ISPRS Journal of Photogrammetry and Remote Sensing 177 (2021) 89–102

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Aerial scene understanding in the wild: Multi-scene recognition via prototype-based memory networks

Yuansheng Hua [a,b], Lichao Mou [a,b], Jianzhe Lin [c], Konrad Heidler [a,b], Xiao Xiang Zhu [a,b,*]

[a] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, 82234 Wessling, Germany
[b] Data Science in Earth Observation (SiPEO), Technical University of Munich (TUM), Arcisstr. 21, 80333 Munich, Germany
[c] Electrical and Computer Engineering (ECE), University of British Columbia (UBC), V6T 1Z2, Canada

ARTICLE INFO

Keywords:
Convolutional neural network (CNN)
Multi-scene recognition in single images
Memory network
Multi-scene aerial image dataset
Multi-head attention-based memory retrieval
Prototype learning

ABSTRACT

Aerial scene recognition is a fundamental visual task and has attracted an increasing research interest in the last few years. Most of current researches mainly deploy efforts to categorize an aerial image into one scene-level label, while in real-world scenarios, there often exist multiple scenes in a single image. Therefore, in this paper, we propose to take a step forward to a more practical and challenging task, namely multi-scene recognition in single images. Moreover, we note that manually yielding annotations for such a task is extraordinarily time- and labor-consuming. To address this, we propose a prototype-based memory network to recognize multiple scenes in a single image by leveraging massive well-annotated single-scene images. The proposed network consists of three key components: 1) a prototype learning module, 2) a prototype-inhabiting external memory, and 3) a multi-head attention-based memory retrieval module. To be more specific, we first learn the prototype representation of each aerial scene from single-scene aerial image datasets and store it in an external memory. Afterwards, a multi-head attention-based memory retrieval module is devised to retrieve scene prototypes relevant to query multi-scene images for final predictions. Notably, only a limited number of annotated multi-scene images are needed in the training phase. To facilitate the progress of aerial scene recognition, we produce a new multi-scene aerial image (MAI) dataset. Experimental results on variant dataset configurations demonstrate the effectiveness of our network. Our dataset and codes are publicly available[1].

1. Introduction

With the enormous advancement of remote sensing technologies, massive high-resolution aerial images are now available and beneficial to a large variety of applications, e.g., urban planning (Marmanis et al., 2018; Audebert et al., 2018; Marcos et al., 2018; Mou and Zhu, 2018b; Li et al., 2018; Qiu et al., 2020; Li et al., 2020b), traffic monitoring (Mou and Zhu, 2018c; Mou and Zhu, 2016), disaster assessment (Vetrivel et al., 2018; Lee et al., 2017), and natural resource management (Lucchesi et al., 2013; Weng et al., 2018; Cheng et al., 2017; Zarco-Tejada et al., 2014; Wen et al., 2017; Mou and Zhu, 2018a; Qiu et al., 2019). Driven by these applications, aerial scene recognition that refers to assigning aerial images scene-level labels is now becoming a fundamental but challenging task.

In recent years, many efforts (Zhu et al., 2017), e.g., developing novel network architectures (Murray et al., 2019; Cheng et al., 2020; Bi et al., 2020; Niazmardi et al., 2017; Lin et al., 2020; Zhu et al., 2018) and pipelines (Byju et al., 2000; Xu et al., 2020; Wang et al., 2019; Zhu et al., 2019), publishing large-scale datasets (Xia et al., 2017; Jin et al., 2018), introducing multi-modal and multi-temporal data (Hu et al., 2020; Tuia et al., 2016; Ru et al., 2020; Li et al., 2020a), have been deployed to address this task, and most of them treat it as a single-label classification problem. A common assumption shared by these researches is that an aerial image belongs to only one scene category, while in real-world scenarios, it is more often that there exist various scenes in a single image (cf. Fig. 1). Furthermore, we notice that aerial images used to learn single-label scene classification models are usually well-cropped so that target scenes could be centered and account for the majority of an

* Corresponding author at: Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, 82234 Wessling, Germany. Data Science in Earth Observation (SiPEO), Technical University of Munich (TUM), Arcisstr. 21, 80333 Munich, Germany.
E-mail addresses: yuansheng.hua@dlr.de (Y. Hua), lichao.mou@dlr.de (L. Mou), jianzhelin@ece.ubc.ca (J. Lin), konrad.heidler@dlr.de (K. Heidler), xiaoxiang.zhu@dlr.de (X.X. Zhu).

[1] https://github.com/Hua-YS/Prototype-based-Memory-Network.

https://doi.org/10.1016/j.isprsjprs.2021.04.006
Received 11 November 2020; Received in revised form 7 April 2021; Accepted 9 April 2021
Available online 16 May 2021

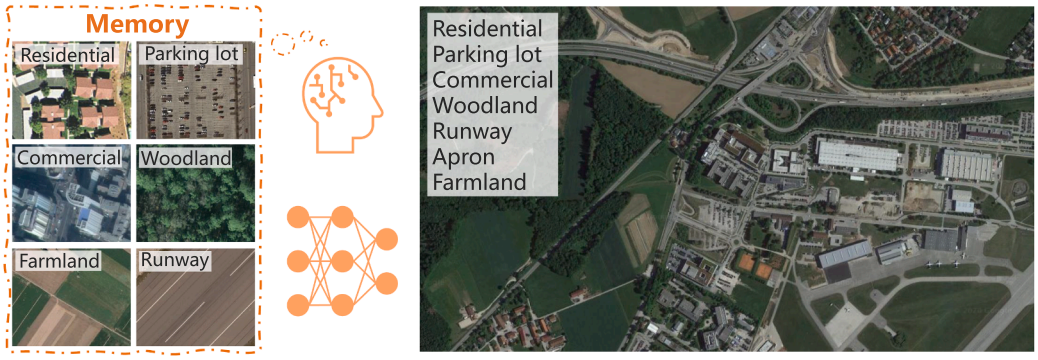


Fig. 1. Illustration of how humans learn to perceive unconstrained aerial images being composed of multiple scenes. We first learn and memorize individual aerial scenes. Then we can possess the capability of understanding complex scenarios by learning from only a limited number of hard instances. We believe by simulating this learning process, a deep neural network can also learn to interpret multi-scene aerial images.

