMBERS OF TIME POINTS ($n = 3, 4, 5$) IN TERMS OF NRSME, PSNR, SSIM [38], AND THE SAM [39] METRIC. OUR MULTITEMPORAL NETWORK WITH SAR GUIDANCE OUTPERFORMS THE MULTITEMPORAL ABLATION MODEL WITHOUT PRIOR SAR INFORMATION

model	NRMSE (all)	NRMSE (cloudy)	NRMSE (clear)	PSNR	SSIM	SAM
ours (n=3)	0.051	0.052	0.040	26.68	0.836	0.186
ours (n=4)	0.049	0.050	0.041	**27.10**	0.845	**0.172**
ours (n=5)	**0.048**	**0.048**	**0.032**	27.07	**0.846**	0.178

TABLE V

PERFORMANCE OF OUR SEQUENCE-TO-POINT CLOUD REMOVAL METHOD ($n = 3$, WITH S1 & WITH PERCEPTUAL LOSS) AS A FUNCTION OF CLOUD COVERAGE. FOR A GIVEN INTERVAL, ALL $n = 3$ INPUT IMAGES ARE SAMPLED TO CONTAIN A CORRESPONDING EXTENT OF CLOUDS. THE OUTCOMES SHOW THAT IMAGE RECONSTRUCTION PERFORMANCE IS HIGHLY DEPENDENT ON THE PERCENTAGE OF CLOUD COVERAGE. WHILE PERFORMANCE DECREASE IS NOT STRICTLY MONOTONOUS WITH AN INCREASE IN CLOUD COVERAGE, A STRONG ASSOCIATION PERSISTS

% cloud coverage	NRMSE (all)	NRMSE (cloudy)	NRMSE (clear)	PSNR	SSIM	SAM
0-10 %	**0.041**	**0.046**	**0.041**	**28.59**	**0.870**	**0.143**
10-20 %	0.044	**0.046**	0.043	27.69	0.848	0.166
20-30 %	0.046	0.047	0.044	27.25	0.841	0.169
30-40 %	0.048	0.050	0.045	26.77	0.830	0.169
40-50 %	0.047	0.048	0.045	26.86	0.830	0.167
50-60 %	0.049	0.494	0.048	26.55	0.825	0.185
60-70 %	0.052	0.052	0.043	26.10	0.817	0.184
70-80 %	0.049	0.050	0.044	26.59	0.816	0.179
80-90 %	0.050	0.050	0.044	26.54	0.820	0.175
90-100 %	0.063	0.063	—	24.79	0.786	0.222

coverage, all $n = 3$ input images are sampled to contain a corresponding extent of clouds. The outcomes show that image reconstruction performance is highly dependent on the percentage of cloud coverage. While performance decrease is not strictly monotonous with an increase in cloud coverage, a strong association persists.

TABLE VI

QUANTITATIVE EVALUATION OF BASELINE METHODS AND THE PROPOSED SEQUENCE-TO-SEQUENCE MODEL IN TERMS OF ROOT MEAN SQUARED ERROR (RSME), PSNR, SSIM, AND THE SAM [39] METRIC. OUR MULTITEMPORAL NETWORK WITH SAR GUIDANCE OUTPERFORMS THE CONSIDERED BASELINES AS WELL AS THE MULTITEMPORAL ABLATION MODEL WITHOUT PRIOR SAR INFORMATION

model	NRMSE (all)	PSNR	SSIM	SAM
RPCP [45]	0.403	7.911	0.264	30.567
NMFISL [46]	0.312	10.262	0.450	29.285
PNMF [47]	0.317	10.135	0.432	29.801
MNMF [48]	0.361	8.945	0.361	28.685
OSTD [49]	0.303	10.853	0.402	35.454
seq2seq (no S1)	0.298	11.434	0.494	28.127
seq2seq (with S1)	**0.274**	**11.590**	**0.512**	**27.733**

D. Sequence-to-Sequence Cloud Removal

A key characteristic of training the sequence-to-sequence cloud removal model described in Section III-B is the model being trained directly on the time series of images one aims to removes clouds from, without the use of any external training data as in [33] and [29]. More specifically, the training procedure teaches the network to replicate cloud-free pixels and inpaint cloud-covered ones in the target sequence S2 according to the cost function $\mathcal{L}_{\text{all}}$ formulated in [29] as

$$\mathcal{L}_{\text{all}} = \lambda_{L2}\mathcal{L}_{\text{L2}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} \tag{4}$$

$$\mathcal{L}_{\text{L2}} = ||\text{S2} \cdot (1-m), \quad \hat{\text{S2}} \cdot (1-m)||_2 \tag{5}$$

$$\mathcal{L}_{\text{perc}} = ||\text{VGG16(S2)} \cdot (1-m), \text{VGG16}(\hat{\text{S2}}) \cdot (1-m)||_2 \tag{6}$$

where $\lambda_{\text{L2}} = 1$ and $\lambda_{\text{perc}} = 0.01$ refer to hyperparameters that linearly combine the terms constituting $\mathcal{L}_{\text{all}}$. $\mathcal{L}_2$ is a pixel-wise reconstruction loss evaluated over the cloud-free pixels via an auxiliary VGG16 network [40] as explained before. The pseudo-code formalizing the intrinsic learning procedure is given in Algorithm 1 described in Section III-B and further justifications are stated in the original work of [33]. For a given target sequence, the network is trained for 20 passes with batches of $n = 5$ samples consisting of temporally adjacent images, for 100 iterations per pass. The network is optimized via ADAM [44] with a learning rate of 0.01 and the hyperparameters of Algorithm 1 set as stated in [29].

To quantitatively evaluate the considered model on SEN12MS-CR-TS, we propose the following protocol for a sequence-to-sequence cloud removal task: For a given target sequence, the least cloud-covered S2 observation is identified and denoted as a target image S2_t. The most cloudy S2 sample is observed and denoted as a source image S2_s. The cloud-covered pixels of S2_s according to a cloud mask m are alpha-blended with the cloud-free pixels of $S2_t$ similar to the approach of [3]. Finally, the cloud-removed prediction $\hat{\text{S2}}_t$ is then compared against the originally cloud-free S2_t in order to get a measure of goodness of cloud removal.

Table VI shows the results of the proposed network on the sequence-to-sequence cloud removal task following the

Fig. 10. Illustration of baseline methods for the sequence-to-sequence cloud removal task. The presented results show a cloudy image to be declouded, as well as the predictions via Riemannian Robust Principal Component Pursuit (RPCP) [45], Nonnegative Matrix Factorization Incremental Subspace Learning (NMFISL) [46], Probabilistic Nonnegative Matrix Factorization (PNMF) [47], Manhattan Nonnegative Matrix Factorization (MNMF) [48], and Online Stochastic Tensor Decomposition (OSTD) [49]. The results indicate that the presence of large and dense clouds poses a severe challenge for the considered methods. Most baselines decloud the image except for some residual artifacts, and some techniques display discolorization. For comparison with ours (no S1), ours (with S1), and the cloud-free target image, see Fig. 11.

aforementioned protocol. Furthermore, the considered model is compared against an ablation model, conditioned on random Gaussian noise as in [33] and [29] in place of the meaningful S1 input observations. Example outcomes of sequence-to-sequence cloud removal on a given ROI are depicted in Figs. 1 and 10. Furthermore, Fig. 11 provides a qualitative comparison between the predictions conditioned on SAR versus no prior information, underlining the benefits of multimodal information. The results highlight that the internal learning approach can learn to reconstruct cloud-covered pixels on a very limited amount of data. Furthermore, the results demonstrate that including SAR data results in performance benefits over the single-sensor baseline.

V. Discussion

The main contribution of this work is in curating and providing SEN12MS-CR-TS, a multimodal multitemporal data set for cloud removal in optical satellite imagery. Our large-scale data set covers a heterogeneous set of ROIs sampled from all over earth, acquired in different seasons throughout the year. Given that the contained observations cover clear-view, filmy, as well as nontransparent dense clouds, the objective of reconstructing cloud-covered information poses a challenging task for the considered methods and future approaches. For the sake of demonstrating the usefulness of the presented data set, we propose a sequence-to-point as well sequence-to-sequence cloud removal network. The considered methods are evaluated in terms of pixel-wise and image-wise metrics. We provide evidence that taking time-series information into account is facilitating the reconstruction of cloudy pixels and that including multisensor measurements does further improve the goodness of the cloud-removed predictions, justifying the design of SEN12MS-CR-TS to include multitemporal and multimodal data. The major difference to the preceding mono-temporal SEN12MS-CR data set [15] for cloud removal is that SEN12MS-CR-TS features a time series of 30 samples per ROI. This allows for developing methods that integrate information across time to more faithfully reconstruct cloud-obscured measurements. The sensitivity to temporal information may be particularly valuable for future research investigating the benefits of cloud removal to time-sensitive applications, such as change detection. On the other side, there is a tradeoff in terms of size, and while SEN12MS-CR-TS is more than twice as large as its mono-temporal precursor, the latter contains about two times as many ROIs sampled over all continents. However, both data sets are fully compatible, meaning that holdout ROIs of one belong to the test split of the other data set and vice versa. As there is no geo-spatial overlap across splits between both data sets, they can be combined for training or validation purposes. Finally, the two data sets exhibit a comparable extent of cloud coverage—about 50% and 48%, respectively, both covering the full spectrum from semitransparent haze to thick and dense clouds. A discrepancy between both data sets is in SEN12MS-CR having between 25% and 50% overlap between neighboring patches (following the design of [43]), whereas SEN12MS-CR-TS has no intersection between adjacent samples. SEN12MS-CR contains 122 218 patch triplets of S1, cloudy S2, and cloud-free S2 data, whereas SEN12MS-CR-TS consists of 30 time samples for each of the 15 578 patch-wise observations, for every S1 and S2 measurement. Due to the differences in preprocessing the two data sets are not coregistered patch-wise but, importantly, they share a common definition of ROIs as well as train and test splits. This way, they are compatible with one another such that SEN12MS-CR-TS can be utilized for time-series cloud removal, while SEN12MS-CR can provide further geospatial coverage of additional ROIs on individual time points. Thanks to the different designs of both data sets, they may prove beneficial facilitating a variety of downstream tasks, such as semantic segmentation [43], scene classification [2], or change detection [5], even in the presence of clouds.

Beyond the design of our novel data set, additional contributions of this work are in introducing the internal learning

Fig. 11. Illustrations on the effect of prior guidance via SAR information. Columns: SAR input to the SAR-conditioned model, cloud-free prediction of the model conditioned on Gaussian noise, cloud-free prediction of the model conditioned on SAR information, and cloud-free observation as a reference image. The structural information provided by the SAR input provides a strong prior to the model, guiding it toward learning to remove clouds in the cloudy input time series.

approach to cloud removal in optical satellite data, as well as demonstrating that SAR-to-optical cloud removal performs better than the original noise-to-optical translation framework. While our data set aims to provide a global distribution of samples, we think that the internal learning approach to cloud removal may be of particular interest for remote-sensing practitioners focusing on a single a spatially confined ROI, as no further external data is necessary.

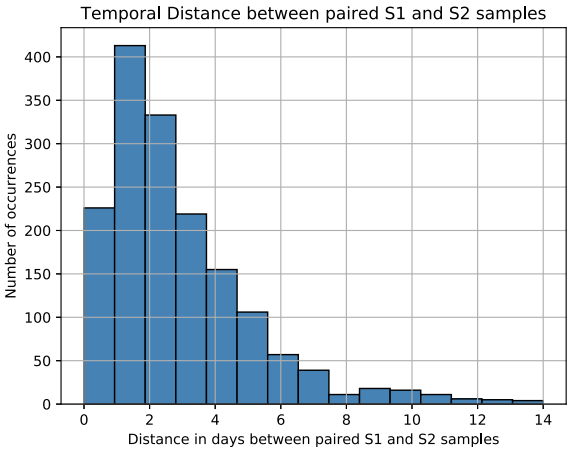


Fig. 12. Histogram of temporal differences between paired observations. The mean time differences across all paired observations are 2.61 ($\pm$ 2.41), indicating a close proximity between paired samples.

VI. Conclusion

As a large extent of our planet is covered by haze or clouds at any given point in time, such atmospheric distortions pose a severe constraint to the ongoing monitoring of earth. To approach this challenge, our work presented SEN12MS-CR-TS, a multimodal and multitemporal data set for training and evaluating global and all-season cloud removal methods. Our data set contains Sentinel-1 and Sentinel-2 observations from over 80000 km^2 of landcover, distributed globally and recorded through the year. The globally distributed ROIs are large-sized and capture a heterogeneous mass of landcover. We demonstrated the practicality of SEN12MS-CR by considering two methods: First, a model for sequence-to-point cloud removal. Second, a network for sequence-to-sequence cloud removal which, to our knowledge, provides the first case a model preserving temporal information is proposed in the context of cloud removal. Both methods benefited from the presence of coregistered and paired SAR measurements contained in our data set. The conducted experiments highlight the contribution of our curated data set to the remote-sensing community as well as the benefits of multimodal and multitemporal information to reconstruct noisy information. SEN12MS-CR is made public to facilitate future research in multimodal and multitemporal image reconstruction.

Appendix

Temporal Coincidence of Paired Observations

Full-scene observations of Sentinel-1 and Sentinel-2 are collected within a 14-day time window in a paired manner, as specified in Section II-A. To further analyze the temporal distance within paired data, Fig. 12 illustrates the empirically observed coincidences within SEN12MS-CR-TS. The mean time differences across all paired observations are 2.61 ($\pm$ 2.41), which is considerably smaller than the interval bound and implies a close proximity between paired samples.

ACKNOWLEDGMENT

The authors would like to thank ESA and the Copernicus program for making the Sentinel observations accessed for this submission publicly available. The authors would also like to thank Rewanth Ravindran for assisting us in the data curation process.

REFERENCES

[1] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 3826–3852, Jul. 2013.

[2] M. Schmitt and Y.-L. Wu, "Remote sensing image classification with the SEN12MS dataset," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. V-2-2021, pp. 101–106, Jun. 2021.

[3] M. U. Rafique, H. Blanton, and N. Jacobs, "Weakly supervised fusion of multiple overhead images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1479–1486.

[4] M. Schmitt, J. Prexl, P. Ebel, L. Liebel, and X. X. Zhu, "Weakly supervised semantic segmentation of satellite images for land cover mapping—Challenges and opportunities," 2020, *arXiv:2002.08254*.

[5] P. Ebel, S. Saha, and X. X. Zhu, "Fusing multi-modal data for supervised change detection," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. B3-2021, pp. 243–249, Jun. 2021.

[6] S. Saha, P. Ebel, and X. X. Zhu, "Self-supervised multisensor change detection," *IEEE Trans. Geosci. Remote Sens.*, early access, Sep. 15, 2022, doi: 10.1109/TGRS.2021.3109957.

[7] K. Enomoto *et al.*, "Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets," 2017, *arXiv:1710.04835*.

[8] C. Grohnfeldi, M. Schmitt, and X. Zhu, "A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from sentinel-2 images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 1726–1729.

[9] P. Singh and N. Komodakis, "Cloud-GAN: Cloud removal for Sentinel-2 imagery using a cyclic consistent generative adversarial networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 1772–1775. [Online]. Available: https://ieeexplore.ieee.org/document/8519033/

[10] J. D. Bermudez, P. N. Happ, R. Q. Feitosa, and D. A. B. Oliveira, "Synthesis of multispectral optical images from SAR/optical multitemporal data using conditional generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1220–1224, Aug. 2019.

[11] Z. Gu, Z. Zhan, Q. Yuan, and L. Yan, "Single remote sensing image dehazing using a prior-based dense attentive network," *Remote Sens.*, vol. 11, no. 24, p. 3008, Dec. 2019.

[12] V. Sarukkai, A. Jain, B. Uzkent, and S. Ermon, "Cloud removal in satellite images using spatiotemporal generative networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1796–1805.

[13] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 333–346, Aug. 2020.

[14] P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu, "Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5866–5878, Jul. 2020.

[15] P. Ebel, M. Schmitt, and X. X. Zhu, "Cloud removal in unpaired Sentinel-2 imagery using cycle-consistent GAN and SAR-optical data fusion," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 2065–2068.

[16] R. Bamler, "Principles of synthetic aperture radar," *Surv. Geophys.*, vol. 21, nos. 2–3, pp. 147–157, 2000.

[17] S. Oehmcke, T.-H.-K. Chen, A. V. Prishchepov, and F. Gieseke, "Creating cloud-free satellite imagery from image time series with deep learning," in *Proc. 9th ACM SIGSPATIAL Int. Workshop Anal. Big Geospatial Data*, Nov. 2020, pp. 1–10.

[18] Q. Zhang, Q. Yuan, Z. Li, F. Sun, and L. Zhang, "Combined deep prior with low-rank tensor SVD for thick cloud removal in multitemporal images," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 161–173, Jul. 2021.

[19] A. Zupanc. (2017). *Improving Cloud Detection With Machine Learning*. Accessed: Oct. 10, 2019. [Online]. Available: https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13

[20] W. Sintarasirikulchai, T. Kasetkasem, T. Isshiki, T. Chanwimaluang, and P. Rakwatin, "A multi-temporal convolutional autoencoder neural network for cloud removal in remote sensing images," in *Proc. 15th Int. Conf. Electr. Engineering/Electronics, Comput., Telecommun. Inf. Technol. (ECTI-CON)*, Jul. 2018, pp. 360–363.

[21] K. Perlin, "Improving noise," in *Proc. 29th Annu. Conf. Comput. Graph. Interact. Techn.*, 2002, pp. 681–682.

[22] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth engine: Planetary-scale geospatial analysis for everyone," *Remote Sens. Environ.*, vol. 202, pp. 18–27, Dec. 2017.

[23] L. Veci, P. Prats-Iraola, R. Scheiber, F. Collard, N. Fomferra, and M. Engdahl, "The Sentinel-1 toolbox," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Oct. 2014, pp. 1–3.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2016, pp. 770–778.

[25] V. Lonjou *et al.*, "MACCS-ATCOR joint algorithm (MAJA)," *Proc. SPIE*, vol. 10001, Oct. 2016, Art. no. 1000107.