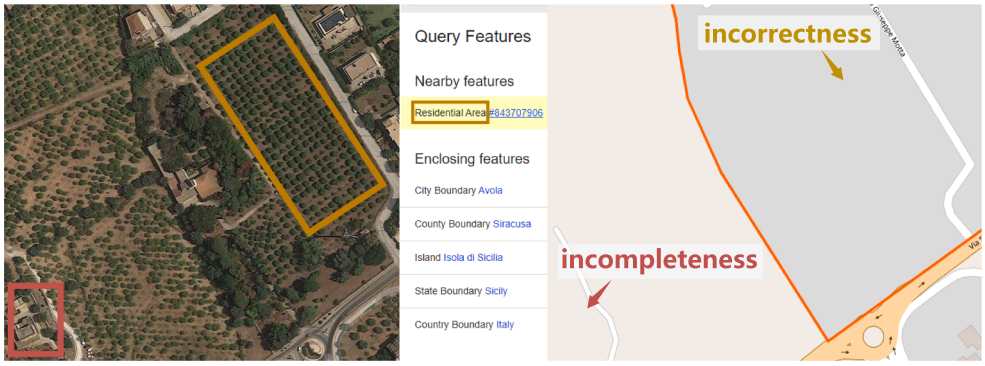


Fig. 2. Examples of incomplete (red) and incorrect (yellow) OSM data. Red: the commercial is not annotated in OSM data. Yellow: the orchard is mislabeled as `residential`.

aerial image. Unfortunately, this might be infeasible for practical applications. Therefore, in this paper, we aim to deal with a more practical and challenging problem, multi-scene classification in a single image, which refers to inferring multiple scene-level labels for a large-scale, unconstrained aerial image. Fig. 1 shows an example image, where we can see that multiple scenes, e.g., `residential`, `parking lot`, and `commercial`, co-exist in one aerial image. We note that there is another research branch of aerial image understanding, multi-label object classification, which refers to the process of inferring multiple objects present in an aerial image. These studies (Sumbul and Demir, 2019; Zegeye and Demir, 2018; Hua et al., 2020; Khan et al., 2019; Hua et al., 2019; Zeggada et al., 2017; Koda et al., 2018) mainly focus on recognizing object-level labels, while in our task, an image is classified into multiple scene categories, which provides a more comprehensive understanding of large-scale aerial images in scene-level. To the best of our knowledge, multi-scene recognition in unconstrained aerial images still remains underexplored in the remote sensing community.

To achieve this task, huge quantities of well-annotated multi-scene images are needed for the purpose of training models. However, we note that such annotations are not easy in the remote sensing community. This could be attributed to the following two reasons. On the one hand, the visual interpretation of multiple scenes is more arduous than that of a single scene in an aerial image, and therefore, labeling multi-scene images requires more work. On the other hand, low-cost annotation techniques, e.g., resorting to crowdsourcing OpenStreetMap (OSM) through keyword searching (Xia et al., 2017; Jin et al., 2018; Long et al., 2020), perform poorly in yielding multi-scene datasets owing to the incompleteness and incorrectness of certain OSM data. Examples of erroneous OSM data are shown in Fig. 2. In addition, manually rectifying annotations generated from crowdsourcing data are inevitable due to error-proneness. Such a procedure is quite labor-consuming, as every scene is required to be checked in case that present ones are mislabeled as absent. Aiming to solve the aforementioned limitations, in this work, we propose to train a network for recognizing complex multi-scene aerial images by using only a small number of labeled multi-scene images but a huge amount of existing, annotated single-scene data. Our motivation is based on an intuitive observation about how humans learn to perceive complex scenes being composed of multiple entities (National Research Council, 2000; Liu et al., 2008; McLaren and Suret, 2000): we first learn and memorize individual objects (through flash cards for example) when we were babies and then possess the capability of understanding complex scenarios by learning from only a limited number of hard instances (cf. Fig. 1). We believe that this learning process also applies to the interpretation of multi-scene aerial images. Driven by this observation, we propose a novel network, termed as prototype-based memory network (PM-Net), which is inspired by recent successes of memory networks in natural language processing (NLP) tasks (Sukhbaatar et al., 2015; Miller et al., 2016) and video analysis (Shi et al., 2019; Park et al., 2020; Lai et al., 2020). To be more specific, we first learn the prototype representation of each aerial scene from single-scene aerial images and then store these prototypes in the external memory of PM-Net. Afterwards, for a given query multi-scene image, a multi-head attention-based memory retrieval module is devised to retrieve scene prototypes that are associated with the query image from the external memory for inferring multiple scene labels (see Fig. 3).

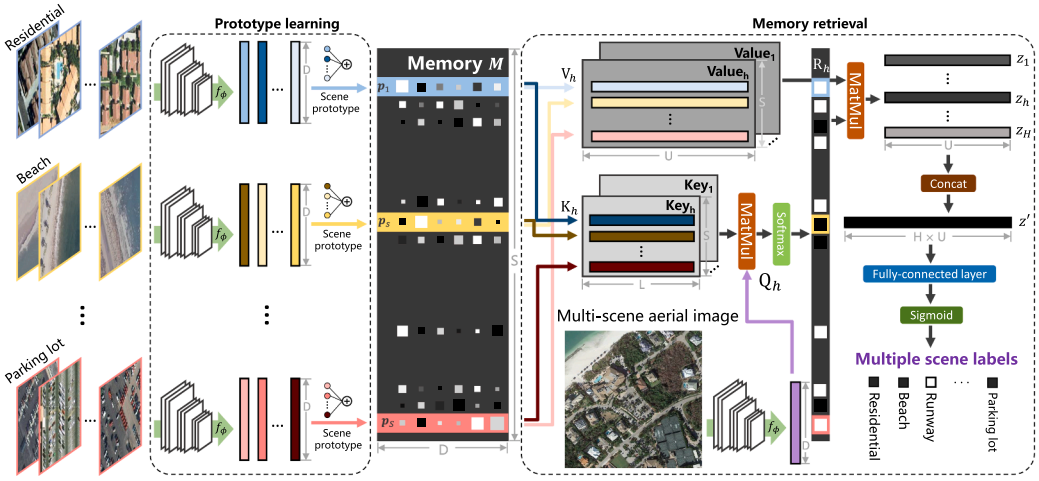


Fig. 3. Architecture of the proposed PM-Net. Particularly, we first learn scene prototypes $\boldsymbol{p}_s$ from well-annotated single-scene aerial images and then store them in the external memory $\boldsymbol{M}$ of PM-Net. Afterwards, given a query multi-scene image, a multi-head attention-based memory retrieval module is devised to retrieve scene prototypes that are relevant to the query image, yielding z' for the prediction of multiple labels. f_ϕ denotes the embedding function, and its output is a D-dimensional feature vector. S and H represent numbers of scenes and heads, respectively. L and U denote channel dimensions of the key and value in the memory retrieval module.

The contributions of this work are fourfold.

- We take a step forward to a more practical and challenging task in aerial scene understanding, namely multi-scene classification in single images, which aims to recognize multiple scenes present in a large-scale, unconstrained aerial image. Such a task is in line with real-world scenarios and capable of providing a comprehensive picture for a given geographic area.
- Given that labeling multi-scene images is very labor-intensive and time-consuming, we propose a PM-Net that can be trained for our task by leveraging large numbers of existing single-scene aerial images and a small number of labeled multi-scene images.
- In order to facilitate the progress of multi-scene recognition in single aerial images, we create a new dataset, multi-scene aerial image (MAI) dataset. To the best of our knowledge, this is the first publicly available dataset for aerial multi-scene interpretation. Compared to existing single-scene aerial image datasets, images in our dataset are unconstrained and contain multiple scenes, which are more in line with the reality.
- We carry out extensive experiments with different configurations. Experimental results demonstrate the effectiveness of the proposed network.

The remaining sections of this paper are organized as follows. Section 2 reviews studies in memory networks and prototypical networks, and the architecture of the proposed prototype-based memory network is introduced in Section 3. Section 4 describes experimental configurations and analyzes results. Eventually, conclusions are drawn in Section 5.

2. Related work

Since very few efforts have been deployed to this task in the remote sensing community, we only review literatures related to our algorithm in this section.

2.1. Memory networks

A memory network takes as input a query and retrieves complementary information from the external memory. In Sukhbaatar et al. (2015), the memory network is first proposed and utilized to address question–answering tasks, where questions are regarded as queries, and statements are stored in the external memory. To retrieve statements for predicting answers, the authors compute relative distances between queries and the external memory through dot product. In the following work, Miller et al. (2016) improves the efficiency of retrieving large memories by pre-selecting small subsets with key hashing. Moreover, the memory network is further applied in video analysis (Shi et al., 2019; Park et al., 2020; Lai et al., 2020) and image captioning (Cornia et al., 2020). In Shi et al. (2019), the authors devise a dual augmented memory network to memorize both target and background features of an video, and use a Long Short-Term Memory (LSTM) to communicate with previous and next frames. In Park et al. (2020), the authors propose a memory network to memorize normal patterns for detecting anomalies in an video. As an attempt in image captioning, Cornia et al. (2020) devise a learnable memory to learn and memorize priori knowledge for encoding relationships between image regions. Inspired by these works, we devise a memory network and store scene prototypes in the memory for recognizing scenes present in multi-scene images.

2.2. Prototypical networks

Prototypical networks are characterized by classifying images according to their distances from class prototypes. In learning with limited training samples, such networks are popular and achieved many successes recently (Snell et al., 2017; Guerriero et al., 2018; Yang et al., 2018; Huang et al., 2020; Zhang et al., 2020; Tang et al., 2019). To be specific, Snell et al. (2017) propose to first learn a prototype representation for each category and then identify images by finding their nearest category prototypes. Guerriero et al. (2018) aim to alleviate the heavy expense of learning prototypes by initializing and updating prototypes with those learned in previous training epochs. Yang et al. (2018) propose to combine prototypical networks and CNNs for tackling the open world recognition problem and improving the robustness and accuracy of networks. Similarly, Huang et al. (2020) propose to integrate prototypical networks and graph convolutional neural networks for learning relational prototypes. Albeit variant, most existing works share a common way to extract prototypes, which is taking average of samples belonging to the same categories. Therefore, we follow this prototype extraction strategy in our work.

3. Methodology

3.1. Overview

The proposed PM-Net consists of three essential components: a prototype learning module, an external memory, and a memory retrieval module. Specifically, the prototype learning module is devised to encode prototype representations of aerial scenes, which are then stored in the external memory. The memory retrieval module is responsible for retrieving scene prototypes related to query images through a multi-head attention mechanism. Eventually, retrieved scene prototypes are

utilized to infer the existence of multiple scenes in the query image.

3.2. Scene prototype learning and writing

Following the observation introduced in Section 1, we propose to learn and memorize scene prototypes with the support of single-scene aerial images. The procedure consists of two stages. We first employ an embedding function to learn semantic representations of all single-scene images. Then, feature representations belonging to the same scene category are encoded into a scene prototype and stored in the external memory.

Formally, let $\boldsymbol{X}_i^s$ denote the i-th single-scene image belonging to scene s, and i ranges from 1 to N_s. N_s is the number of samples annotated as s. The embedding function f_ϕ can be learned via the following objective function:

$$\mathscr{L}(\boldsymbol{X}_i^s, \boldsymbol{y}^s) = -\boldsymbol{y}^s \log \frac{\exp(-\boldsymbol{g}_\theta(\boldsymbol{f}_\phi(\boldsymbol{X}_i^s)))}{\sum_s \sum_i \exp(-\boldsymbol{g}_\theta(\boldsymbol{f}_\phi(\boldsymbol{X}_i^s)))}, \tag{1}$$

where ϕ represents learnable parameters of f_ϕ, and $\boldsymbol{y}^s$ is a one-hot vector denoting the scene label of $\boldsymbol{X}_i^s$. g_θ is a multilayer perceptron (MLP) with parameters θ and its outputs are activated by a softmax function to predict probability distributions. Following the overwhelming trend of deep learning, here we employ a deep CNN, e.g., ResNet-50 (He et al., 2016), as the embedding function f_ϕ and learn its parameters on public single-scene aerial image datasets. After sufficient training, f_ϕ is expected to be capable of learning discriminative representations for different aerial scenes.