[26] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote Sens. Environ.*, vol. 229, pp. 247–259, Aug. 2019.

[27] D. López-Puigdollers, G. Mateo-García, and L. Gómez-Chova, "Benchmarking deep learning models for cloud detection in Landsat-8 and Sentinel-2 images," *Remote Sens.*, vol. 13, no. 5, p. 992, Mar. 2021.

[28] P. Ebel, M. Schmitt, and X. X. Zhu, "Internal learning for Sequence-to-Sequence cloud removal via synthetic aperture radar prior information," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2021, pp. 2691–2694.

[29] H. Zhang, L. Mai, N. Xu, Z. Wang, J. Collomosse, and H. Jin, "An internal learning approach to video inpainting," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 2720–2729.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-319-24574-4_28

[31] M. Fuentes Reyes, S. Auer, N. Merkle, C. Henry, and M. Schmitt, "SAR-to-optical image translation based on conditional generative adversarial networks-optimization, opportunities and limits," *Remote Sens.*, vol. 11, no. 17, p. 2067, Sep. 2019.

[32] L. Wang, X. Xu, Y. Yu, R. Yang, R. Gui, Z. Xu, and F. Pu, "SAR-to-optical image translation using supervised cycle-consistent adversarial networks," *IEEE Access*, vol. 7, pp. 129136–129149, 2019.

[33] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9446–9454.

[34] F. De la Torre and M. J. Black, "Robust principal component analysis for computer vision," in *Proc. Int. Conf. Comput. Vis.*, vol. 1. Jul. 2001, pp. 362–369.

[35] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[36] X. X. Zhu and R. Bamler, "Tomographic SAR inversion by L_1-norm regularization-the compressive sensing approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3839–3846, Feb. 2010.

[37] X. X. Zhu and R. Bamler, "A sparse image fusion algorithm with application to pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2827–2836, May 2013.

[38] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[39] F. A. Kruse *et al.*, "The spectral image processing system (SIPS)-interactive visualization and analysis of imaging spectrometer data," *AIP Conf.*, vol. 283, no. 1, pp. 192–201, 1993.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[41] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 694–711. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46475-6_43

[42] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[43] M. Schmitt, L. Hughes, C. Qiu, and X. Zhu, "SEN12MS–A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 4, pp. 153–160, Oct. 2019.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
[45] M. Hintermüller and T. Wu, "Robust principal component pursuit via inexact alternating minimization on matrix manifolds," *J. Math. Imag. Vis.*, vol. 51, no. 3, pp. 361–377, 2015.
[46] S. S. Bucak, B. Gunsel, and O. Gursoy, "Incremental nonnegative matrix factorization for background modeling in surveillance video," in *Proc. IEEE 15th Signal Process. Commun. Appl.*, Oct. 2007, pp. 1–4.
[47] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2000, pp. 556–562. [Online]. Available: https://papers.nips.cc/paper/2000/hash/f9d1152547c0bde01830b7e8bd60024c-Abstract.html
[48] N. Guan, D. Tao, Z. Luo, and J. Shawe-Taylor, "MahNMF: Manhattan non-negative matrix factorization," 2012, *arXiv:1207.3438*.
[49] A. Sobral, S. Javed, S. K. Jung, T. Bouwmans, and E.-H. Zahzah, "Online stochastic tensor decomposition for background subtraction in multispectral video sequences," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 106–113.

Patrick Ebel received the B.Sc. degree in cognitive science from the University of Osnabrück, Osnabrück, Germany, in 2015, and the M.Sc. degree in cognitive neuroscience and the M.Sc. degree in artificial intelligence from Radboud University Nijmegen, Nijmegen, The Netherlands, in 2018. He is currently pursuing the Ph.D. degree with the Data Science in Earth Observation Laboratory, Department of Aerospace and Geodesy, Technical University of Munich (TUM), Munich, Germany.

His research interests include machine learning as well as its applications in computer vision and remote sensing. Specifically, he is working on multimodal and multitemporal data fusion and automated image reconstruction methods.

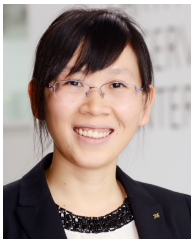
Yajin Xu received the B.Sc. degree (Hons.) in engineering from Wuhan University, Wuhan, China, in 2018. He is currently pursuing the joint M.Sc. degree in earth-oriented space science and technology (ESPACE) with Wuhan University and the Technical University of Munich (TUM), Munich, Germany.

His study focus is machine learning applied to remote-sensing data. In 2021, he was a Research Assistant with the Data Science in Earth Observation (SiPEO) Group, TUM and Remote Sensing Technology Institute of DLR, investigating deep learning-based approaches for cloud removal in optical satellite data. His interest includes geospatial data analysis.

Michael Schmitt (Senior Member, IEEE) received the Dipl.-Ing. (Univ.) degree in geodesy and geoinformation, the Dr.-Ing. degree in remote sensing, and the Habilitation degree in data fusion from the Technical University of Munich (TUM), Munich, Germany, in 2009, 2014, and 2018, respectively.

Since 2021, he has been the Chair of Earth Observation at the Department of Aerospace Engineering, Bundeswehr University Munich, Neubiberg, Germany. Before that, he was a Full Professor of applied geodesy and remote sensing with the Department of Geoinformatics, Munich University of Applied Sciences, Munich. From 2015 to 2020, he was a Senior Researcher and the Deputy Head at the Professorship for Data Science in Earth Observation at TUM. In 2019, he was additionally appointed as an Adjunct Teaching Professor with the Department of Aerospace and Geodesy, TUM. In 2016, he was a Guest Scientist at the University of Massachusetts, Amherst, MA, USA. His research focuses on image analysis and machine learning applied to the extraction of information from multimodal remote sensing observations. In particular, he is interested in remote-sensing data fusion with a focus on SAR and optical data.

Dr. Schmitt is a Co-Chair of the Working Group "SAR and Microwave Sensing" of the International Society for Photogrammetry and Remote Sensing and also the Working Group "Benchmarking" of the IEEE-Geoscience and Remote Sensing Society (GRSS) Image Analysis and Data Fusion Technical Committee. He frequently serves as a reviewer for a number of renowned international journals and conferences and has received several best reviewer awards. He is an Associate Editor of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.



Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.-Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is a Professor of data science in earth observation (former: signal processing in earth observation) at TUM and the Head of the Department "EO Data Science" at the Remote Sensing Technology Institute, German Aerospace Center (DLR). Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School and also the Head of the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the Director of the International Future AI Laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since October 2020, she has also been the Co-Director of the Munich Data Science Institute (MDSI), TUM. She was a Guest Scientist or a Visiting Professor at the Italian National Research Council (CNR-IREA), Naples, Italy; Fudan University, Shanghai, China; the University of Tokyo, Tokyo, Japan; and the University of California, Los Angeles, CA, USA; in 2009, 2014, 2015, and 2016, respectively. She is currently a Visiting AI Professor at ESA's Phi-Laboratory, Frascati, Italy. Her main research interests are remote sensing and earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She serves in the Scientific Advisory Board in several research organizations, among others the German Research Center for Geosciences (GFZ) and the Potsdam Institute for Climate Impact Research (PIK). She is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and serves as the Area Editor responsible for special issues of *IEEE Signal Processing Magazine*.

A.5 Explaining the Effects of Clouds on Remote Sensing Scene Classification

Reference: J. Gawlikowski*, P. Ebel*, M. Schmitt, and X. X. Zhu. *Explaining the effects of clouds on remote sensing scene classification.* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15:9976–9986, 2022.
* Authors contributed equally to this work.

Explaining the Effects of Clouds on Remote Sensing Scene Classification

Jakob Gawlikowski, Patrick Ebel, Michael Schmitt, *Senior Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*—Most of Earth is covered by haze or clouds, impeding the constant monitoring of our planet. Preceding works have documented the detrimental effects of cloud coverage on remote sensing applications and proposed ways to approach this issue. However, up to now, little effort has been spent on understanding how exactly atmospheric disturbances impede the application of modern machine learning methods to Earth observation data. Specifically, we consider the effects of haze and cloud coverage on a scene classification task. We provide a thorough investigation of how classifiers trained on cloud-free data fail once they encounter noisy imagery—a common scenario encountered when deploying pretrained models for remote sensing to real use cases. We show how and why remote sensing scene classification suffers from cloud coverage. Based on a multistage analysis, including explainability approaches applied to the predictions, we work out four different types of effects that clouds have on scene prediction. The contribution of our work is to deepen the understanding of the effects of clouds on common remote sensing applications and consequently guide the development of more robust methods.**

***Index Terms*—Classification, clouds, deep learning, explainability, remote sensing, robustness.**

Manuscript received 19 January 2022; revised 7 August 2022 and 24 October 2022; accepted 26 October 2022. Date of publication 21 November 2022; date of current version 30 November 2022. This work was supported by the Federal Ministry for Economic Affairs and Energy of Germany in the project "AI4Sentinels—Deep Learning for the Enrichment of Sentinel Satellite Imagery" under Grant FKZ50EE1910. The work of X. Zhu was jointly supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under Grant (ERC-2016-StG-714087, Acronym: *So2Sat*), in part by the Helmholtz Association through the Framework of Helmholtz AI under Grant ZT-I-PF-5-01—Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)," in part by the Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research" under Grant W2-W3-100, in part by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001, and in part by the German Federal Ministry for Economic Affairs and Climate Action in the framework of the "National Center of Excellence ML4Earth" under Grant 50EE2201C. *(Jakob Gawlikowski and Patrick Ebel contributed equally to this work.) (Corresponding author: Xiao Xiang Zhu.)*

Jakob Gawlikowski is with the Institute of Data Science, German Aerospace Centre (DLR), 07745 Jena, Germany, and also with the Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: jakob.gawlikowski@dlr.de).

Patrick Ebel is with the Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: patrick.ebel@tum.de).

Michael Schmitt was with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany. He is now with the Department of Aerospace Engineering, University of the Bundeswehr Munich, 85577 Neubiberg, Germany (e-mail: michael.schmitt@unibw.de).

Xiao Xiang Zhu is with the Chair of Data Science in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: xiaoxiang.zhu.ieee@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2022.3221788

I. Introduction

CLOUD coverage is detrimental to common remote sensing applications, such as remote sensing scene classification [1], [2], [3] and semantic segmentation [4], [5]. While clouds are characterized in great detail [6], [7] and different approaches for handling them have been investigated, less effort has been spent to investigate what exactly its effects on remote sensing applications are. The existing approaches range from learning cloud removal for preprocessing [8], [9], [10], [11], [12], [13] to familiarizing neural networks with clouds by including cloud-covered observations in the training dataset, such that the models learn to ignore clouds irrelevant to the task at hand [3], [4], [14]. Such approaches that include cloudy images in the training process are limited to samples with transparent clouds or samples where the crucial features for classification are not covered. Although recent work demonstrated that explicitly performing cloud removal may improve model robustness [15], the coverage of important features or the misinterpretation of features induced by clouds still poses a significant problem for remote sensing tasks sensitive to inter- and intraclass feature differences [16], [17]. Furthermore, the majority of curated optical satellite datasets are explicitly cleaned from clouds and remote sensing models are subsequently (pre-) trained on (predominantly) clear-view data [1], [2], [18]. This common practice, however, is in contrast to the application of networks typically trained on noncloudy datasets to data in the wild, which is to a large extent polluted by haze or clouds [6]. Fig. 1 illustrates the possible negative effects of cloud cover on scene classification. Fine-tuning such models on cloudy observations would require the post-hoc collection of new data plus task-related labels, which may thus be impracticable for the remote sensing practitioner. Hence, the issue of cloud-agnostic networks confronted with out-of-distribution data at test time commonly persists. That is, classifiers trained on cloud-free data may in practice still encounter samples significantly deviating from the distribution of data that the model has been trained on.

In order to understand the causes of the experienced drops in task performances [3], [14], we provide detailed insights into how clouds affect every single part of the remote sensing pipeline—from raw data to a model's predictions. To our knowledge, the only prior study explaining neural network's scene classifications focuses on clear data without taking the effects of clouds into account [19]. In our work, we explain the causes of overconfident miss-classifications resulting from scenes fully or partially covered by clouds. Specifically, we consider single-label scene classification on the SEN12MS

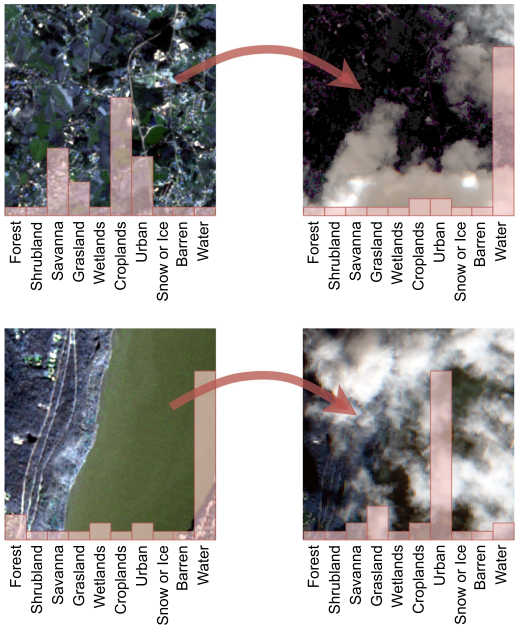


Fig. 1. Two examples of the effect of clouds on single-label scene classification. The visualization shows two examples of clear images, cloudy images, and the corresponding predicted class probabilities. While in both cases, the cloud-free image is classified correctly with respect to the ground truth, the cloudy version is misclassified. In the upper example, much of the croplands are obscured by cloud shadow, which causes the misclassification as a water body with a high soft-max probability. In the lower example, the clouds cover a large range of the water but keep a part of a city visible such that the sample containing clouds is misclassified as Urban with a high conviction. The cloud coverage of the samples is 19% and 77%, respectively. Although parts of the images are still visible, the classifier's predictions are misguided by the clouds and the resulting shadows.

dataset [2]. We use the Sentinel-2 images of the dataset, which have a resolution of 256×256 pixels and are assigned to one of 10 classes of land cover types. For cloud-covered samples, we utilize the corresponding and co-registered observations of the SEN12MS-CR dataset [20]. Our analysis is fourfold, as we consider the effects of clouds on the following.

1) *Data distribution*, by describing the effects of clouds on the statistics of the input dataset and how this affects individual land cover types.
2) *Classification performance*, by evaluating the impact of cloud coverage on a task performance level with respect to the considered single-label classification task, including individual class confusions.
3) *Effects on the network output*, by investigating the changes in the network predictions and the capability to separate cloudy samples from clear samples based on the network's output.
4) *Feature importance and network focus*, by analyzing which parts of an image drive a classifier's predictions and how this changes in the presence of clouds.

In sum, the contribution of this work is to provide a more thorough qualitative as well as quantitative analysis and interpretation of the effects of clouds on remote sensing applications, to subsequently allow further research to handle cloud-covered data more gracefully than currently feasible. The code base for the presented results and experiments can be found in our github repository: https://github.com/JakobCode/explaining_cloud_effects.

II. Data

A. Remote Sensing Data

To assess the effects of clouds on the scene classification task, both cloudy observations and patchwise land cover class annotations are required. For single-class labels and cloud-free observations, this work builds on the SEN12MS dataset of globally sampled Sentinel-1 and Sentinel-2 data [2], [21]. The Sentinel-2 data correspond to the Level-1 C top-of-atmosphere reflectance products. Semantic land cover annotations are given by the MODIS-derived [22] simplified IGBP scheme of [21], which consists of 10 different land cover types. For single-class labels, we use the provided target values in [2] which, for any sample, are given by the mode of its pixel-based simplified IGBP land cover type map. For every 252 globally distributed regions of interest, a large-scale observation is acquired within a given meteorologically defined season for each of the three sensors and collected semiautomatically via Google Earth Engine [23]. Each region on average covers an area of approximately 52×40 km^2 land surface, equating to images of about 5200×4000 pixels. All full-scene observations are translated into the Universal Transverse Mercator coordinate reference system. Afterward, the images are sliced into patches of sizes 256×256 pixels with a stride of 128 pixels, such that neighboring patches have an overlap of 25% to 50%. Patches that contain invalid pixels, either due to sensor noise or due to the coordinate transformation, are automatically removed from the dataset. For cloud-covered data, we utilize the compatible and co-registered SEN12MS-CR dataset of cloudy Sentinel-2 data [20].[1] The additional cloud-covered full-scene observations are acquired in the same year and season as their respective cloud-free counterparts to minimize surface changes and are preprocessed analogously. For training and testing data of this study, we use the intersection of both datasets' splits, respectively. That is, for each considered testing sample a cloud-free and a co-registered, potentially cloud-covered version exists.

In order to compute statistics on the extent of cloud coverage in the considered dataset, a pixelwise cloud map is required. We utilize s2cloudless [24] to compute binary cloud masks. The resulting distribution of cloud coverage on the considered test split is depicted in Fig. 2. The statistics indicate that the complete range of cloud coverage is present in the test split, from clear view to fully obscured. The distribution exhibits a concentration at high cloud coverage, implying an often impossible classification task. For hard or even impossible classification tasks, the predictions should be given with a larger entropy among the predicted soft-max probability vectors.

B. Data Distribution

The distribution of land cover types in the test split is reported in Fig. 3. The globally sampled land cover types are unbalanced,

[1] https://patrickTUM.github.io/cloud_removal

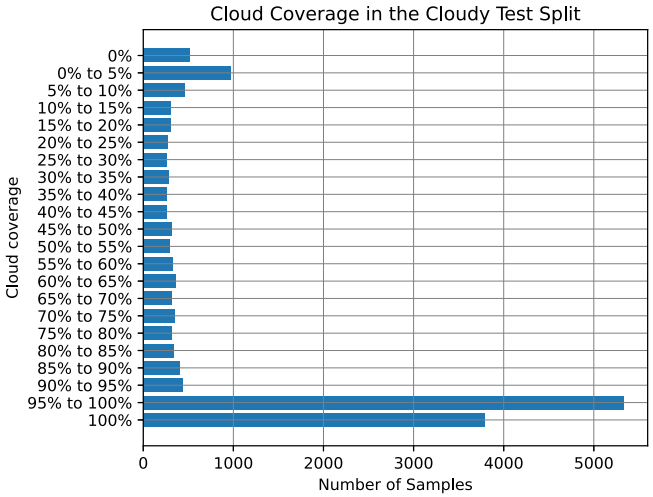


Fig. 2. Histogram of test split samples per percentage of cloud coverage. All extent of cloud coverage is present in the test split. The distribution exhibits a concentration at high cloud coverage, implying a challenging or even infeasible classification task.

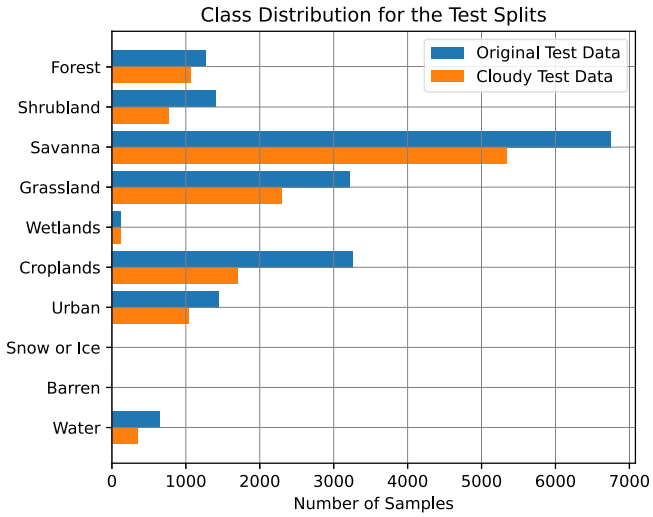


Fig. 3. Histogram of the land cover class distribution in the original test split of the SEN12MS dataset and the considered test split that is based on the intersection with the cloudy SEN12MSCR dataset. The globally sampled land cover types are unbalanced, with majority classes like *Savanna* while other classes hardly occur.

with majority classes like *Savanna* while other classes (*Snow, Barren*) hardly occur. The distribution of land cover in the training split is comparable, which makes it representative of the holdout data.

The bandwise statistics of each class's spectral properties are illustrated in Fig. 4. The illustrated band intensities are computed by calculating the grand mean across all samples, averaging spatial dimensions for each class and band separately. The statistics show that the presence of clouds results in an average increase in band intensities as well as a considerable increase in standard deviations. That is, clouds result in land cover types being less separable based solely on their spectral properties. Furthermore, the considerable shift in the data distribution makes the behavior

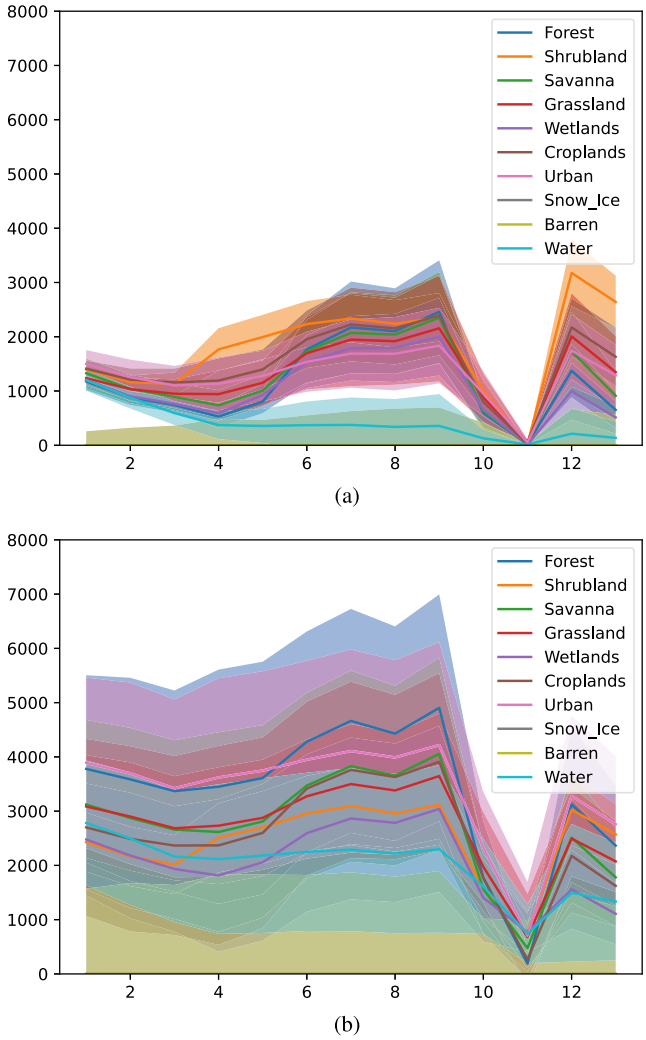


Fig. 4. Bandwise spectral fingerprint of each land cover class. The figures illustrate amplitude as a function of spectral bands and land cover type. Band intensities are computed as the grand mean across all samples, averaging across spatial dimensions for each class and band separately. The presence of clouds results in an average increase in band intensities and standard deviation. This indicates that, in the presence of clouds, land cover types become less separable on the basis of their spectral fingerprint. (a) Statistics of cloud-free data. (b) Statistics of 95% cloud-covered data.

of neural networks unreliable and sensitive to misinterpretations caused by very confident but false predictions [25], [26].

III. Scene Classification Under Cloudy and Noncloudy Conditions

A. Scene Classification Models

We investigate the scene classification performance of a ResNet50 as well as a ResNet101 [27], a DenseNet121 [28], a VGG-16, and a VGG-19 model [29], which were already previously considered for this task [2]. Other than [2], we make use of all Sentinel-2 bands to include atmospheric information, which is of particular relevance in the presence of clouds. We trained on the cloud-free SEN12MS training data and randomly held out 10% of the training data for a validation set. The models were trained for 30 epochs and the models with the

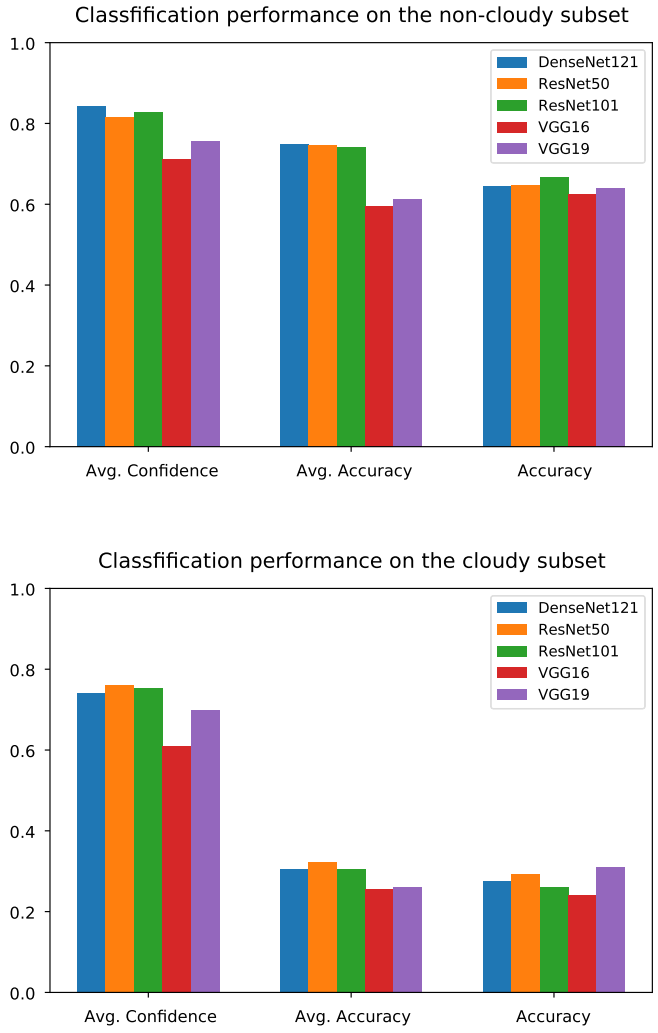


Fig. 5. Classification performances on cloudy and cloud-free data. The evaluated metrics demonstrate comparable performances for the different architectures considered in [2].

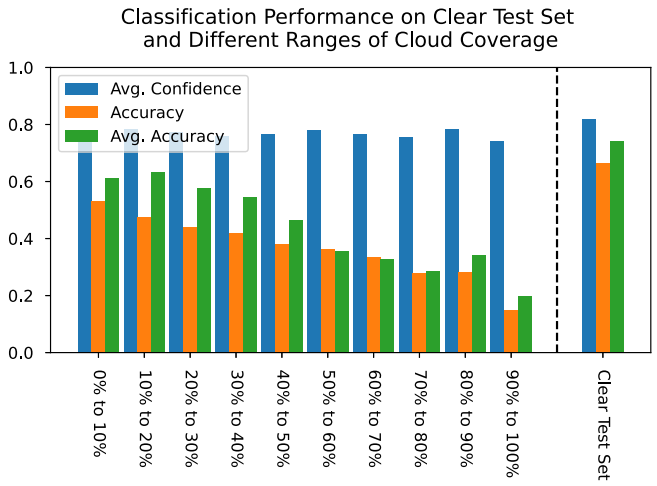


Fig. 6. Performance of the ResNet50 architecture as a function of varying ranges of cloud coverage. While accuracies are detriment with increasing cloud coverage, the network's confidence remains consistently high.

best performance on the validation set were saved during the training. For the optimization procedure, we utilized the ADAM optimizer [30] with a learning rate and weight decay of 10^{-5}. For the implementation, we extended the PyTorch [31] implementation provided by Schmitt and Wu.[2] The trained networks perform comparably to the baselines proposed in [2], which were trained on the 10 surface-relevant bands of Sentinel-2 only.

B. Classification Performance

Our trained networks achieve an average accuracy score between 0.61 and 0.75 (see also Fig. 5), which is comparable to the performance of the available networks pretrained on only 10 bands of Sentinel-2 [2]. In the following parts, we take the ResNet50 network as a representative use case for our further evaluations. The network can be seen as representative in a way that the presented findings based on the application of GradCam hold for all the trained networks. In contrast to the subset of clear images, the networks achieve only average accuracy scores between 0.26 and 0.32 on the cloudy test data. This denotes a considerably detrimental effect of clouds on the model's classification performance, in line with the high cloud coverage rates reported in Section II-A. In Fig. 5, the effects of the clouds on the accuracy, the average accuracy, and the confidence are illustrated. In general, the largest value within a network's soft-max output vector can be interpreted as the model confidence. Networks where the predicted probability represents the actual fraction of correct predictions are called calibrated while uncalibrated networks lead to over- or underconfident predictions [25]. We indicated the confidence by the average over the highest probabilities received from the network for the single samples. While there is a clear drop in classification performances, there is considerably less decrease in confidence.

Complementary, Fig. 6 details the performance of the ResNet50 network for different ranges of cloud coverage. The analysis shows that classification performances decrease with an increase in cloud coverage while confidence stays high.

To attribute the decrease in performance to specific land types, we analyze the confusion matrices for clear and for cloudy observations shown in Fig. 7(a) and (b), respectively. For the cloud-free data, class 4 (*Grasslands*) is often confused with other types, specifically with class 6 (*Croplands*). The presence of clouds generally results in more misclassifications, but, in particular, reinforces the bias of predicting class 5 (*Wetlands*). Remarkably, especially the already harder-to-differentiate vegetation classes are much more distracted by the (partial) cloud cover with a clear bias toward class 4 and class 6.

IV. Analysis of Cloud Effects

A. Separability and Out-of-Distribution Analysis

The eventual occurrence of clouds poses the question of whether a given set of samples can be divided into cloudy and noncloudy images, solely based on a neural network's output.

[2]https://github.com/schmitt-muc/SEN12MS

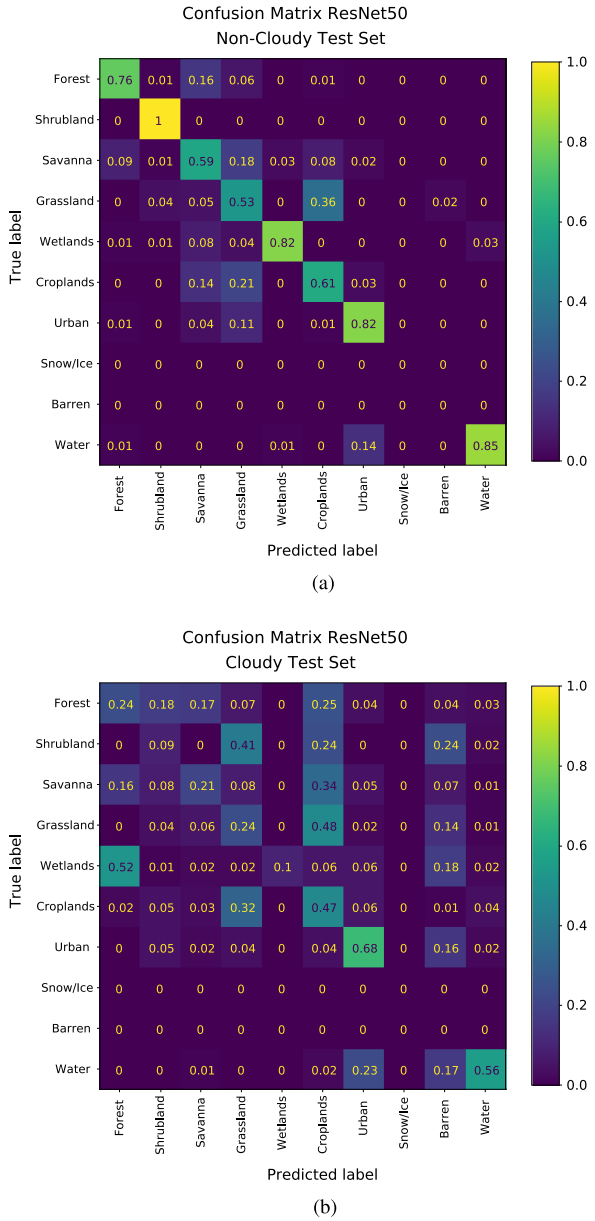


Fig. 7. Confusion matrices of the cloudy and cloud-free test samples resulting from the intersection of SEN12MS and SEN12MS-CR. The true class labels are plotted versus the predicted class labels, with the row-normalized probabilities color-coded. Specifically, class 4 (*Grasslands*) is often confused with others, in particular with class 6 (*Croplands*). The presence of clouds generally results in more misclassifications, but in particular reinforces the bias of predicting class 6 (*Croplands*). Remarkably, the already harder-to-differentiate vegetation classes are much more distracted by the (partial) cloud cover with a clear bias toward classes 4 and 6. (a) Confusion matrix on cloud-free data. (b) Confusion matrix on cloud-covered data.

This can be seen as a case of Out-of-Distribution detection, which is a broadly studied topic in the field of machine learning [25], [32] and also applied in different remote sensing scenarios [26]. In order to evaluate the out-of-distribution detection performance of a classifier, one in general evaluates how well metrics can be used to separate a given test dataset into so-called in-distribution samples (in our case the noncloudy samples) and out-of-distribution samples (in our case the cloudy samples). For every classification neural network, one can apply different metrics on the logit values as well as on the predicted probability vector. The motivation behind this analysis is driven by findings that predictions for data points from unknown data distributions might give a very confident prediction, but often differ considerably in the pure network output, the so-called logits [26]. An ideal model confronted with cloudy samples would express its uncertainty for example by a low confidence value or a high entropy in the resulting probability vector. Also, the features derived from a cloudy sample would fit relatively bad to the possible classes, and therefore, the predicted logit values should be small for all classes. Popular metrics are for example the *maximum probability* (or confidence), the *mutual information*, the *entropy*, the sum of the logit values (*log-sum*), and the *precision*. The precision is motivated by the Dirichlet distribution (a multivariate generalization of the Beta distribution) and can be interpreted as a description of the certainty on the predicted probability vector [25]. The precision is computed as the sum of the exponential of the logit values and the larger the precision value, the less variation in the prediction is assumed. In this article, we investigate the separability of cloudy and noncloudy samples based on the maximum probability, the entropy, the mutual information, the sum of the logit values, and the precision value.

B. Grad-CAM for Saliency Map Computation

Complementary to analyzing the effects of clouds on the scene classification performance via established statistics, we use Gradient-weighted Class Activation Mapping (Grad-CAM) [33] to inspect the workings of the considered classifier when facing noisy optical data. Grad-CAM is a popular method to analyze which input region of an image contributed most to a given prediction. Grad-CAM can be applied post-hoc to a trained network to provide heat maps M_c of the models' attention on the image conditioned on a specific target class c, so-called saliency maps. To do so, the derivative $\frac{\delta y_c}{\delta A_k}$ of the output logit y_c for the conditioned class c with respect to the feature maps A_k is computed. The gradients are then global average pooled across the spatial dimensions H and W to obtain mapwise attention weightings

$$\alpha_{c,k} = \frac{1}{H \times W} \Sigma_{i=1,\ldots,H} \Sigma_{j=1,\ldots,W} \frac{\delta y_c}{\delta A_{k,i,j}}$$

which can be interpreted as the attribution of feature map A_k to drive the classification of c. The feature maps A_k at that layer are averaged across all output channels and the gradients for each channel are weighted by the respective layer's activations $\alpha_{c,k}$ in a simple linear combination. On the resulting pixelwise attribution of activations, a rectified linear unit σ is applied

$$M_c = \sigma(\Sigma_k \alpha_{c,k} A_k)$$

and the saliency map M_c is upsampled via bilinear interpolation to the dimensions of the input image. The resulting attention map specifies which areas in a given input to the network drive its classification as a scene of class c. We utilize Grad-CAM to analyze which regions of a land cover are salient in classifier's

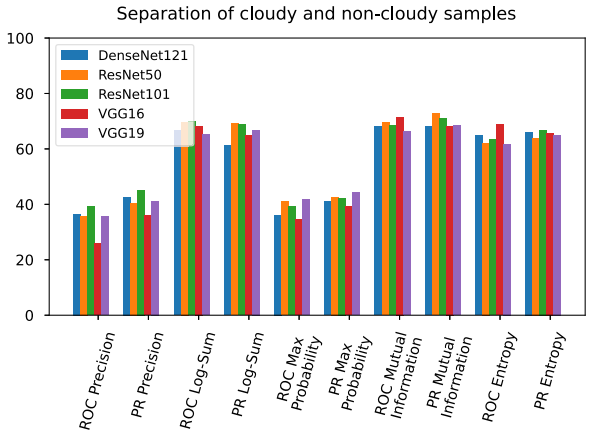


Fig. 8. Separability of samples from the cloudy and from the clear dataset, based on the PR and receiving-operating-characteristic (ROC) of different metrics applied to the output of different network architectures proposed in [2]. The evaluation shows that cloudy and noncloudy samples affect the output of different architectures differently. The best separability is reached with the VGG16 and the ResNet50 architecture and the mutual information metric, followed by the ResNet101 and the DenseNet121 architectures.

receptive fields, and how the presence of clouds affects these saliency maps.