Once f_ϕ is learned, the scene prototype can be computed by averaging representations of all aerial images belonging to the same scene (Snell et al., 2017; Guerriero et al., 2018; Yang et al., 2018). Let $\boldsymbol{p}_s$ be the prototype representation of scene s. We calculate $\boldsymbol{p}_s$ with the following equation:

$$\boldsymbol{p}_s = \frac{\boldsymbol{1}}{N_s} \sum_{i=1}^{N_s} \boldsymbol{f}_\phi(\boldsymbol{X}_i^s). \tag{2}$$

By doing so, in the single-scene classification, an image closely around $\boldsymbol{p}_s$ in the common embedding space is supposed to belong to scene s. Similarly, in the multi-scene scenario, the representation of an aerial image comprising scene s should show high relevance with $\boldsymbol{p}_s$. After encoding all scene prototypes, the external memory $\boldsymbol{M}$ can be formulated as follows:

$$\boldsymbol{M} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_S]^T, \tag{3}$$

where S denotes the number of scenes. $[\cdots, \cdots]$ represents the concatenation operation. Given that $\boldsymbol{p}_s$ is a D-dimensional vector, $\boldsymbol{M}$ is a matrix of $S \times D$. Note that D varies when using different backbone CNNs as embedding functions.

3.3. Multi-head attention-based memory retrieval

Inspired by successes of the multi-head self-attention mechanism (Vaswani et al., 2017) in natural language processing tasks (Radford et al., 2018; Radford et al., 2019; Devlin et al., 2018; Wolf et al., 2020), we develop a multi-head attention-based memory retrieval module to retrieve scene prototypes from the memory $\boldsymbol{M}$ for a given query image $\boldsymbol{X}$. Given a query multi-scene aerial image $\boldsymbol{X}$, to retrieve relevant scene prototypes from $\boldsymbol{M}$, we develop a multi-head attention-based memory retrieval module. In particular, we first extract the feature representation of $\boldsymbol{X}$ through the same embedding function f_ϕ and linearly project it to an L-dimensional query $\mathrm{Q}(\boldsymbol{X})$. Similarly, we transform the external memory $\boldsymbol{M}$ into key $\mathrm{K}(\boldsymbol{M})$ and value $\mathrm{V}(\boldsymbol{M})$, and both are implemented as MLPs. The channel dimension of the key is L, while that of the value is U. The relevance between $\boldsymbol{X}$ and each scene prototype $\boldsymbol{p}_s$ can be measured by dot product similarity and a softmax function as follows:

$$\mathrm{R}(\boldsymbol{X}, \boldsymbol{M}) = \mathrm{softmax}\left(\frac{\mathrm{Q}(\boldsymbol{f}_\phi(\boldsymbol{X})) \cdot \mathrm{K}(\boldsymbol{M})^T}{\sqrt{\boldsymbol{L}}}\right). \tag{4}$$

The output is an S-dimensional vector, where each component represents a relevance probability that a specific scene prototype is related to the query image. Subsequently, the retrieved scene prototypes are computed by weight-summing all values with the following equation:

$$z = \mathrm{R}(\boldsymbol{X}, \boldsymbol{M}) \cdot \mathrm{V}(\boldsymbol{M}). \tag{5}$$

Since the memory retrieval is designed in a multi-head fashion, the final retrieved prototype is reformulated as follows:

$$z' = [z_1, z_2, \ldots, z_H], \tag{6}$$

where H denotes the number of heads, and each head yields a retrieved prototype z_h by transforming $\boldsymbol{X}$ and $\boldsymbol{M}$ to the variant query $\mathrm{Q}_h(f_\phi(\boldsymbol{X}))$, key $\mathrm{K}_h(\boldsymbol{M})$, and value $\mathrm{V}_h(\boldsymbol{M})$. Eventually, the output z' is fed into a fully-connected layer followed by a sigmoid function for inferring presences of aerial scenes.

3.4. Implementation details

For a comprehensive assessment of our PM-Net, we implement the embedding function with various backbone CNNs. Specifically, we conduct experiments on four CNN architectures, and details are as follows:

- PM-VGGNet: f_ϕ is built on VGG-16 (Simonyan and Zisserman, 2014) by replacing all layers after the last max-pooling layer in *block5* with a global average pooling layer.
- PM-Inception-V3: Inception-V3 (Szegedy et al., 2015) is utilized, and layers before and including the global average pooling layer are employed as f_ϕ.
- PM-ResNet: We modify ResNet-50 (He et al., 2016) by discarding layers after the global average pooling layer and using the remaining layers as f_ϕ.
- PM-NASNet: The backbone of f_ϕ is mobile NASNet (Zoph and Le, 2017). As with the modification in PM-ResNet, only layers before and including the global average pooling layer are used.

In our experiments, we train original deep CNNs on single-scene aerial image datasets and then take them as the embedding function f_ϕ following the aforementioned points. Subsequently, we yield scene prototypes $\boldsymbol{p}_s$ and concatenate all of them along the first axis to form $\boldsymbol{M}$.

4. Experiments and discussion

In this section, we introduce a newly produced multi-scene aerial image dataset, MAI dataset, and two single-scene datasets, i.e., UCM and AID datasets, which are used in experiments. Then network configurations and training schemes are detailed in SubSection 4.2. The remaining subsections discuss and analyze the performance of the proposed network thoroughly.

4.1. Dataset description and configuration

4.1.1. MAI dataset

To facilitate the progress of aerial scene interpretation in the wild, we yield a new dataset, MAI dataset, by collecting and labeling 3923 large-scale images from Google Earth imagery that covers the United

Fig. 4. Example images in our MAI dataset. Each image is 512 × 512 pixels, and their spatial resolutions range from 0.3 m/pixel to 0.6 m/pixel. We list their scene-level labels here: (a) farmland and residential; (b) baseball, woodland, parking lot, and tennis court; (c) commercial, parking lot, and residential; (d) woodland, residential, river, and runway; (e) river and storage tanks; (f) beach, woodland, residential, and sea; (g) farmland, woodland, and residential; (h) apron and runway; (i) baseball field, parking lot, residential, bridge, and soccer field.