V. Results

A. Effects on the Network Output

This section details the effects of clouds on the network output, including the predictions before applying the soft-max function to compute the categorical probability vectors. We utilize the metrics defined in Section IV-A and analyze the separability of the set of cloudy samples (with a coverage of at least 10%) to their cloud-free pairings and present in Fig. 8 the outcomes for the considered metrics in terms of the average under the curve of precision recall (PR) as well as the receiver operating characteristic curve (ROC). There is an effect of the different architectures on separability, dependent on the considered metric. Overall, separability works best for the mutual information and the entropy metric and the VGG16 architecture, followed by the ResNet50 and the DenseNet121 architecture. It is important to realize that a perfect separability, i.e., a value of 100, is unrealistic to reach in our setup, since several samples are only covered by clouds on a small fraction or do not contain any thick clouds at all (cmp. Fig. 2).

B. Feature Importance and Network Focus

To further analyze what drives misclassifications in the presence of clouds, we apply Grad-CAM to compute saliency maps as detailed in Section IV-B. Within our investigation, we encountered four different manners in which clouds affect the network's attention, presented in the following.[3]

[3]Please note that these chosen examples are exemplary in the sense that their class labels and the classifier's predictions are indeed representatives according to the land cover distribution of Fig. 3 and the confusion matrices of Fig. 7: The analyzed cases feature prominent land cover types such as Grassland, Croplands, Urban, and Forest—which, according to Fig. 2, make up a large proportion of the overall test data. Moreover, the considered cases are representative of salient changes to the network's performance. For instance, in the presence of clouds, the TPR of classifying Forest, the ground truth class in Fig. 13, drops drastically from 0.76 to 0.24. Meanwhile, the FPR to confuse Forest with croplands increases from 0.01 to 0.25, as shown in Fig. 7. As another example, Fig. 11 illustrates a confusion between the ground truth Grassland and the prediction of Cropland. In the presence of clouds, the FPR of this confusion is at 0.48, which is twice as large as the TPR of predicting Grassland correctly as shown in Fig. 7(b).

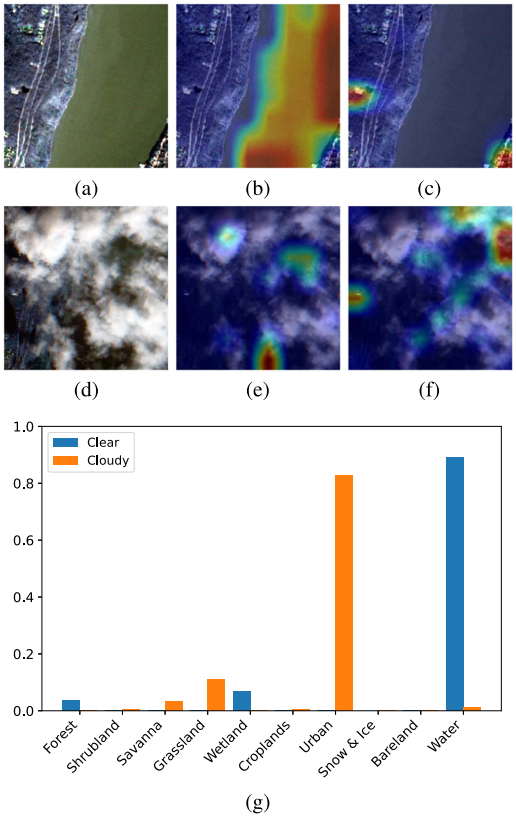


Fig. 9. (a) Clear and (d) 77% cloud-covered image with ground truth class water corresponding saliency maps with respect to the (b) and (e) classes water and (c) and (f) urban. In (g), the network's predictions are shown. This is an example of data where clouds partially cover the image such that homogeneous features are covered but "small feature classes" are still visible. Specifically, the few small buildings visible on the very edge of the image, and the small clouds, cause this confident misclassification.

1) Clouds Partially Cover the Image Such That Homogeneous Features are Covered But "Small Feature Classes" are Still Visible: Depending on the type of cloud coverage, a few clear features can already be enough to make the network predict a specific class with a high confidence value. Especially the urban class is an example of such behavior. Complementing Fig. 1 with the corresponding Grad-CAM results, Fig. 9 illustrates the saliency maps of a water-type land cover scene for both cloudy and cloud-free views. Evidently, the correct *Water* classification focuses on the whole water body, whereas the *Urban* misprediction is driven by the peripheral urban parts not covered by clouds. In both cases, the scenes are (in-)correctly classified at very high confidence, as shown in Fig. 1. Interestingly, the confidence of the network on the cloudy sample prediction is 86%, compared to 90% for the water prediction on the clean image.

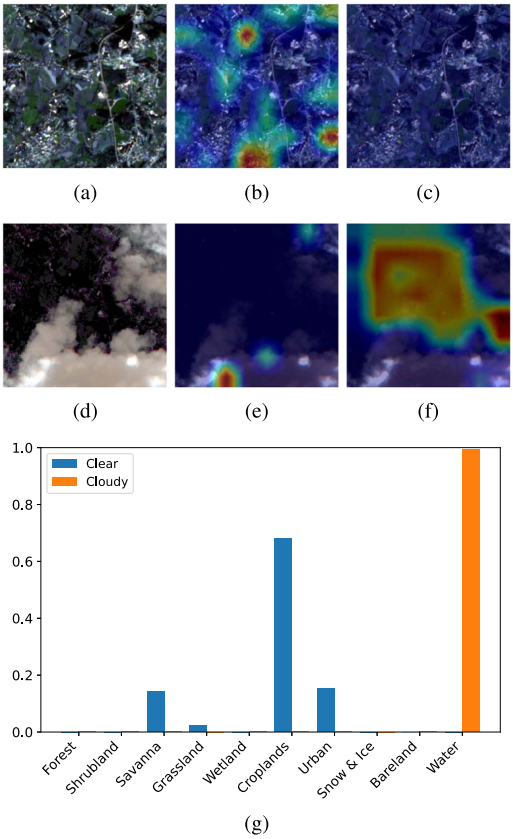


Fig. 10. (a) Clear and (d) 19% cloud-covered input images with ground truth class croplands and corresponding saliency maps for the (b) and (e) classes croplands and (c) and (f) water. In (g), the network's predictions on the clear and on the cloudy image are visualized. The illustrated example shows the case of large cloud shadow regions causing a confident misclassification. It is representative in featuring the majority class "croplands," constituting a large part of our dataset.

2) Structures are Hidden by Shadows: Even clouds that cover an image only partially on a small fraction can still have a considerable effect on the image caused by their shadow. Optical sensors are sensitive to illumination and large shadows impact the illumination. Based on this, shadows can hide structures and characteristics on the floor, leading to a more homogeneous-looking area. In Fig. 10, a very inhomogeneous side is visualized. As shown in Fig. 1, the confidence in the predictions is hence not very large. In contrast to this, the cloudy version covers most of the picture in a very dark monotonic-looking side. As a result, the network predicts the sample as a water body with high confidence. While the saliency map for the clear image shows several single regions that caused the correct prediction, the saliency map of the cloudy version clearly shows that the shadow caused the false prediction as a water body.

3) Small Clouds and Their Shadows Make the Ground Look Less Homogeneous: Clouds and their shadows cannot only cause homogeneity but also make images look more inhomogeneous. Especially many small clouds with many corresponding shadows make the image indicates more structure in the land side as their actually is. In Fig. 11, the cloud-free patch is accurately classified as "grassland." The cloudy patch of 40% cloud coverage is misclassified as "croplands." The corresponding saliency maps clearly show that while for the correct prediction on the clear image, most of the image is taken into account, the false prediction on the cloudy image is based mainly on cloudy and shadow parts of the image.

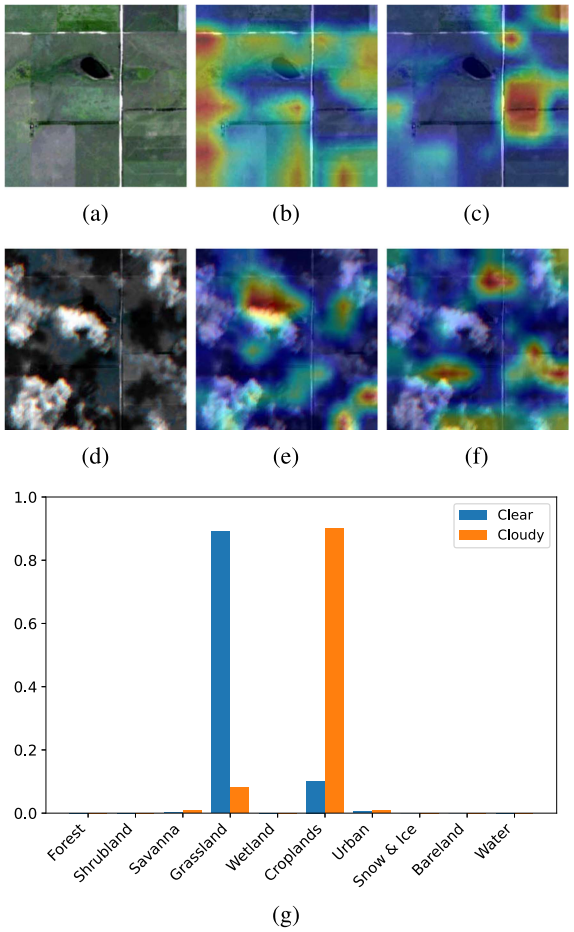


Fig. 11. (a) Clear and (d) 40% cloud-covered input images with ground truth class grasslands and corresponding saliency maps for the (b) and (e) classes grasslands and (c) and (f) croplands. In (g), the network's predictions are shown. This is a case of small clouds and their shadows making the ground look less homogeneous. Altogether, one can clearly see that the intensity of cloudy pixels and their high-contrast neighborhood capture the network's attention and result in misclassification. The shown misclassification is representative for many cases, as croplands are erroneously predicted twice as often as the correct class of grassland in the presence of clouds, according to Fig. 7(b).

4) Homogeneous and Semitransparent Clouds Make Ground Look More Homogeneous: Besides the above-considered non-transparent clouds with clear shapes and shadows, there also exist semitransparent and very homogeneous clouds. In Figs. 12 and 13, two examples are shown where these types of clouds lead to a wrong water and a wrong croplands prediction, respectively.

VI. Discussion

Following the four levels of analysis provided in Section V, this section communicates an interpretation of the observed results. The provided interpretations follow the preceding four stages of analysis to detail our views on the effects of clouds, from the raw data to network decisions and clarify how each step relates to one another.

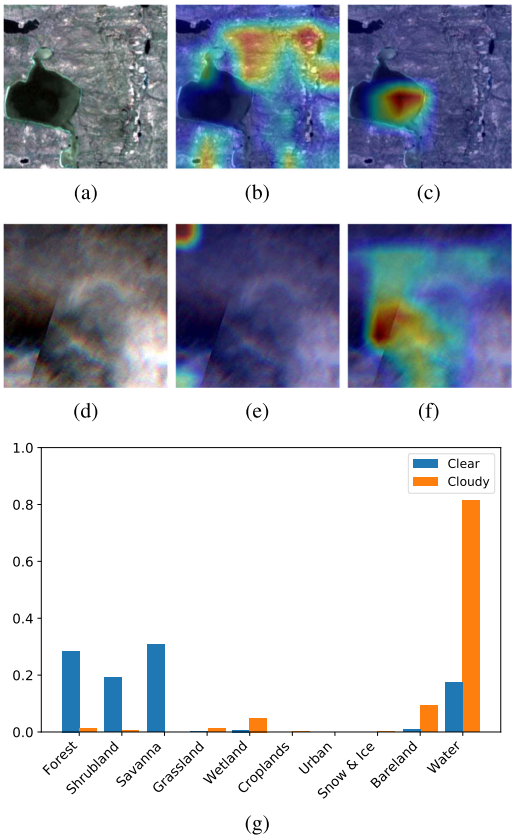


Fig. 12. (a) Clear and (d) 100% cloud-covered input images with ground truth class savanna and corresponding saliency maps for the (b) and (e) classes savanna and (c) and (f) water. In (g), the network's predictions are shown. The shown samples represent the case of homogeneous and semi-transparent clouds making ground appear more homogeneous. Altogether, one can clearly see that the lower contrast and the dark water shimmering through the clouds result in the water prediction.

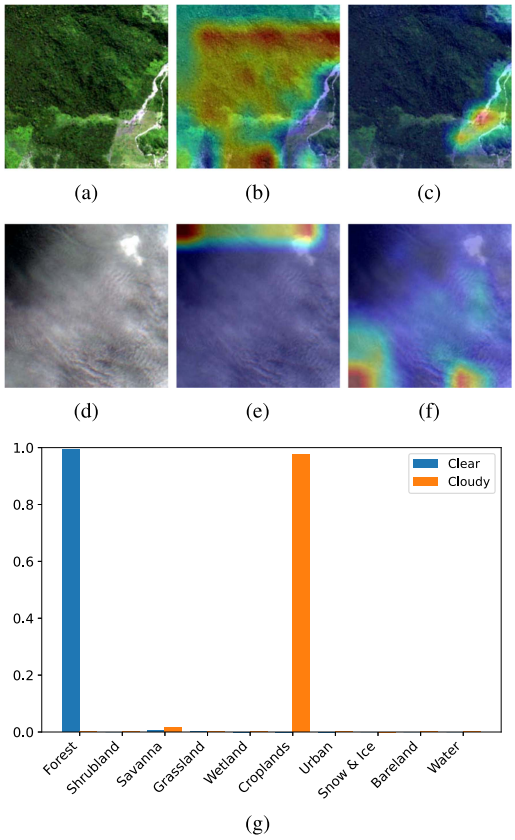


Fig. 13. (a) Clear and (d) 87% cloud-covered input images with ground truth class forest and corresponding saliency maps for the (b) and (e) classes forest and (c) and (f) croplands. In (g), the network's predictions are shown. This is a case of homogeneous and semitransparent clouds making ground appear more homogeneous. Specifically, some small regions with structured clouds result in the croplands prediction. This sample is represented as, in the presence of clouds, the correct classification of "forest" drops to a third of the original rate. Moreover, the probability of misclassifying "forest" as "croplands" outgrows the chance of a correct prediction, as analyzed in Fig. 7.

1) Distribution Shift: As presented in Section II-B, the presence of clouds changes the bandwise data statistics. That is, an overall shift in the data distribution is observable. Distribution shifts have previously been shown to make the behavior of neural networks unreliable and sensitive to misinterpretations caused by very confident but false predictions [25], [26]. Moreover, the bandwise standard deviations increased considerably. This, in return, causes the individual land cover classes to be less separable on their spectral statistics alone. While convolutional neural networks do also incorporate spatial information via local context, the spectral statistics of a sample become less indicative of its class belongings. Finally, preprocessing pipelines based on statistics priorly computed on the cloud-free training data (as in [2]), are no longer appropriate as they do not match the cloudy data distribution and thus do not normalize the cloud-covered data.

2) Classification Performance and Overconfidence: The performance and confidence metrics presented in Section III-B indicate that the classifier is oblivious to the presence of previously unencountered clouds and their effects caused by the shift in the input data distribution as described in Section II-B. Interestingly, the drop in the accuracy is not uniformly distributed, but the confusion matrix in Fig. 7(b) shows a bias toward particular classes. Moreover, this bias is not toward the class with the most training samples (savanna). In addition to the biased decrease in classification performance, the classifier's high overconfidence in the cloudy samples is an undesirable effect caused by clouds. Even though the data are very different from the data known from the training (as seen in the band statistics), the network still gives predictions with high confidence. This behavior is in line with prior observations that neural networks are overly confident in their predictions even in the presence of noise and on changing data domains and distributions [25], [34].

3) Cloudy Noncloudy Separability: Even though the clouds have such a strong effect on the classification performance, the results in Section V-A showed that the separation between cloudy and noncloudy images based on different metrics on the network output is only possible to a certain extent. Even the most discriminative network architectures and measures can only separate in-distribution from out-of-distribution samples in roughly two-thirds of the considered cases. This behavior was also observed when the threshold for the cloud coverage was increased from 10% to a larger value or even to 100%. Besides this, the classifiers and metrics also differ in the extent to which

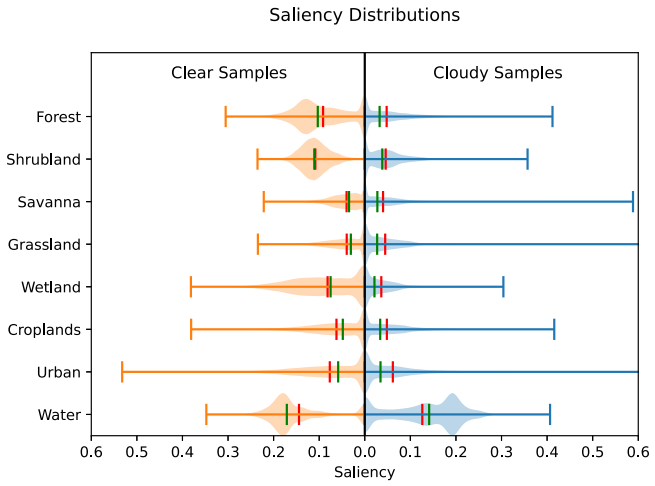


Fig. 14. Violin plot visualization of the pixelwise saliencies with respect to the correct class over the noncloudy (left) and cloudy (right) test data. The violin bodies indicate a smoothed empirical probability distribution, per class. The mean and the median intensities for each class are given by the red (mean) and the green (median) line, respectively. For the clear case, the plots clearly show classwise differences in the average number of pixels and intensities contributing to the prediction. For the cloudy case, they show a reduction of saliency for all classes but the water class.

they can grasp differences between representations of cloudy and cloud-free images. Especially the performances based on the precision value and the maximum probability underline the findings that networks are overly confident and further support the interpretation that the scene classifier is oblivious to the presence of previously unencountered clouds.

4) Outliers as Distractors: As evidenced by the Grad-CAM analysis in Section V-B, cloud coverage poses an obstacle to land cover classification in four different kinds: First, clouds partially cover the image such that large areas are covered but relatively irrelevant "small feature classes" may still be visible. Second, otherwise apparent structures may be hidden by cloud shadows. Third, small clouds and their shadows make the ground look less homogeneous. Fourth, transparent clouds lead to a different representation of the (often already on clear images hard to differentiate) classes of land cover. These four cases can be directly related to the shift in the confusion matrix represented in Fig. 7(b), as, for example, the large shift from water to urban classes can be explained and the interplay of houses and water was presented as shown in Fig. 9. Moreover, samples across all four cases highlight that the network's spatial attention often shifts toward clouds, their shadows, or the transition between both. That is, outstandingly bright, very dark, or high-contrast areas often coincide with a focus of attention. As these are oftentimes entailed by the presence of clouds, we interpret that out-of-distribution image intensities function as a distractor. In sum, clouds and their shadows distract classifiers on a macroscale by obscuring large areas—but also on a per-pixel level, as cloud or cloud-shadow induced intensity changes equally distract the classifier from the actual land cover. Moreover, the evaluation of the pixelwise saliencies in Fig. 14 shows that all areas of the water land cover type contribute to a relatively high relevance. In contrast, for the more feature-based urban class, the majority of the class is not that relevant for the prediction. At the same time, the values for the urban saliency sometimes reach larger values than for all other classes. Interestingly, an equivalent but less significant trend of larger areas of an image into account also appears for the forest and the shrubland class. For the Wetland class, the relevance is not that concentrated on single values but seems to also take a variety of areas into account. Those classes also have the largest relative drop in the true positives, indicating that the coverage of clouds harms these types of classes more than those, which focus on smaller areas.

Hence, clouds and shadows covering parts of an image and hiding information for specific areas affect the scenewise classification of regions differently. When comparing the pixelwise saliency of cloud-free data to one of cloudy samples, a clear decrease in saliency is visible while the outliers become more extreme. That is, on average, a smaller fraction of a scene's pixels contributes to its classification in the presence of clouds, except for a few extrema. This finding validates the hypothesis that less information covering multiple pixels leads to class prediction but mainly local information. This is also what the presented saliency maps, except of the false water prediction in Fig. 10, indicate. Fewer pixels driving a classification are in contrast to the majority principle that the most prominent class (i.e., the one covering the largest area) defines a scene's label. The only exception from the trend of shrinking saliencies is the Water class, for which larger areas of cloud shadows in other scene types tend to be misclassified as water. Altogether, the presented analysis clearly shows that conventionally (pre-)trained networks are not fit for domain shifts in data common in remote sensing. Specifically, the derived features cannot be used to give a strong idea of the underlying class, even if only parts of the image are covered by clouds.

Overall, our multistage analysis reveals that the effects of clouds on remote sensing applications manifest in many different aspects of the pipeline, from the raw data to the information a trained network extracts from these images. As the visualizations and evaluations of the Grad-CAM images underlined, the structure caused by clouds and their shadows contain misleading information leading to very confident but false predictions.

While our analyzed data comprise a large cohort of globally distributed regions acquired through several seasons that should be sufficiently heterogeneous and representative, our analysis may nonetheless be dependent on, e.g., the choice of datasets and cloud detection algorithms. For instance, future work may conduct our analysis focused on a single-country level, e.g., on the dataset in [35]. Moreover, recent publications have provided novel large-scale datasets for cloud detection or removal in time series [12], [13], [36], which may serve as an extended version of our analysis. With respect to the cloud detector algorithm, s2cloudless was chosen for being commonly deployed, easily applicable, and performing well [37], [38]. However, many alternative approaches exist [35], [39], [40], [41], [42], whose variable sensitivity thresholds may result in qualitatively different cloud masks and thus different downstream analysis results. The chosen s2cloudless algorithm is reported to show a fair "balance (within 10%) between commission and omission errors" [38], which may avoid any one-sided biases to either false alarms or misses of clouds in our subsequent analysis.

VII. Conclusion

With over 50% of our planet's surface covered by clouds at any point [6], haze and clouds pose a considerable obstacle to the continuous monitoring of Earth. In this work, we investigated in detail the effects of clouds on a deep neural network performing remote sensing scene classification. To start with, clouds considerably alter the spectral characteristics of data and make individual land cover types less separable from one another. In terms of performance, we observed a considerable drop in overall classification accuracy to almost half of the rates at clear views. A confusion matrix analysis revealed that existing biases toward predicting certain classes are reinforced in the presence of clouds. Even though the network remains highly confident in its predictions, it cannot separate between cloud-free and cloud-covered observations—indicating the classifier's unawareness of clouds. Finally, we complemented the reported statistics with a qualitative analysis of the classifier's attention maps. The saliency maps highlighted that clouds distract the network from the actual land cover surface. That is, rather than focusing on the actual land cover, previously unseen noise is so salient that it becomes the focus of the classifier's attention. These insights contribute to a better understanding of the effects of clouds on remote sensing applications and may consequently guide the future development of more robust models. We plan to continue our research and develop a methodology that is more robust to the effects of outliers and noise detailed in this contribution. For future approaches, evaluating the distribution of image regions relevant to the prediction is an interesting way to identify misconceptions and misclassifications. In addition, training methods that incorporate clouds and shadowy regions and can express the uncertainty and the lack of knowledge due to obscured parts of the image are a promising route to more robust approaches in the future.

Acknowledgment

The authors would like to thank Dharani Deivasihamani for the support in implementing the evaluation pipeline and visualizing preliminary results.

References

[1] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

[2] M. Schmitt and Y.-L. Wu, "Remote sensing image classification with the SEN12MS dataset," *ISPRS Ann. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. V- 2-2021, pp. 101–106, 2021.

[3] H. Sun, Y. Lin, Q. Zou, S. Song, J. Fang, and H. Yu, "Convolutional neural networks based remote sensing scene classification under clear and cloudy environments," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Oct. 2021, pp. 713–720, doi: 10.1109/ICCVW54120.2021.00085.

[4] M. Rafique, H. Blanton, and N. Jacobs, "Weakly supervised fusion of multiple overhead images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1479–1486.

[5] W. Kang, Y. Xiang, F. Wang, and H. You, "CFNet: A cross fusion network for joint land cover classification using optical and SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1562–1574, Jan. 2022, doi: 10.1109/JSTARS.2022.3144587.

[6] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by MODIS onboard the terra and aqua satellites," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 3826–3852, Jul. 2013.

[7] D. Spänkuch, O. Hellmuth, and U. Görsdorf, "What is a cloud? Toward a more precise definition?," *Bull. Amer. Meteorol. Soc.*, vol. 103, pp. E1894–E1929, Mar. 2022, doi: 10.1175/BAMS-D-21-0032.1.

[8] Y. Chen, L. Tang, X. Yang, R. Fan, M. Bilal, and Q. Li, "Thick clouds removal from multitemporal ZY-3 satellite images using deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 143–153, Dec. 2020, doi: 10.1109/JSTARS.2019.2954130.

[9] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 333–346, 2020.

[10] Y. Zi, F. Xie, N. Zhang, Z. Jiang, W. Zhu, and H. Zhang, "Thin cloud removal for multispectral remote sensing images using convolutional neural networks combined with an imaging model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3811–3823, Mar. 2021, doi: 10.1109/JSTARS.2021.3068166.

[11] J. Li, Z. Wu, Z. Hu, Z. Li, Y. Wang, and M. Molinier, "Deep learning based thin cloud removal fusing vegetation red edge and short wave infrared spectral information for sentinel-2a imagery," *Remote Sens.*, vol. 13, no. 1, 2021, Art. no. 157.

[12] P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu, "SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5222414, doi: 10.1109/TGRS.2022.3146246.

[13] A. Sebastianelli et al., "PLFM: Pixel-level merging of intermediate feature maps by disentangling and fusing spatial and temporal data for cloud removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5412216, doi: 10.1109/TGRS.2022.3208694.

[14] M. Rußwurm and M. Körner, "Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 11–19.

[15] Z. Gu, P. Ebel, Q. Yuan, M. Schmitt, and X. X. Zhu, "Explicit haze & cloud removal for global land cover classification," in *Proc. Comput. Vis. Pattern Recognit. Conf. Workshop Multimodal Learn. Earth Environ.*, Jul. 2022, pp. 1–6. [Online]. Available: https://elib.dlr.de/186738/

[16] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[17] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, Jun. 2020, doi: 10.1109/JSTARS.2020.3005403.

[18] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigEarthNet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5901–5904.

[19] I. Kakogeorgiou and K. Karantzalos, "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, 2021, Art. no. 102520.

[20] P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu, "Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5866–5878, Jul. 2021.

[21] M. Schmitt, L. Hughes, C. Qiu, and X. Zhu, "SEN12MS—A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Ann. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. IV-2/W7, pp. 153–160, 2019.

[22] M. A. Friedl et al., "Global land cover mapping from MODIS: Algorithms and early results," *Remote Sens. Environ.*, vol. 83, no. 1/2, pp. 287–302, 2002.

[23] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sens. Environ.*, vol. 202, pp. 18–27, 2017.

[24] A. Zupanc, "Improving cloud detection with machine learning," *Sentinel-Hub*, 2017, Accessed: Oct. 10, 2019. [Online]. Available: https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13

[25] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," 2021, *arXiv:2107.03342*.

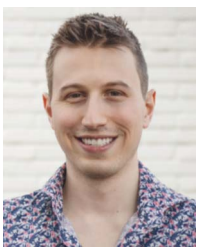
[26] J. Gawlikowski, S. Saha, A. Kruspe, and X. X. Zhu, "An advanced Dirichlet prior network for out-of-distribution detection in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5616819, doi: 10.1109/TGRS.2022.3140324.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015. [Online]. Available: https://iclr.cc/archive/www/doku.php%3Fid=iclr2015:main.html
[30] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, May 2015. [Online]. Available: https://iclr.cc/archive/www/doku.php%3Fid=iclr2015:main.html
[31] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8026–8037.
[32] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 7047–7058.
[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
[34] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, Dept. Eng., Univ. Cambridge, Cambridge, U.K., 2016.
[35] J. Li et al., "A lightweight deep learning-based cloud detection method for Sentinel-2A imagery fusing multiscale spectral and spatial features," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5401219, doi: 10.1109/TGRS.2021.3069641.
[36] C. Aybar et al., "CloudSEN12–A global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2," *Zenodo*, Aug. 2022.
[37] J. Braaten, K. Schwehr, and S. Ilyushchenko, "More accurate and flexible cloud masking for Sentinel-2 images," *Medium*, 2020, Accessed: Oct. 16, 2022. [Online]. Available: https://medium.com/google-earth/more-accurate-and-flexible-cloud-masking-for-sentinel-2-images-766897a9ba5f
[38] S. Skakun et al., "Cloud mask intercomparison exercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2," *Remote Sens. Environ.*, vol. 274, 2022, Art. no. 112990.
[39] Z. Li, H. Shen, Q. Weng, Y. Zhang, P. Dou, and L. Zhang, "Cloud and cloud shadow detection for optical satellite imagery: Features, algorithms, validation, and prospects," *ISPRS J. Photogrammetry Remote Sens.*, vol. 188, pp. 89–108, 2022.
[40] L. Sun et al., "A new cloud detection method supported by GlobeLand30 data set," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 10, pp. 3628–3645, Oct. 2018.
[41] H. Guo, H. Bai, and W. Qin, "ClouDet: A dilated separable CNN-based cloud detection framework for remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9743–9755, Sep. 2021, doi: 10.1109/JSTARS.2021.3114171.
[42] D. López-Puigdollers, G. Mateo-García, and L. Gómez-Chova, "Benchmarking deep learning models for cloud detection in Landsat-8 and Sentinel-2 images," *Remote Sens.*, vol. 13, no. 5, 2021, Art. no. 992.

Jakob Gawlikowski received the B.Sc. and M.Sc. degrees in mathematics in 2015 and 2019 from the Technical University of Munich, Munich, Germany, where he is currently working toward the Ph.D. degree in robust data fusion with the Chair of Data Science in Earth Observation, Department of Aerospace and Geodesy.

He is currently a Researcher with the German Aerospace Center's (DLR) Institute of Data Science, Jena, Germany. His research focuses on data fusion machine learning approaches with a special focus on uncertainty quantification and robustness in deep learning models.



Patrick Ebel received the B.Sc. degree in cognitive science from the University of Osnabrü, Osnabrück, Germany, in 2015, and the M.Sc. degree in cognitive neuroscience and the M.Sc. degree in artificial intelligence from Radboud University Nijmegen, Nijmegen, The Netherlands, in 2018. He is currently working toward the Ph.D. degree in optical satellite image reconstruction with the SiPEO Lab, Department of Aerospace and Geodesy, Technical University of Munich, Munich, Germany.

His research interests include machine learning and its applications in computer vision and to remote sensing data.

Michael Schmitt (Senior Member, IEEE) received the Dipl.-Ing. (Univ.) degree in geodesy and geoinformation, the Dr.-Ing. degree in remote sensing, and the habilitation in data fusion from the Technical University of Munich (TUM), Munich, Germany, in 2009, 2014, and 2018, respectively.

Since 2021, he has held the Chair for Earth Observation with the Department of Aerospace Engineering, University of the Bundeswehr Munich, Neubiberg, Germany. Before that, he was a Professor in applied geodesy and remote sensing with the Department of Geoinformatics, Munich University of Applied Sciences. From 2015 to 2020, he was a Senior Researcher and Deputy Head with the Professorship for Signal Processing in Earth Observation, TUM; in 2019, he was additionally appointed as an Adjunct Teaching Professor with the Department of Aerospace and Geodesy, TUM. In 2016, he was a Guest Scientist with the University of Massachusetts, Amherst. His research focuses on image analysis and machine learning applied to the extraction of information from multimodal remote sensing observations, particularly interested in remote sensing data fusion with a focus on SAR and optical data.

Dr. Schmitt is a Co-Chair of the Working Group "Active Microwave Remote Sensing" of the International Society for Photogrammetry and Remote Sensing, and also of the Working Group "Benchmarking" of the IEEE-GRSS Image Analysis and Data Fusion Technical Committee. He frequently serves as a Reviewer for a number of renowned international journals and conferences and has received several Best Reviewer awards.

Xiao Xiang Zhu (Fellow, IEEE) received the master's (M.Sc.), Doctor of Engineering (Dr.-Ing.), and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is currently the Chair Professor for Data Science in Earth Observation with TUM and was the founding Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR). Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School (www.mu-ds.de). Since 2019, she has also been the head of the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the PI and Director of the International Future AI Lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since October 2020, she has also been a Co-Director of the Munich Data Science Institute (MDSI), TUM. She was a Guest Scientist or Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, the University of Tokyo, Tokyo, Japan, and the University of California, Los Angeles, USA, in 2009, 2014, 2015, and 2016, respectively. She is currently a Visiting AI Professor with ESA's Phi-lab. Her main research interests include remote sensing and Earth observation, signal processing, machine learning, and data science, with their applications in tackling societal grand challenges, e.g., Global Urbanization, UN's SDGs and Climate Change.

Dr. Zhu is a member of young academy (Junge Akademie/Junges Kolleg) with the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is on the scientific advisory board of several research organizations, among others the German Research Center for Geosciences (GFZ) and Potsdam Institute for Climate Impact Research (PIK). She is an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and the Area Editor responsible for special issues of *IEEE Signal Processing Magazine*.

A.6 UnCRtainTS: Uncertainty Quantification for Cloud Removal in Optical Satellite Time Series

Reference: P. Ebel, V. Garnot, M. Schmitt, J. Wegner and X. X. Zhu. *UnCRtainTS: Uncertainty Quantification for Cloud Removal in Optical Satellite Time Series.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2023

UnCRtainTS: Uncertainty Quantification for Cloud Removal in Optical Satellite Time Series

Patrick Ebel* patrick.ebel@tum.de
Vivien Sainte Fare Garnot† vsaint@ics.uzh.ch
Michael Schmitt‡ michael.schmitt@unibw.de
Jan Dirk Wegner† jandirk.wegner@uzh.ch
Xiao Xiang Zhu* xiaoxiang.zhu@tum.de

* Technical University of Munich † University of Zurich ‡ University of the Bundeswehr Munich

Abstract

Clouds and haze often occlude optical satellite images, hindering continuous, dense monitoring of the Earth's surface. Although modern deep learning methods can implicitly learn to ignore such occlusions, explicit cloud removal as pre-processing enables manual interpretation and allows training models when only few annotations are available. Cloud removal is challenging due to the wide range of occlusion scenarios—from scenes partially visible through haze, to completely opaque cloud coverage. Furthermore, integrating reconstructed images in downstream applications would greatly benefit from trustworthy quality assessment. In this paper, we introduce UnCRtainTS, a method for multi-temporal cloud removal combining a novel attention-based architecture, and a formulation for multivariate uncertainty prediction. These two components combined set a new state-of-the-art performance in terms of image reconstruction on two public cloud removal datasets. Additionally, we show how the well-calibrated predicted uncertainties enable a precise control of the reconstruction quality.