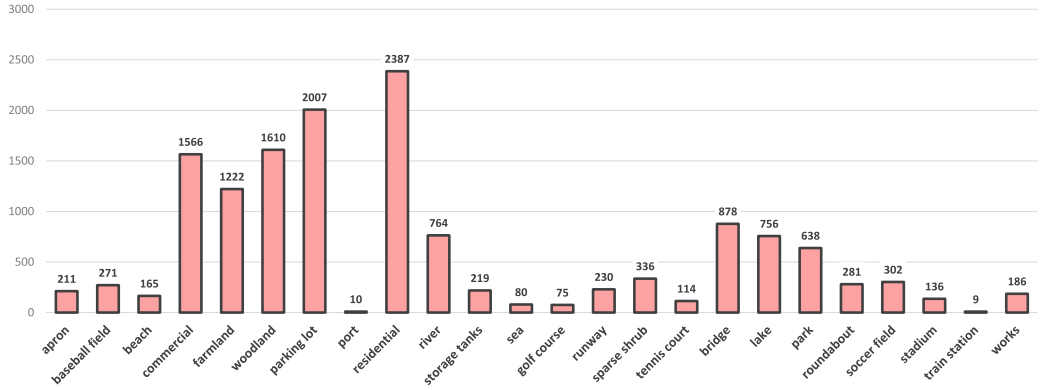


Fig. 5. Statistics of the proposed MAI dataset for multi-scene classification in single aerial images.

States, Germany, and France. The size of each image is 512 × 512, and spatial resolutions vary from 0.3 m/pixel to 0.6 m/pixel. After capturing aerial images, we manually assign each image multiple scene-level labels from in total 24 scene categories, including apron, baseball, beach, commercial, farmland, woodland, parking lot, port, residential, river, storage tanks, sea, bridge, lake, park, roundabout, soccer field, stadium, train station, works, golf course, runway, sparse shrub, and tennis court. Notably, OSM data associated with the collected images cannot be directly employed as reference owing to the problems presented in Section 1. Such a labeling procedure is extremely time- and labor-consuming, and annotating one image costs around 20 s, which is ten times more than labeling a single-scene image. Several example multi-

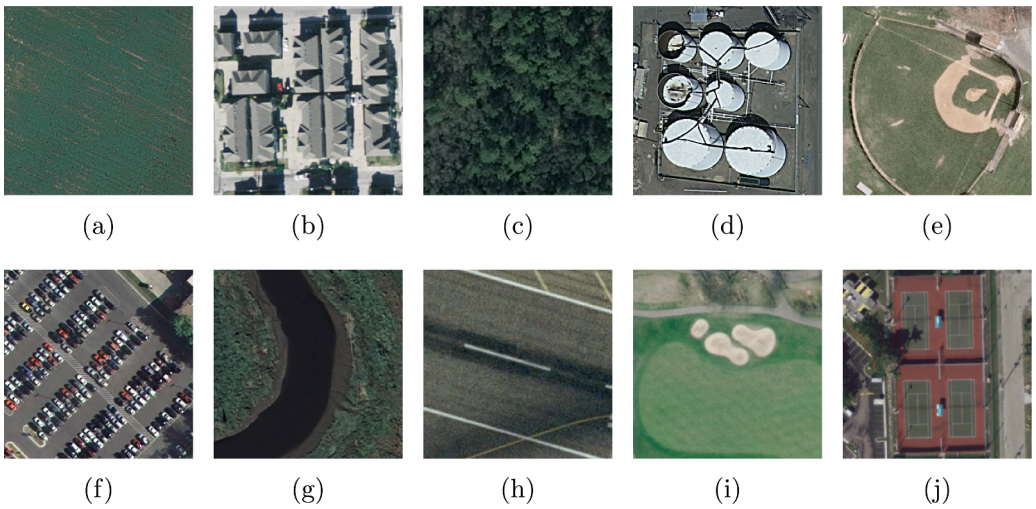


Fig. 6. Example single-scene aerial categories in the UCM dataset: (a) agricultural, (b) dense residential, (c) forest, (d) storage tanks, (e) baseball field, (f) parking lot, (g) river, (h) runway, (i) golf course, and (j) tennis court.

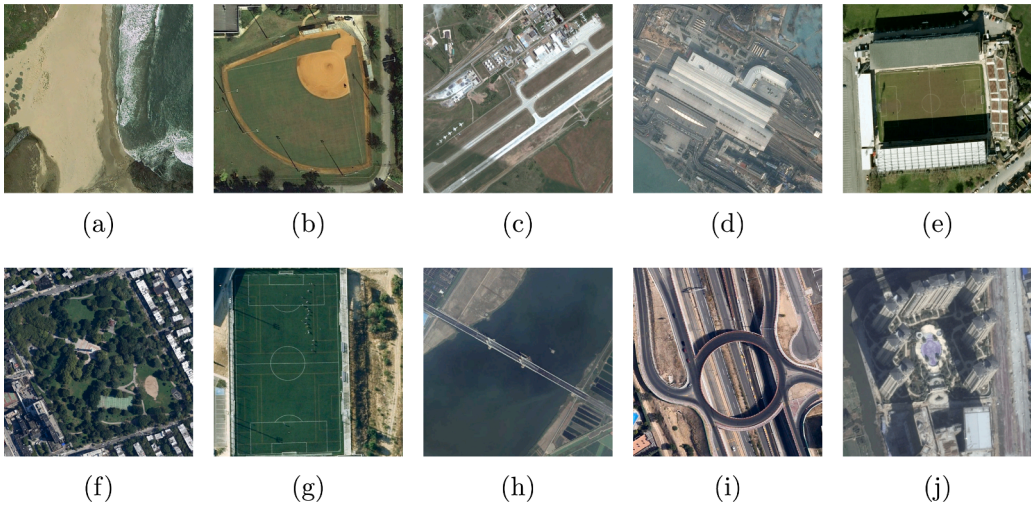


Fig. 7. Example single-scene aerial categories in the AID dataset: (a) beach, (b) baseball field, (c) airport, (d) railway station, (e) stadium, (f) park, (g) playground, (h) bridge, (i) viaduct, and (j) commercial.

scene images are shown in Fig. 4. Numbers of aerial images related to various scenes are reported in Fig. 5. Among existing datasets, BigEarthNet (Sumbul et al., 2019) is one of the most relevant datasets, which consists of Sentinel-2 images acquired over the European Union with spatial resolutions ranging from 10 m/pixel to 60 m/pixel. Spatial sizes of images vary from 20 × 20 pixels to 120 × 120 pixels, and each is assigned multiple land-cover labels provided from the CORINE Land Cover map[2]. Compared to BigEarthNet, our dataset is characterized by its high-resolution large-scale aerial images and worldwide coverage.