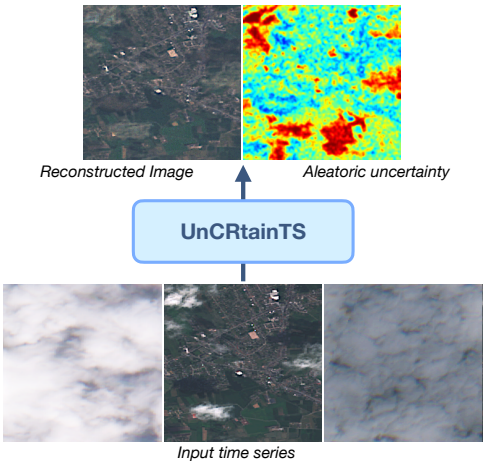


Figure 1. **Overview:** Our attention-based UnCRtainTS architecture predicts a single cloud-free image from a sequence of cloudy observations. For each reconstructed pixel, our method also estimates the aleatoric uncertainty of the prediction. Note how higher uncertainties (in red) are associated with persistent occlusion, cloud shadow, or with specific land cover types.

1. Introduction

Multispectral, optical satellite imagery allows for large-scale assessments of the environment like crop monitoring [58,71] and global vegetation height estimation [45,46]. Clouds, haze and other atmospheric disturbances, however, often occlude large parts of optical satellite images, particularly during meteorological winter season [40] and over landcover such as rainforests [4]. Neural networks trained on extensive amounts of annotated data may implicitly learn to ignore task-irrelevant cloudy observations [55, 58, 59]. Yet, explicit cloud removal as a pre-processing step can further improve model performance and is valuable if ground truth annotations for supervised training are scarce [30]. Cloud removal prior to training or applying a pre-trained task-specific model also permits a seamless analysis using traditional non-learning methods or visualisation [51].

Hence, cloud removal is an active field of research boasting a large body of literature on image reconstruction methods to recover cloud-free observations [4, 12, 17, 20, 29, 54, 61, 62]. Such methods are typically evaluated in terms of image restoration metrics, e.g. mean squared error or structural similarity (SSIM), providing an aggregated measure of reconstruction quality. These metrics, however, provide little insight into how reliable a given reconstruction is on a pixel-wise or image-by-image basis. To address this shortcoming, we introduce uncertainty estimation to satellite im-

age reconstruction, specifically to the task of multi-temporal cloud-removal in optical satellite images. Predicting uncertainties that correlate with the empirical errors of a neural net is at the core of the growing field of probabilistic deep learning [39, 65, 68]. By modelling the uncertainty and training for a negative log likelihood (NLL) objective, such approaches allow to jointly learn a model for making a prediction and estimate the prediction's variances. If well-calibrated, the predicted uncertainties can be very valuable for downstream usage by providing a measure of a reconstruction's confidence. Uncertainty quantification has been successfully applied in univariate remote sensing regression problems such as canopy height regression [46] or flood risk estimation [8]. Here, we extend uncertainty quantification to multivariate regression for satellite image reconstruction. We obtain experimentally well-calibrated uncertainties that enable flagging poorly reconstructed images. We also show that multivariate uncertainty prediction requires a multivariate uncertainty model for better calibration.

Aleatoric uncertainty prediction implies training with a pixel-based Negative Log Likelihod (NLL) loss. On the other hand, image reconstruction losses like SSIM or perceptual loss are typically used in existing cloud removal methods to better retrieve high-frequency details [10, 12, 74]. Here, we introduce a novel neural architecture that operates on feature maps at full resolution. It leverages attention-based temporal encoding, allowing it to outperform previous state-of-the-art approaches even when trained via a pixel-based loss. In sum, our contributions are:

- We introduce multivariate uncertainty quantification to the task of multispectral satellite image reconstruction, to obtain both reconstructions and variance estimates.
- We propose a novel neural network architecture achieving state-of-the-art results on two challenging benchmark datasets for optical satellite cloud removal.
- We obtain well-calibrated uncertainties that allow to measure and control the quality of reconstructed images for risk-mitigation in downstream applications.

2. Related Work

2.1. Cloud Removal in Satellite Image Time Series

Optical satellite image reconstruction [64], and specifically cloud removal, pose a long-standing challenge in remote sensing [15, 33, 35, 49, 50]. Contemporary deep learning approaches can be categorised into mono-temporal [4, 17, 20, 56, 75], mono-temporal & multi-modal [12, 29, 54], multi-temporal [61] and multi-temporal & multi-modal methods [14, 62]. Here, we consider the reconstruction task in a multi-temporal & multi-modal setting.

Spatial encoding of image reconstruction is either done with UNet-like encoder-decoder backbones [37, 57, 76] that spatially down-sample the intermediate representations [12, 17, 29], or with architectures preserving the full resolution of the images [44, 54]. While the first are computationally more efficient especially in the multi-temporal setting, the latter tend to better preserve the spatial structure in the reconstructed images. In fact, downsampling architectures often necessitate auxiliary perceptual [12, 13, 36, 38] or structural similarity losses [72, 73] to recover high-frequency information. The combination of such cost functions with a probabilistic training objective for uncertainty prediction is not straightforward. Therefore, we design an architecture that operates on full resolution feature maps and make design choices to reduces its computational complexity. For temporal encoding, we draw inspiration from recent work in satellite time series encoding [21, 22, 59] and rely on self-attention to integrate the temporal information.

2.2. Uncertainty Quantification

Uncertainty can be partitioned into *epistemic* or model uncertainty, and *aleatoric* or data uncertainty. Epistemic uncertainty accounts for the uncertainty on the model's weights, and can be estimated for instance with ensemble methods [43, 70], or monte-carlo dropout [19] in deep nets. Aleatoric uncertainty captures the randomness inherent to the data. In the case of optical satellite image reconstruction, aleatoric uncertainty may thus help flagging restorations based on too little evidence. In the recent deep learning literature, aleatoric uncertainty estimation is achieved via likelihood maximization with a parametric model of the noise distribution [1, 63, 65, 67, 68]. This is a common technique in safety-critical applications, such as solving inverse problems in biomedical imaging [2, 5, 9, 16, 27, 47, 48, 69]. Uncertainty quantification is of growing interest in remote sensing [26], with applications to forest assessments, flood hazard monitoring, geophysical modeling, landcover classification and out-of-distribution detection [8, 24, 25, 45, 46, 52]. As prior remote sensing work covers uncertainty quantification for univariate regression problems, the multivariate extension has yet to be explored. To our knowledge, the aforementioned contributions are either on image reconstruction in the biomedical domain or target specific remote sensing downstream tasks, such that ours is the first work to investigate uncertainty quantification for multispectral satellite image reconstruction. The current lack of uncertainty quantification in the cloud removal literature is a significant research gap because reconstructed satellite images may guide safety-critical downstream applications or human judgement alike, such that pixel-wise measures of confidence would be beneficial.

3. Methods

We follow the problem statement of the public cloud removal benchmark SEN12MS-CR-TS [14]. Each sam-

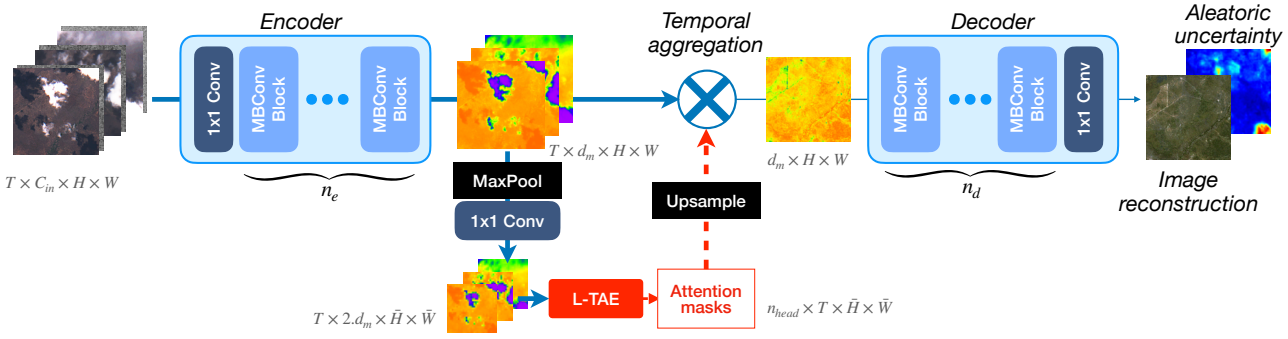


Figure 2. **UnCRtainTS.** The network consists of three main parts, applied along a main branch of MBConv blocks [60] that is processing feature maps at full input resolution: First, an *encoder* is applied in parallel to the T time points. Then, an *attention-based temporal aggregator* computes attention mask by applying an L-TAE to downsampled feature maps, used to aggregate the sequence of observations. Finally, the temporally integrated feature map is processed by a *decoding block*, yielding the image reconstruction and aleatoric uncertainty.

ple i of the N-sized dataset consists of a pair $(\boldsymbol{X}^i, Y^i)$, where $\boldsymbol{X}^i = [X_1^i, \cdots, X_T^i]$ is the input time series of size $[T \times C_{in} \times H \times W]$ containing cloudy pixels, and Y^i is the target cloud-free image of shape $[K \times H \times W]$. T denotes the number of dates in the input sequence, C_{in} and K the number of input and output channels, and $H \times W$ the two spatial dimensions of the images. As in [14], we set $T = 3$, $C_{in} = 15$, $K = 13$, $H = W = 256$. Note that $C_{in} \neq K$ because Sentinel-1 radar observations are utilized as additional input. Furthermore, aleatoric uncertainty quantification introduces additional output channels to describe the modeled noise distribution. For convenience, we drop the i superscript in the rest of this section.

3.1. Network Architecture

Our proposed UnCRtainTS network architecture maps a cloudy input time series to a single cloud-free optical image. As explained in Sec. 2.1, we make the explicit choice to perform spatial encoding only on full-resolution feature maps to allow for good performance when training with a pixel-based loss. To ease the impact of this choice on the computational load of the architecture, we rely on efficient MBConv blocks [60]. They combine depthwise convolution and regular pointwise convolutions for computationally efficient spatial encoding. We perform temporal encoding on downsampled feature maps via the attention-based L-TAE [21], which is designed for satellite image time series and computationally more efficient than transformers. The network architecture is illustrated in Fig. 2 and further described in the following paragraphs.

Pre-aggregation shared encoder The T different input images are processed in parallel by a shared spatial encoding branch. This encoder is composed of a pointwise convolution $C_{in} \rightarrow d_m$, followed by a specifiable number n_e of MBConv blocks. Following [22] we use group normalisation in the encoding branch. All MBConv blocks map to $d_m \rightarrow 2 \times d_m \rightarrow d_m$ channels and contain Squeeze-Excitation layers [34]. Ultimately, each input image X_t is mapped to a feature map f_t of the same resolution.

Attention-based temporal aggregation Following recent literature, we employ self-attention to aggregate a sequence of feature maps $[f_1, \cdots, f_T]$ into a single one. We first down-sample features f_t with a single max-pooling operation to low resolution feature maps $\overline{f_t}$ of size $\left[d_m \times \overline{H} \times \overline{W}\right]$. We set $\overline{H} = \overline{W} = 32$, to limit computation while providing sufficient resolution to group cloudy pixels, which typically cluster in space. We re-project the downsampled features via a linear layer $d_m \rightarrow 2 \times d_m$. Next, as in [22], the low-resolution features $\overline{f_t}$ are processed pixel-wise with an L-TAE [21, 23]: we obtain attention masks over the T observations for each pixel position of the low resolution feature maps. Contrary to previous work, we only use the L-TAE's attention masks, and omit attention-weighting of the sequence of low resolution feature maps. We upsample the attention masks to the full resolution via bilinear interpolation, and apply them to the sequence of high resolution feature maps $[f_1, \cdots, f_T]$. This results in a single feature map $\hat{f}$ of shape $[d_m \times H \times W]$. We use a dropout rate of 0.1 on the attention masks after upsampling, and the temporal aggregation is done with L-TAE's channel grouping strategy [21].

Post-aggregation decoding The temporally aggregated feature map $\hat{f}$ is processed by a decoding branch, which consists of a specifiable number n_d of batch-normalized MBConv blocks and a final $d_m \rightarrow C_{out}$ pointwise convolution followed by a non-linearity. For every channel predicting image reconstruction, we use a sigmoidal function to squash the outputs into the data's valid range. For channels predicting aleatoric uncertainty (see next section), we use a

softplus activation to ensure positivity, as in [32, 63, 67].

3.2. Aleatoric uncertainty prediction

Here, we explain how our UnCRtainTS method predicts an aleatoric uncertainty value for each reconstructed pixel. As UnCRtainTS is trained with pixel-wise losses, we henceforth adopt a pixel-based notation. We consider the set of pixels of cardinal n contained in the dataset. We denote each pixel reconstruction by $\hat{\boldsymbol{y}}_j$ and the corresponding ground truth by $\boldsymbol{y}_j$, both vectors of dimension K.

Image reconstruction In the default setting of satellite image reconstruction, the network only regresses the target pixel values. Hence, in this setting, $C_{out} = K$ and the predictions are typically supervised with L2 loss [3, 11]:

$$\mathcal{L}_2(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{n}\sum_{j=1}^{n} \|\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j\|_2^2 \,. \tag{1}$$

Multivariate negative log-likelihood loss Predicting aleatoric uncertainty assumes a parametric noise distribution with a likelihood function. We then optimise the likelihood of the observed data as a function of the input and the distribution's parameters, using a negative log-likelihood (NLL) cost function [6]. Following the literature [39], we model aleatoric uncertainty on the reconstructed pixel with a K-variate Normal distribution centered at the predicted value $\hat{\boldsymbol{y}}_j$ and with positive definite covariance matrix $\boldsymbol{\Sigma}$:

$$\mathcal{N}(\boldsymbol{y}_j|\hat{\boldsymbol{y}}_j, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}(2\pi)^{\frac{K}{2}}} exp\left(-\frac{1}{2}\|\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j\|_M\right) \,, \tag{2}$$

with $\|.\|_M$ the Mahalanobis distance, defined as:

$$\|\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j\|_M = (\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j)^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j) \,. \tag{3}$$

Subsequently, the negative log likelihood loss writes as:

$$\mathcal{L}_{NLL}(\boldsymbol{y_j}|\hat{\boldsymbol{y}}_{\boldsymbol{j}}, \boldsymbol{\Sigma}) \propto \sum_{j=1}^{n} log(|\boldsymbol{\Sigma}_j|) + \|\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j\|_M \,. \tag{4}$$

Fitting a multivariate distribution raises the question of whether a full description of the covariance matrix should be pursued or if any structural constraints on Σ are preferable. NLL optimization does become notoriously difficult when involving full covariance matrices [63, 65].

Diagonal covariance matrix We define $\boldsymbol{\Sigma}$ as a diagonal matrix with diagonal elements $\boldsymbol{\sigma^2} = (\sigma_1^2, \cdots, \sigma_K^2)$. This greatly simplifies the inverse and determinant computations in Eq. 4. The diagonal model allows for different variance predictions per channel, which we experimentally find to be beneficial. However, cross-channel interactions in aleatoric predictions are not captured under this assumption, and such modelling is left for further research. To predict the variances, we set C_{out} to $2 \times K = 26$. The diagonal entries of $\boldsymbol{\Sigma}$ serve as aleatoric uncertainty prediction for the corresponding output channel:

$$\boldsymbol{u_j} = [u_j^1, \cdots, u_j^K] = [\sigma_1^2, \cdots, \sigma_K^2] \,. \tag{5}$$

4. Experiments

4.1. Data

We conduct our experiments on the SEN12MS-CR [12] and SEN12MS-CR-TS [14] datasets for mono-temporal and multi-temporal cloud removal. Both are challenging image reconstruction benchmark datasets with about 50% cloud coverage over regions distributed across the whole planet and all seasons. The datasets contain ground range detected dual-polarization C-band $S1$ measurements as well as co-registered level-1C top-of-atmosphere reflectance $S2$ products, curated from Google Earth Engine [28] and subsequently handled as documented in the two associated publications. The mono-temporal dataset contains 169 regions, whereas SEN12MS-CR-TS focuses on a global subset of 53 large areas. All regions of the datasets are utilized for training, validation and testing, with the respective splits as originally defined. Unless specified otherwise, experiments on SEN12MS-CR-TS are run on $T = 3$ time points, which is a reasonable number of revisits for the cloud removal task and has been a prevalent choice in prior work [14, 61, 62]. All data are of spatial dimensions $H = W = 256$ px and we use the full spectrum of all 13 optical bands. Analogous to preceding studies combining information of SAR and optical imagery [14, 15, 35, 54, 75] we use both Sentinel-1 and Sentinel-2 data to reconstruct images of the latter (i.e., $C_{S1} = 2$, $C_{S2} = C_{out} = 13$, and $C_{in} = C_{S1} + C_{S2} = 15$). $S1$ data are preprocessed as in [12, 14] and $S2$ pixel-values are divided by 1000. Finally, binary cloud masks are calculated via s2cloudless [77]—a lightweight and commonly deployed cloud detector [7, 66]. The cloud masks are used for sampling cloud-free target images at train time, statistical evaluations of results, and in prior work for losses that are cloud-sensitive [54].

4.2. Implementation details

Architectures We train the proposed UnCRtainTS in its default setting with $n_e = 1$ pre- and and $n_d = 5$ post-aggregation MBConv blocks. The input convolution maps to $d_m = 128$ channels, so that MBConv blocks map to $128 \rightarrow 256 \rightarrow 128$ channels with the default expansion factor 0.25 in their Squeeze-Excitation layers. The L-TAE's parameters are kept to their default values $n_{head} = 16$, and key dimension $d_k = 4$. For mono-temporal considera-

tions, we use the same architecture and simply discard the unnecessary L-TAE-based aggregation. We compare our architecture against the baselines already evaluated on the SEN12MS-CR [12] and SEN12MS-CR-TS [14] datasets. We also evaluate the performance of U-TAE [22] a state-of-the-art satellite image time series encoder, using the official implementation with minor adaptations to our task [1].