4.1.2. UCM dataset

UCM dataset (Yang and Newsam, 2010) is a commonly used single-scene aerial image dataset produced by Yang and Newsam from the University of California Merced. This dataset comprises 2100 aerial images cropped from aerial ortho imagery provided by the United States Geological Survey (USGS) National Map, and the spatial resolution of the collected images is one foot. The size of each image is 256 × 256 pixels, and all image samples are classified into 21 scene-level classes: overpass, forest, beach, baseball diamond, building, airplane, freeway, intersection, harbor, golf course, runway, agricultural, storage tank, mobile home park, medium residential, sparse residential, chaparral, river, tennis courts, dense residential, and parking lot. The number of aerial images collected for each scene is 100, and several example images are shown in Fig. 6. To learn scene prototypes from these single-scene images, we randomly choose 80% of image samples per scene category to train and validate the embedding function and utilize the rest for testing.

4.1.3. AID dataset

AID dataset (Xia et al., 2017) is a another popular single-scene aerial image dataset which consists of 10000 aerial images with a size of 600 × 600 pixels. These images are captured from Google Earth imagery that is taken over China, the United States, England, France, Italy, Japan, and Germany, and spatial resolutions of the collected images vary from 0.5 m/pixel to 8 m/pixel. In total, there are 30 scene categories, including viaduct, river, baseball field, center, farmland, railway station, meadow,

[2] https://land.copernicus.eu/pan-european/corine-land-cover.

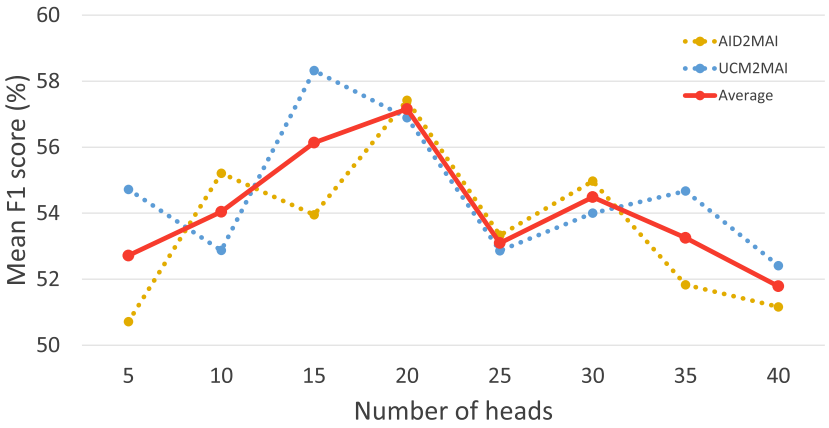


Fig. 8. The influence of the number of heads on both dataset configurations. Blue and yellow dot lines represent mean F_1 scores on UCM2MAI and AID2MAI. The Red line indicates the average of them.

Table 1
The number of images associated with each scene.

	UCM2MAI		AID2MAI	
Scene Category	UCM	MAI	AID	MAI
apron	100	194	360	54
baseball field	100	75	220	235
beach	100	94	400	130
commercial	100	607	350	1391
farmland	100	680	370	983
woodland	100	762	250	1312
parking lot	100	708	390	1777
port	100	3	380	9
residential	200	958	700	2082
river	100	209	410	686
storage tanks	100	89	360	193
sea	100*	51	400*	59
golf course	100	75	-	-
runway	100	230	-	-
sparse shrub	100	336	-	-
tennis court	100	114	-	-
bridge	-	-	360	878
lake	-	-	420	756
park	-	-	350	638
roundabout	-	-	420	281
soccer field	-	-	370	302
stadium	-	-	290	136
train station	-	-	260	9
works	-	-	390	186
All	1600	1649	7050	3239

* indicates that the number of images is not counted in total amounts, as the scene prototype of `beach` and `sea` are learned from the same images.

bare land, storage tanks, beach, mountain, park, bridge, playground, church, commercial, desert, forest, parking, industrial, square, sparse residential, pond, medium residential, port, resort, airport, school, stadium, and dense residential. The number of images in different classes ranges from 220 to 420. Similar to the data split in the UCM dataset, 20% of images are chosen from each scene as test samples, while the remaining images are utilized to train and validate the embedding function. Some example images of the AID dataset are exhibited in Fig. 7.

4.1.4. Dataset configuration

In order to widely evaluate the performance of our method, we utilize two variant dataset configurations, UCM2MAI and AID2MAI, based on common scene categories shared by UCM/AID and MAI. Specifically, the UCM2MAI configuration consists of 1600 single-scene aerial images from the UCM dataset and 1649 multi-scene images from our MAI dataset. 16 aerial scenes that are commonly included in both two datasets are considered in UCM2MAI, and numbers of their associated images are listed in Table 1. Besides, the AID2MAI configuration is composed of 7050 and 3239 aerial images from the AID and MAI datasets, respectively. 20 common scene categories are taken into consideration, and the number of images related to each scene is present in Table 1. Although such configurations might limit the number of recognizable scene classes, we believe this limitation can be addressed by collecting more single-scene images by crawling OSM data and producing large-scale multi-scene aerial image datasets. We select only 90 and 120 multi-scene aerial images from UCM2MAI and AID2MAI as training instances, respectively, and test networks on the remaining multi-scene images. For rare scenes (e.g., port and train station), we select all associated training images, while for common scenes, we randomly select several of their training samples. It is noteworthy that we yield the scene prototype of `residential` by taking an average of high-level representations of aerial images belonging to scene `medium residential` and `dense residential`. Besides, although the UCM and AID datasets do not contain images for `sea`, their images for `beach` often comprise both sea and beach (cf. (c) in Fig. 7). Therefore, we make use of training samples labeled as `beach` to yield the prototype representation of `sea`.

Table 2
Differences between two training phases.

Phase	Learnable Module	Dataset		Memory
		Pretraining f_ϕ	Fine-tuning module	
1	prototype learning	ImageNet	UCM/AID	updated
2	memory retrieval	UCM/AID	MAI	frozen

4.2. Training details

The training procedure consists of two phases: 1) learning the embedding function f_ϕ on large quantities of single-scene aerial images and 2) training the entire PM-Net on a limited number of multi-scene images in an end-to-end manner. Thus, various training strategies are applied to each phase and detailed as follows.

In the first training phase, the embedding function f_ϕ is initialized with the corresponding deep CNNs pretrained on ImageNet (Deng et al., 2009), and weights in g_θ are initialized by a Glorot uniform initializer. Eq. (1) is employed as the loss of the network, and Nestrov Adam (Dozat, 2015) is chosen as the optimizer, of which parameters are set as

Table 3
Numerical Results on UCM2MAI (%).

Model	m. F_1	m. F_2	m. p_e	m. r_e	m. p_l	m. r_l
VGGNet* (Simonyan and Zisserman, 2014)	32.16	32.79	35.08	34.35	21.74	22.57
VGGNet (Simonyan and Zisserman, 2014)	51.42	49.04	62.00	48.38	36.80	**27.44**
Mem-N2N-VGGNet (Sukhbaatar et al., 2015)	52.16	50.93	57.26	**50.73**	20.79	22.58
K-Branch CNN (Sumbul and Demir, 2019)	47.04	43.15	64.57	41.83	37.93	22.28
proposed PM-VGGNet	**54.42**	**51.16**	**67.35**	49.95	**47.24**	26.79
Inception-V3* (Szegedy et al., 2015)	48.03	44.37	62.22	42.80	47.36	20.43
Inception-V3 (Szegedy et al., 2015)	53.96	51.28	**65.47**	50.49	**51.03**	**32.88**
Mem-N2N-Inception-V3 (Sukhbaatar et al., 2015)	56.06	55.27	62.95	55.92	47.90	30.48
proposed PM-Inception-V3	**58.56**	**58.06**	64.17	**58.73**	46.44	26.47
ResNet* (He et al., 2016)	48.36	45.00	63.90	43.84	53.63	28.35
ResNet (He et al., 2016)	51.39	48.31	65.33	47.37	51.89	**30.54**
Mem-N2N-ResNet (Sukhbaatar et al., 2015)	54.31	51.45	63.97	50.31	44.33	24.58
proposed PM-ResNet	**56.89**	**54.11**	**69.85**	**53.38**	**55.93**	29.76
NASNet* (Zoph and Le, 2017)	43.64	39.94	58.56	38.39	46.01	19.69
NASNet (Zoph and Le, 2017)	52.03	49.43	64.24	48.75	49.99	**33.75**
Mem-N2N-NASNet (Sukhbaatar et al., 2015)	55.17	53.05	64.71	52.65	49.60	29.14
proposed PM-NASNet	**60.13**	**59.57**	**67.04**	**60.42**	**58.60**	**35.04**

CNN* is initialized with weights pretrained on ImageNet.
CNN, Mem-N2N, and PM-Net are initialized with parameters pretrained on the UCM dataset.
m.F_1 and m.F_2 indicate the mean F_1 and F_2 score.
m. p_e and m. r_e indicate mean example-based precision and recall.
m. p_l and m. r_l indicate mean label-based precision and recall.

recommended: $\beta_1 = 0.9, \beta_2 = 0.999$, and $\epsilon = 1e-08$. The learning rate is set as $2e-04$ and decayed by $\sqrt{0.1}$ when the validation loss fails to decrease for two epochs.

In the second learning phase, we initialize f_ϕ with parameters learned in the previous training stage and employ the Glorot uniform initializer to initialize all weights in Q_h, V_h, K_h, and the last fully-connected layer. L and U are set to the same value of 256, and the number of heads is defined as 20. Notably, all weights are trainable, and the embedding function is tuned during the second training phase as well. Differences between two training phases are summarized in Table 2. Multiple scene-level labels are encoded as multi-hot vectors, where 0 indicates the absence of the corresponding scene while 1 refers to existing scenes. Accordingly, the loss is defined as binary cross-entropy. The optimizer is the same as that in the first training phase, but here we make use of a relatively large learning rate, $5e-4$. The network is implemented on TensorFlow and trained on one NVIDIA Tesla P100 16 GB GPU for 100 epochs. We set the size of training batch to 32 for both training phases.

4.3. Evaluation metrics

For the purpose of evaluating the performance of networks quantitatively, we utilize example-based F_1 (Wu and Zhou, 2016) and F_2 (Van Rijsbergen, 1979) scores as evaluation metrics and calculate them with the following equation:

$$F_\beta = (1+\beta^2)\frac{p_e r_e}{\beta^2 p_e + r_e}, \beta = 1, 2, \tag{7}$$

where p_e and r_e denote example-based precision and recall (Tsoumakas and Vlahavas, 2007). We calculate p_e and r_e as follows:

$$p_e = \frac{TP_e}{TP_e + FP_e}, r_e = \frac{TP_e}{TP_e + FN_e}, \tag{8}$$

where FN_e, FP_e, and TP_e represent numbers of false negatives, false positives, and true positives in an example, respectively. In our case, an example is a multi-scene aerial image, and by averaging scores of all examples in the test set, the mean example-based F scores, precision, and recall can be eventually computed. In addition to example-based evaluation metrics, we also calculate label-based precision p_l and recall r_l with Eq. 8 but replace FN_e,FP_e, and TP_e with numbers of false negatives, false positives, and true positives in respect of each scene category. The mean p_l and r_l can then be calculated. Note that principle indexes are the mean F_1 and F_2 scores.