Training To assess the contribution of uncertainty modelling we train two variants: *UnCRtainTS - no σ*, trained with L2 loss only, i.e., without uncertainty prediction, and *UnCRtainTS* trained with the NLL loss of Eq. 4 predicting uncertainties together with the reconstructed image. We use the ADAM optimizer [41] with an initial learning rate of 0.001, at a batch size of 4 as in [22]. All models are trained for 20 epochs with an exponential learning rate decay of 0.8, such that the rate decays by roughly one order of magnitude every 10 epochs. Models are evaluated on the validation split each epoch and the checkpoint with best validation loss is used for testing.

Evaluation For image reconstruction performance, we report the Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) as well as Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM) [73] and the Spectral Angle Mapper (SAM) metric [42]. We assess the quality of the uncertainty predictions via Uncertainty Calibration Error (UCE) [31]

$$UCE(e,u) = \sum_{p=1}^{P} \frac{N_p}{N} |e(B_p) - u(B_p)| \,, \tag{6}$$

where $e(B_p)$ denotes the RMSE of N_p pixel predictions in bin B_p, $P = 20$ is the bin count and a bin's uncertainty $u(B_p)$ is given in terms of Root Mean Variance (RMV):

$$u(B_p) = \sqrt{\frac{1}{N_p} \sum_{j \in B_p} \frac{1}{K} \sum_{k=1}^{K} u_j^k} \,. \tag{7}$$

UCE quantifies the deviation between the predicted uncertainty and the empirical reconstruction error. Low UCE corresponds to well-calibrated uncertainties. We also report a patch-wise calibration metric termed UCE_{im}, where RMSE and RMV are spatio-spectrally averaged across all pixels of a given image before calculating calibration.

4.3. UnCRtainTS

In this section we show the experimental performance of our approach, both in terms of image reconstruction and aleatoric uncertainty prediction.

[1] github.com/VSainteuf/utae-paps

Table 1. **Multi-temporal image reconstruction experiment.** We evaluate models for $T = 3$ inputs on SEN12MS-CR-TS benchmark. UnCRtainTS outperforms all learnable approaches on every metric, and performs best on all measures while predicting well calibrated uncertainties (bottom table).

Model	↓ RMSE	↑ PSNR	↑ SSIM	↓ SAM
least cloudy	0.079	—	0.815	12.204
DSen2-CR [54]	0.060	26.04	0.810	12.147
STGAN [61]	0.057	25.42	0.818	12.548
CR-TS Net [14]	*0.051*	26.68	0.836	10.657
U-TAE [22]	*0.051*	27.05	0.849	11.649
UnCRtainTS - no σ (ours)	**0.049**	*27.23*	*0.859*	*10.168*
UnCRtainTS (ours)	*0.051*	**27.84**	**0.866**	**10.160**

	UCE_{im}	UCE
UnCRtainTS (ours)	0.010	0.007

Multi-temporal image reconstruction We benchmark our method against established heuristics and baselines of [14,22,54,61]. We report the performance of these methods in Table 1. UnCRtainTS sets a new state-of-the-art performance in terms of PSNR, SSIM, and SAM. Our architecture trained without uncertainty prediction (UnCRtainTS - no σ) scores second best on all those metrics and first in RMSE. This shows that our neural architecture alone outperforms existing approaches, and uncertainty prediction further improves the reconstruction performance. Compared to U-TAE, the architecture improves by 1pt SSIM while the uncertainty prediction increases the performance by another 0.7pt. Note that uncertainty prediction has a slightly detrimental impact on RMSE performance (-0.002). This is in line with recent evidence that NLL optimization involves a trade-off between mean and variance estimate optimization that may hinder regression performance [63, 65]. However this does not impact the image similarity metrics. Lastly, in terms of parameter efficiency, our model counts 0.5M parameters. For comparison, the competitive U-TAE baseline [22] which performs third-best consists of 1.2M trainable weights, such that UnCRtainTS is relatively lightweight.

Aleatoric uncertainty prediction We show the uncertainty calibration metrics of our method at image and pixel level in Table 1. Those values should be compared to the test RMSE: at the pixel (resp. image) level the average error made on the reconstruction uncertainty is around 7 (resp. 5) times smaller than the average reconstruction error, showing satisfactory calibration. In other words, our method predicts uncertainty values that correlate well with the empirical reconstruction error. To demonstrate how uncertainty predictions can be useful in practice, we show how they allow filtering bad predictions. We rank all reconstructed images of the test set sorted by increasing UCE_{im} and accumulate squared errors from the least to the most uncertain samples. The monotonous curve in Fig. 3 displays a linear relation

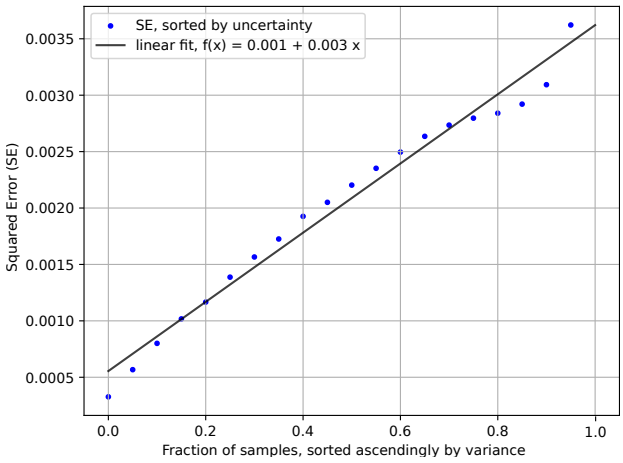


Figure 3. **Controlling error** on the test split by discarding top uncertain samples. Discarding the top 50% of uncertain reconstructions almost halves prediction error, enabling risk management.

between error and uncertainty, such that error can be stepwise decreased by uncertainty-based filtering. In practice, this enables controlling risk in downstream applications on the restored satellite images.

4.4. Architecture design

To support the previous results and our architecture design choices, we systematically investigate UnCRtainTS' hyper-parameter sensitivity. Here, all model instances are trained with L2 loss only. Because UnCRtainTS operates on feature maps at full resolution, computational complexity is an important design criterion. In addition to its image reconstruction metrics, we report each model's number of trainable parameters and Floating Point Operations Per Second (GFLOPS), estimated via FAIR's *fvcore* package [18].

Table 2. **Block setup.** Evaluation of the UnCRtainTS backbone for varying numbers of pre- and post-aggregation MBConv blocks.

MBConv		params (k)	GFLOPS	↓ RMSE	↑ PSNR	↑ SSIM	↓ SAM
n_e	n_d						
1	3	400	29.3	0.052	27.03	0.859	11.614
1	4	483	34.0	0.050	27.00	0.851	11.771
1	5	568	38.7	*0.049*	*27.23*	0.859	**10.168**
1	6	654	43.4	0.050	**27.55**	*0.860*	10.471
1	7	740	48.1	0.049	27.21	0.859	*10.300*
0	5	483	24.6	0.052	26.97	0.853	11.002
1	5	568	38.7	*0.049*	*27.23*	0.859	**10.168**
2	5	654	52.9	**0.048**	**27.55**	**0.864**	10.641

Spatial processing We explore the influence of the number of MBConv blocks before (n_e) and after (n_d) temporal aggregation in Table 2. Using $n_e = 2$ blocks in the encoder instead of one, brings a 0.5pt increase in SSIM, while the performance gain is marginal on the three other metrics. More pressingly, due to the parallel processing of the input sequence of feature maps, this setup incurs the highest computational complexity of 52.9 GFLOPS. In terms of post-aggregation blocks, performance peaks around $5-6$ modules, with 5 modules being best on one metric and a close second on two more. For these reasons we choose $n_e = 1$ pre and $n_d = 5$ post aggregation blocks as default configuration. We also note that the ($n_e = 0$) model performs competitively while being very lightweight and directly aggregating the input features. Indeed, it performs comparable to the U-TAE baseline. This secondary result shows that competitive performance can be obtained with very light architectures.

Table 3. **Head count.** Quantitative evaluation of the UnCRtainTS backbone with varying number of self-attention heads.

n_{head}	params (k)	GFLOPS	↓ RMSE	↑ PSNR	↑ SSIM	↓ SAM
1	556	38.7	**0.049**	**27.56**	0.856	10.497
4	559	38.7	0.052	*27.40*	0.856	10.825
8	563	38.7	0.051	27.00	0.851	11.131
16	568	38.7	**0.049**	27.23	*0.859*	**10.168**
32	588	38.8	0.051	27.12	**0.861**	*10.245*
64	621	38.9	0.051	27.24	0.858	11.054

Temporal aggregation Second, we explore the effect of the number of attention heads on the reconstruction quality. Table 3 shows that performances are closeby and differences in computational costs are negligible. We opt for 16 heads, in line with the literature [22].

Mono-temporal image reconstruction To validate our resolution-preserving network design, we re-train and evaluate UnCRtainTS on the mono-temporal SEN12MS-CR dataset for cloud removal. That is, we consider the special case of $T = 1$ to investigate the model's spatio-spectral restoration qualities and benchmark against the competitive baselines of [4, 17, 20, 29, 54, 56, 75]. Albeit being primarily designed for time series cloud removal, UnCRtainTS achieves best performances on all metrics except for SSIM, where it ranks second best following the recently published mono-temporal vision transformer architecture of [75]. The competitive performance achieved by the spatial encoding part of our architecture supports our choice of relying on MBConv blocks operating on full resolution feature maps.

4.5. Uncertainty Modelling

In this section, we provide additional experiments and ablations on the uncertainty prediction part of our method.

Comparison of covariance models UnCRtainTS predicts aleatoric uncertainties using a diagonal covariance

Table 4. **Mono-temporal image reconstruction experiment.** Evaluation of models for $T = 1$ inputs on the SEN12MS-CR benchmark. UnCRtainTS is best on all metrics except SSIM, where it is second following the recent vision transformer of [75].

Method	↓ MAE	↑ PSNR	↑ SSIM	↓ SAM
McGAN [17]	0.048	25.14	0.744	15.676
SAR-Opt-cGAN [29]	0.043	25.59	0.764	15.494
SAR2OPT [4]	0.042	25.87	0.793	14.788
SpA GAN [56]	0.045	24.78	0.754	18.085
Simulation-Fusion GAN [20]	0.045	24.73	0.701	16.633
DSen2-CR [54]	0.031	27.76	0.874	9.472
GLF-CR [75]	*0.028*	*28.64*	**0.885**	*8.981*
UnCRtainTS (ours)	**0.027**	**28.90**	*0.880*	**8.320**

Table 5. **Uncertainty models.** Evaluation of different uncertainty models and of two ensembles of 5 UnCRtainTS instances (bottom), with and without SAR measurements as auxiliary input data.

model	↓ RMSE	↑ PSNR	↑ SSIM	↓ SAM	↓ UCE_{im}	↓ UCE
isotropic Σ	0.053	26.74	0.842	11.77	0.029	0.023
UnCRtainTS	0.051	27.84	0.866	**10.16**	**0.010**	*0.007*
ensemble	*0.049*	**28.19**	**0.872**	*10.18*	*0.012*	**0.002**
ensemble_{noSAR}	**0.048**	*27.97*	*0.869*	10.76	0.018	0.014

model, enabling different uncertainty predictions across channels. Here, this choice is compared to the simpler option of an isotropic covariance model. In the isotropic setting, we model the covariance matrix as $\mathbf{\Sigma} = \sigma^2 \boldsymbol{I_K}$ where σ^2 is scalar and $\boldsymbol{I_K}$ the K-dimensional identity matrix. This model assumes that the aleatoric uncertainty across channels can be described with a single value. We compare the performance of those two methods in Table 5. The diagonal matrix model is best overall, outperforming on all metrics. These results clearly demonstrate that uncertainty prediction for satellite image reconstruction requires channel-specific uncertainty predictions. Indeed, modeling a diagonal covariance matrix over a simplistic isotropic description entails a three-fold reduction of the final uncertainty calibration error.

Combined epistemic and aleatoric modelling To give a full picture of uncertainty, we complement aleatoric uncertainty modelling with epistemic uncertainty estimation. We re-train the diagonal model with different weight initializations and samples of training batches to obtain a deep ensemble of $M = 5$ member networks [43]. The members' reconstructions and uncertainty predictions are averaged via:

$$\hat{\boldsymbol{y}}^M = \frac{1}{M} \sum_{m=1}^{M} \hat{\boldsymbol{y}}^m \tag{8}$$

$$(\sigma^M)^2 = \frac{1}{M} \sum_{m=1}^{M} (\sigma^m)^2 + \frac{1}{M} \sum_{m=1}^{M} (\hat{\boldsymbol{y}}^m)^2 - (\hat{\boldsymbol{y}}^M)^2 \tag{9}$$

to obtain the ensemble reconstruction $\hat{\boldsymbol{y}}^M$ and total uncertainty $(\sigma^M)^2$. As shown on Table 5, the 5-member ensemble achieves the best reconstruction performances overall. The full ensemble also achieves the best pixel-based calibration at 0.002 UCE, Deep ensembles come at a computational cost both at training and inference time, but can prove valuable for the integration in downstream applications.

Table 6. **Repeated Measures.** Evaluation of our ensemble of UnCRtainTS models with varying numbers of input time points.

input length T	↓ RMSE	↑ PSNR	↑ SSIM	↓ SAM	↓ UCE_{im}	↓ UCE
2	0.051	27.78	0.861	10.86	0.012	0.004
3	*0.049*	*28.19*	*0.872*	*10.18*	0.012	*0.002*
4	**0.047**	**28.41**	**0.875**	**9.99**	0.013	**0.001**

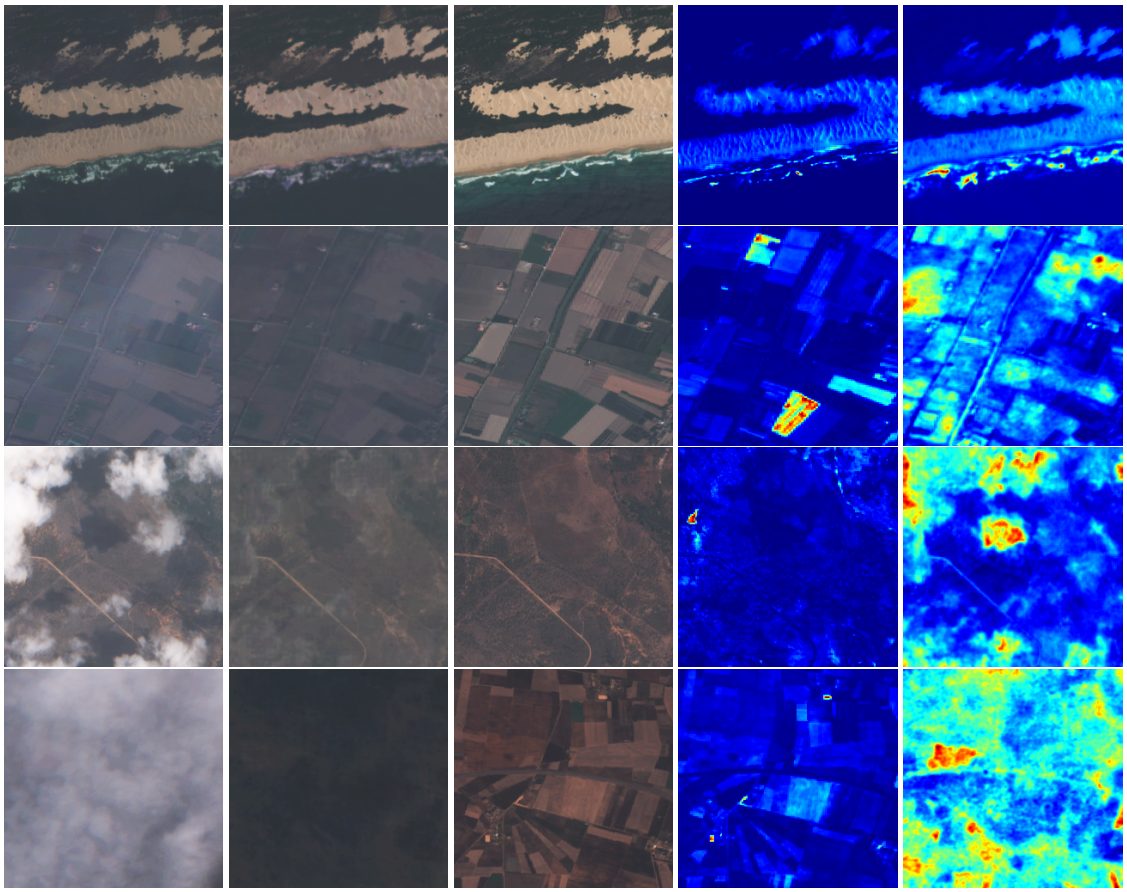
Uncertainty vs. sequence length To evaluate the effect of the number of input time points T on performances, we perform inference with the UnCRtainTS ensemble on input time series of lengths $T = 2, 3, 4$. Table 6 shows that longer sequences help achieve both better image reconstruction quality and uncertainty calibration. This confirms the intuition that longer sequences, where additional samples are likely cloud-free, facilitate the restoration task and provide growing evidence for better calibration. Table 6 also underlines that the $T = 3$ case considered in the main experiments makes for a challenging setting.

SAR reduces uncertainty We obtain a second ensemble trained without using SAR as auxiliary inputs, to explore the benefits of radar data. We show its performance on the bottom row of Table 5. The single-sensor ensemble achieves a considerably higher UCE at both image and pixel level. This suggests that the additional information contained in the SAR inputs is beneficial to improve the trustworthiness of the reconstructions.

Qualitative results Complementary to the quantitative measures, Fig. 4 shows UnCRtainTS' image restorations and uncertainty maps across varying levels of cloud coverage. Of particular interest is the uncertainty predictions not only being sensitive to clouds and cloud shadows, but also capturing other dynamics such waves breaking on a shore or the coloring of maturing crops. UnCRtainTS attends to differences in the input time series—not entirely unlike sequence-based cloud detectors explicitly designed for spotting transients across repeated measures [53]—and then, due to their temporary nature, attributes them an elevated aleatoric uncertainty.

5. Conclusion

We introduced UnCRtainTS, a novel method for combining uncertainty quantification with cloud removal from



(a) Cloudy input. (b) Prediction. (c) Target. (d) Error map. (e) Uncertainty map.

Figure 4. **Exemplary images.** Detail view on exemplary satellite images and predictions by UnCRtainTS with a diagonal covariance matrix model. Rows: Four different samples from the test split. The illustrated cases show mild atmospheric distortions, semi-transparent haze, partly dense cloud coverage and cloud coverage with no visibility at all. Columns: The input sequence's least-cloudy image ($T = 3$), UnCRtainTS' image reconstruction, the clear-view target image, the map of squared error residuals as well as the map of UnCRtainTS' variance predictions. Note the model's sensitivity to transients captured in the input time series, such as the ocean's white wash, changing crops as well as clouds and cloud shadow. UnCRtainTS captures these changing circumstances as data-inherent, aleatoric uncertainty.

optical satellite image time series. While prior contributions applied uncertainty prediction in biomedical imaging or to univariate remote sensing downstream applications, our work is the first to investigate multivariate uncertainty quantification for multispectral satellite image reconstruction. UnCRtainTS features an attention-based neural architecture that outperforms all competitors benchmarked on the satellite image reconstruction task. Our proposed method includes a formulation of aleatoric uncertainty prediction for image reconstruction based on diagonal covariance matrices, as well as an estimation of epistemic uncertainty via deep ensembles. The conducted experiments show that both of our contributions, the new architecture combined with uncertainty quantification, set a new state-of-the-art image reconstruction performance on SEN12MS-CR-TS. Finally, the outcomes highlight how our well-calibrated uncertainties can effectively serve as a measure to control reconstruction quality and help integration in risk-sensitive downstream applications. Our results encourage further explorations of more complex multivariate uncertainty models for image reconstructions. Our code is provided at https://patrickTUM.github.io/cloud_removal/.

Acknowledgements This work is jointly supported by the Federal Ministry for Economic Affairs and Energy of Germany in the project "AI4Sentinels– Deep Learning for the Enrichment of Sentinel Satellite Imagery" (FKZ50EE1910), by the German Federal Ministry of Education and Research (BMBF) in the framework "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (01DD20001) and by the German Federal Ministry of Economics and Technology in the framework of the "national center of excellence ML4Earth" (50EE2201C).

References

[1] Navid Ansari, Hans-peter Seidel, Nima Vahidi Ferdowsi, and Vahid Babaei. Autoinverse: Uncertainty aware inversion of neural networks. In *Advances in Neural Information Processing Systems*, 2022. 2

[2] Javier Antorán, Riccardo Barbano, Johannes Leuschner, José Miguel Hernández-Lobato, and Bangti Jin. Uncertainty estimation for computed tomography with a linearised deep image prior. *arXiv preprint arXiv:2203.00479*, 2022. 2

[3] Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *ACM Computing Surveys*, 2020. 4

[4] J. D. Bermudez, P. N. Happ, D. A. B. Oliveira, and R. Q. Feitosa. SAR to optical image synthesis for cloud removal with generative adversarial networks. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2018. 1, 2, 6, 7

[5] Sayantan Bhadra, Varun A Kelkar, Frank J Brooks, and Mark A Anastasio. On hallucinations in tomographic image reconstruction. *IEEE Transactions on Medical Imaging*, 2021. 2

[6] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006. 4

[7] Justin Braaten, Kurt Schwehr, and Simon Ilyushchenko. More accurate and flexible cloud masking for Sentinel-2 images. `Medium.` https://medium.com/google-earth/more-accurate-and-flexible-cloud-masking-for-sentinel-2-images-766897a9ba5f, 2020. Accessed: 2022-10-16. 4

[8] Priyanka Chaudhary, João P Leitão, Tabea Donauer, Stefano D'Aronco, Nathanaël Perraudin, Guillaume Obozinski, Fernando Perez-Cruz, Konrad Schindler, Jan Dirk Wegner, and Stefania Russo. Flood uncertainty estimation using deep ensembles. *Water*, 2022. 2

[9] Hyungjin Chung, Eun Sun Lee, and Jong Chul Ye. Mr image denoising and super-resolution using regularized reverse diffusion. *arXiv preprint arXiv:2203.12621*, 2022. 2

[10] Faramarz Naderi Darbaghshahi, Mohammad Reza Mohammadi, and Mohsen Soryani. Cloud removal in remote sensing images using generative adversarial networks and sar-to-optical image translation. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 2

[11] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. 4

[12] Patrick Ebel, Andrea Meraner, Michael Schmitt, and Xiao Xiang Zhu. Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2020. 1, 2, 4, 5

[13] Patrick Ebel, Michael Schmitt, and Xiao Xiang Zhu. Internal learning for sequence-to-sequence cloud removal via synthetic aperture radar prior information. In *International Geoscience and Remote Sensing Symposium*, 2021. 2

[14] Patrick Ebel, Yajin Xu, Michael Schmitt, and Xiao Xiang Zhu. SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 2, 3, 4, 5

[15] Robert Eckardt, Christian Berger, Christian Thiel, and Christiane Schmullius. Removal of optically thick clouds from multi-spectral satellite images using multi-frequency SAR data. *Remote Sensing*, 2013. 2, 4

[16] Vineet Edupuganti, Morteza Mardani, Shreyas Vasanawala, and John Pauly. Uncertainty quantification in deep mri reconstruction. *IEEE Transactions on Medical Imaging*, 40(1):239–250, 2020. 2

[17] Kenji Enomoto, Ken Sakurada, Weimin Wang, Hiroshi Fukui, Masashi Matsuoka, Ryosuke Nakamura, and Nobuo Kawaguchi. Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 1, 2, 6, 7

[18] FAIR. Flop Counter for PyTorch Models. `Github.` https://github.com/facebookresearch/fvcore, 2019. Accessed: 2023-01-05. 6

[19] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*. PMLR, 2016. 2

[20] Jianhao Gao, Qiangqiang Yuan, Jie Li, Hai Zhang, and Xin Su. Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks. *Remote Sensing*, 2020. 1, 2, 6, 7

[21] Vivien Sainte Fare Garnot and Loic Landrieu. Lightweight temporal self-attention for classifying satellite images time series. In *International Workshop on Advanced Analytics and Learning on Temporal Data*. Springer, 2020. 2, 3

[22] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 2, 3, 5, 6

[23] Vivien Sainte Fare Garnot, Loic Landrieu, Sebastien Giordano, and Nesrine Chehata. Satellite image time series classification with pixel-set encoders and temporal self-attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[24] Jakob Gawlikowski, Sudipan Saha, Anna Kruspe, and Xiao Xiang Zhu. An advanced dirichlet prior network for out-of-distribution detection in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 2

[25] Jakob Gawlikowski, Sudipan Saha, Julia Niebling, and Xiao Xiang Zhu. Robust distribution-shift aware sar-optical data fusion for multi-label scene classification. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 911–914. IEEE, 2022. 2

[26] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021. 2

[27] Jan Glaubitz, Anne Gelb, and Guohui Song. Generalized sparse bayesian learning and application to image reconstruction. *SIAM/ASA Journal on Uncertainty Quantification*, 2023. 2

[28] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017. 4

[29] Claas Grohnfeldt, Michael Schmitt, and Xiaoxiang Zhu. A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images. In *IEEE International Geoscience and Remote Sensing Symposium*, 2018. 1, 2, 6, 7

[30] Ziqi Gu, Patrick Ebel, Qiangqiang Yuan, Michael Schmitt, and Xiao Xiang Zhu. Explicit haze & cloud removal for global land cover classification. *CVPR 2022 Workshop on Multimodal Learning for Earth and Environment*, 2022. 1

[31] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*. PMLR, 2017. 5

[32] Ali Harakeh, Jordan Hu, Naiqing Guan, Steven L Waslander, and Liam Paull. Estimating regression predictive distributions with sample networks. *arXiv preprint arXiv:2211.13724*, 2022. 4

[33] Gensheng Hu, Xiaoyi Li, and Dong Liang. Thin cloud removal from remote sensing images using multidirectional dual tree complex wavelet transform and transfer least square support vector regression. *Journal of Applied Remote Sensing*, 2015. 2

[34] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[35] Bo Huang, Ying Li, Xiaoyu Han, Yuanzheng Cui, Wenbo Li, and Rongrong Li. Cloud removal from optical satellite imagery with SAR imagery using sparse representation. *IEEE Geoscience and Remote Sensing Letters*, 2015. 2, 4

[36] Jieon Hwang, Chushi Yu, and Yoan Shin. SAR-to-optical image translation using ssim and perceptual loss based cycle-consistent gan. In *International Conference on Information and Communication Technology Convergence*. IEEE, 2020. 2

[37] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[38] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*. Springer, 2016. 2

[39] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 2017. 2, 4

[40] Michael D. King, Steven Platnick, W. Paul Menzel, Steven A. Ackerman, and Paul A. Hubanks. Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites. *IEEE Transactions on Geoscience and Remote Sensing*, 2013. 1

[41] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 5

[42] Fred A Kruse, AB Lefkoff, JW Boardman, KB Heidebrecht, AT Shapiro, PJ Barloon, and AFH Goetz. The spectral image processing system (SIPS)-interactive visualization and analysis of imaging spectrometer data. In *AIP Conference Proceedings*. American Institute of Physics, 1993. 5

[43] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017. 2, 7

[44] Charis Lanaras, José Bioucas-Dias, Silvano Galliani, Emmanuel Baltsavias, and Konrad Schindler. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018. 2

[45] Nico Lang, Walter Jetz, Konrad Schindler, and Jan Dirk Wegner. A high-resolution canopy height model of the earth. *arXiv preprint arXiv:2204.08322*, 2022. 1, 2

[46] Nico Lang, Nikolai Kalischek, John Armston, Konrad Schindler, Ralph Dubayah, and Jan Dirk Wegner. Global canopy height regression and uncertainty estimation from GEDI LIDAR waveforms with deep ensembles. *Remote Sensing of Environment*, 2022. 1, 2

[47] Max-Heinrich Laves, Sontje Ihler, Jacob F Fast, Lüder A Kahrs, and Tobias Ortmaier. Well-calibrated regression uncertainty in medical imaging with deep learning. In *Medical Imaging with Deep Learning*. PMLR, 2020. 2

[48] Max-Heinrich Laves, Malte Tölle, and Tobias Ortmaier. Uncertainty estimation in medical image denoising with bayesian deep image prior. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*. Springer, 2020. 2

[49] Xinghua Li, Huanfeng Shen, Liangpei Zhang, Hongyan Zhang, Qiangqiang Yuan, and Gang Yang. Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning. *IEEE Transactions on Geoscience and Remote Sensing*, 2014. 2

[50] Chao-Hung Lin, Po-Hung Tsai, Kang-Hua Lai, and Jyun-Yuan Chen. Cloud removal from multitemporal satellite images using information cloning. *IEEE Transactions on Geoscience and Remote Sensing*, 2012. 2

[51] Han Liu, Peng Gong, Jie Wang, Xi Wang, Grant Ning, and Bing Xu. Production of global daily seamless data cubes and quantification of global land cover change from 1985 to 2020 - imap world 1.0. *Remote Sensing of Environment*, 2021. 1

[52] Mingliang Liu, Dario Grana, and Leandro Passos de Figueiredo. Uncertainty quantification in stochastic inversion with dimensionality reduction using variational autoencoder. *Geophysics*, 87(2):M43–M58, 2022. 2

[53] Vincent Lonjou, Camille Desjardins, Olivier Hagolle, Beatrice Petrucci, Thierry Tremas, Michel Dejus, Aliaksei Makarau, and Stefan Auer. MACCS-ATCOR joint algorithm (MAJA). In Adolfo Comerón, Evgueni I. Kassianov, and Klaus Schäfer, editors, *Remote Sensing of Clouds and the Atmosphere XXI*. International Society for Optics and Photonics, SPIE, 2016. 7

[54] Andrea Meraner, Patrick Ebel, Xiao Xiang Zhu, and Michael Schmitt. Cloud removal in Sentinel-2 imagery using a deep

residual neural network and SAR-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020. 1, 2, 4, 5, 6, 7

[55] Nando Metzger, Mehmet Ozgur Turkoglu, Stefano D'Aronco, Jan Dirk Wegner, and Konrad Schindler. Crop classification under varying cloud cover with neural ordinary differential equations. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 1

[56] Heng Pan. Cloud removal for remote sensing imagery via spatial attention generative adversarial network. *arXiv preprint arXiv:2009.13015*, 2020. 2, 6, 7

[57] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015. 2

[58] Marc Rußwurm and Marco Körner. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 1

[59] Marc Rußwurm and Marco Körner. Self-attention for raw optical satellite time series classification. *ISPRS journal of photogrammetry and remote sensing*, 2020. 1, 2

[60] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[61] Vishnu Sarukkai, Anirudh Jain, Burak Uzkent, and Stefano Ermon. Cloud removal from satellite images using spatiotemporal generator networks. In *The IEEE Winter Conference on Applications of Computer Vision*, 2020. 1, 2, 4, 5

[62] Alessandro Sebastianelli, Erika Puglisi, Maria Pia Del Rosso, Jamila Mifdal, Artur Nowakowski, Pierre Philippe Mathieu, Fiora Pirri, and Silvia Liberata Ullo. PLFM: Pixel-level merging of intermediate feature maps by disentangling and fusing spatial and temporal data for cloud removal. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 1, 2, 4

[63] Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*, 2021. 2, 4, 5

[64] Huanfeng Shen, Xinghua Li, Qing Cheng, Chao Zeng, Gang Yang, Huifang Li, and Liangpei Zhang. Missing information reconstruction of remote sensing data: A technical review. *IEEE Geoscience and Remote Sensing Magazine*, 2015. 2

[65] Nicki Skafte, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance networks. *Advances in Neural Information Processing Systems*, 2019. 2, 4, 5

[66] Sergii Skakun, Jan Wevers, Carsten Brockmann, Georgia Doxani, Matej Aleksandrov, Matej Batič, David Frantz, Ferran Gascon, Luis Gómez-Chova, Olivier Hagolle, et al. Cloud mask intercomparison exercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. *Remote Sensing of Environment*, 2022. 4

[67] Andrew Stirn, Hans-Hermann Wessels, Megan Schertzer, Laura Pereira, Neville E Sanjana, and David A Knowles. Faithful heteroscedastic regression with neural networks. *arXiv preprint arXiv:2212.09184*, 2022. 2, 4

[68] Hiroshi Takahashi, Tomoharu Iwata, Yuki Yamanaka, Masanori Yamada, and Satoshi Yagi. Student-t variational autoencoder for robust density estimation. In *International Joint Conference on Artificial Intelligence*, 2018. 2

[69] Malte Tölle, Max-Heinrich Laves, and Alexander Schlaefer. A mean-field variational inference approach to deep image prior for inverse problems in medical imaging. In *Medical Imaging with Deep Learning*. PMLR, 2021. 2

[70] Mehmet Ozgur Turkoglu, Alexander Becker, Hüseyin Anil Gündüz, Mina Rezaei, Bernd Bischl, Rodrigo Caye Daudt, Stefano D'Aronco, Jan Dirk Wegner, and Konrad Schindler. FiLM-ensemble: Probabilistic deep learning via feature-wise linear modulation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2

[71] Mehmet Ozgur Turkoglu, Stefano D'Aronco, Gregor Perich, Frank Liebisch, Constantin Streit, Konrad Schindler, and Jan Dirk Wegner. Crop mapping from image time series: Deep learning with multi-scale label hierarchies. *Remote Sensing of Environment*, 2021. 1

[72] Xiaoke Wang, Guangluan Xu, Yang Wang, Daoyu Lin, Peiguang Li, and Xiujing Lin. Thin and thick cloud removal on remote sensing image by conditional generative adversarial network. In *International Geoscience and Remote Sensing Symposium*, 2019. 2

[73] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. 2, 5

[74] Zhaobin Wang, Yikun Ma, and Yaonan Zhang. Hybrid cgan: Coupling global and local features for sar-to-optical image translation. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 2

[75] Fang Xu, Yilei Shi, Patrick Ebel, Lei Yu, Gui-Song Xia, Wen Yang, and Xiao Xiang Zhu. GLF-CR: SAR-enhanced cloud removal with global–local fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022. 2, 4, 6, 7

[76] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2

[77] Anze Zupanc. Improving cloud detection with machine learning. `Sentinel-Hub`. https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13, 2017. Accessed: 2019-10-10. 4

B Appendix: Related Publications

The following published and peer-reviewed works, while not strictly being part of the main contributions constituting this dissertation, are nonetheless more or less closely related. Subsequently, these publications are listed below for any interested reader.

- **P. Ebel**, M. Schmitt, and X. X. Zhu. Cloud removal in unpaired Sentinel-2 imagery using cycle-consistent GAN and SAR-optical data fusion. In IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium, pages 2065–2068. IEEE, 2020.
- **P. Ebel**, M. Schmitt, and X. X. Zhu. Internal learning for sequence-to-sequence cloud removal via synthetic aperture radar prior information. In 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, pages 2691–2694. IEEE, 2021.
- **P. Ebel**, M. Schmitt, and X. X. Zhu. Multi-sensor time series cloud removal fusing optical and SAR satellite information. In IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, pages 5381-5384. IEEE, 2022.
- Z. Gu, **P. Ebel**, M. Schmitt, and X. X. Zhu. Explicit haze and cloud removal for global land cover classification. CVPR 2022 Workshop on Multimodal Learning for Earth and Environment, pages 1–6, 2022.

- **P. Ebel**, S. Saha, and X. X. Zhu. Fusing multi-modal data for supervised change detection. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 43:243–249, 2021.
- S. Saha, **P. Ebel**, and X. X. Zhu. Self-supervised multisensor change detection. IEEE Transactions on Geoscience and Remote Sensing, 60:1–10, 2022.
- S. Saha, M. Shahzad, **P. Ebel**, and X. X. Zhu. Supervised change detection using prechange optical-SAR and postchange SAR data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15:8170–8178, 2022.