4.4. Results on UCM2MAI

For a comprehensive evaluation, we compare the proposed PM-Net with two baselines, CNN* and CNN. The former is initialized with parameters pretrained on ImageNet, and the latter is pretrained on single-scene datasets. Besides, we compare our network with a memory network, Mem-N2N (Sukhbaatar et al., 2015). Since Mem-N2N was proposed for the question answering task, we adapt it to our task by replacing its inputs, i.e., embeddings of *questions* and *statements*, with *query image representations* $f_\phi(\boldsymbol{X})$ and *scene prototypes* $\boldsymbol{p}_s$, respectively. To

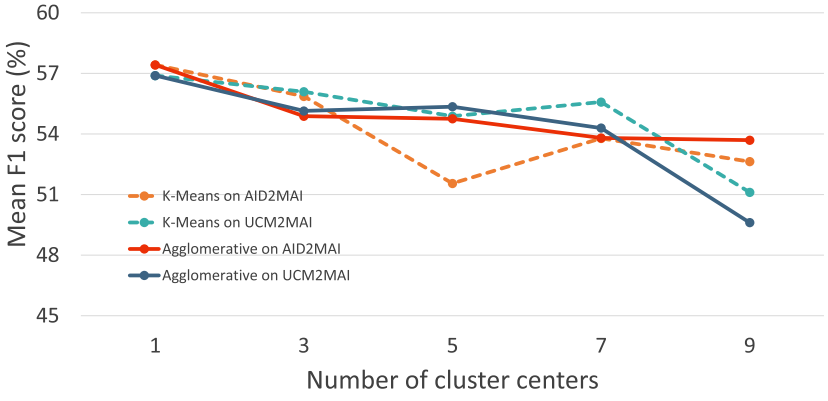


Fig. 9. The influence of the number of cluster centers on both dataset configurations. K-Means (turquoise and orange dash lines) and Agglomerative (blue and red lines) clustering algorithms are tested with PM-ResNet on both UCM2MAI and AID2MAI, respectively.

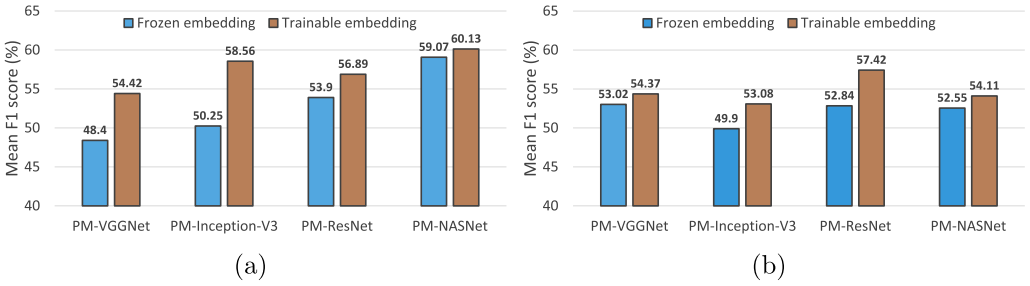


Fig. 10. Comparisons between freezing and fine-tuning embedding functions on (a) UCM2MAI and (b) AID2MAI, respectively. Blue bars represent the performance of PM-Net with frozen embedding functions, and brown bars denote the performance of PM-Net with trainable embedding functions.

be more specific, we feed $\boldsymbol{X}$ to a CNN backbone and take its output as the input of Mem-N2N. Scene prototypes are stored in the memory of Mem-N2N and retrieved according to $f_\phi(\boldsymbol{X})$. The initialization of f_ϕ is the same as that of our network, and the entire Mem-N2N is trained in an end-to-end manner. Various backbones of embedding functions are test, and quantitative results are reported in Table 3. Besides, we also compare with a multi-attention driven multi-label classification network, termed as K-Branch CNN (Sumbul and Demir, 2019). K-Branch samples images into K spatial resolutions and extracts their features with separate branches. Afterwards, a bidirectional recurrent neural network is employed to encode their relationships for inferring multiple labels. In our experiments, K is set as default, 3, and input sizes of the three branches are $224 \times 224, 112 \times 112$, and 56×56, respectively. Here we analyze results from the following three perspectives.

4.4.1. The effectiveness of learnt single-scene prototypes

To demonstrate the effectiveness of the prototype-inhabiting external memory, here we focus on comparisons between PM-Net and standard CNNs. In Table 3, PM-VGGNet increases the mean F_1 and F_2 scores by 3.00% and 2.12%, respectively, with respect to VGGNet, and PM-ResNet obtains increments of 5.50% and 5.80% in the mean F_1 and F_2 scores compared to ResNet. Besides, it is interesting to observe that PM-NASNet achieves not only the best mean F_1 and F_2 scores (60.13% and 59.57%) but also relatively high example-based precision and recall in comparison with other competitors. This demonstrates that employing NASNet as the embedding function can enhance the robustness of PM-Net. Comparisons between PM-Inception-V3 with Inception-V3 show that the external memory module contributes to improvements of 4.60% and 6.78% in the mean F_1 and F_2 scores, respectively. To summarize, memorizing and leveraging scene prototypes learned from huge quantities of single-scene images can improve the performance of network in multi-label scene recognition when limited training samples are available. For a deep insight, we further conduct ablation studies on the prototype modality and embedding function.

Single- vs. multi-prototype representations. We note that images collected over variant countries show high intra-class variability, and therefore, we wonder whether learning multi-prototype scene representations could improve the effectiveness of PM-Net. Specifically, instead of yielding scene prototypes via Eq. 2, we partition representations of single-scene aerial images belonging to the same scene into several clusters and take cluster centers as multi-prototype representations of each scene. In our experiments, we test two clustering methods, K-Means (Lloyd, 1982) and Agglomerative (Zepeda-Mendoza and Resendis-Antonio, 2013), with PM-ResNet on both UCM2MAI and AID2MAI, and results are shown in Fig. 9. We can see that the performance of PM-ResNet is decreased with the increasing number of cluster centers either using K-Means or Agglomerative clustering algorithms. Explanations could be that there are no obvious subclusters within each scene category (cf. Fig. 13), and thus PM-Net does not benefit from fine-grained multi-prototype representations.

Frozen vs. trainable embedding function. The embedding function plays a key role in both scene prototype learning and memory retrieval. In the former, we train the embedding function on single-scene images, while in the latter, the function is fine-tuned on multi-scene images. To explore the effectiveness of fine-tuning, we conduct experiments on freezing the embedding function when learning the memory retrieval module. The comparisons between PM-Net learned with frozen and trainable embedding functions are shown in Fig. 10. It can be observed that PM-Net with a trainable embedding function shows higher performance on both UCM2MAI and AID2MAI configurations. The reason could be that sources of single- and multi-scene images are variant, and fine-tuning can narrow their gaps.

Triplet vs. cross-entropy loss. Triplet loss (Schroff et al., 2015) is known as learning discriminative representations by minimizing distances between embeddings of the same class while pushing away those of different classes. To study its performance in our task, we train the embedding function by replacing Eq. 1 with the following equation:

$$\mathscr{L}(\boldsymbol{X}_i^s) = \boldsymbol{max}(||\boldsymbol{f}_\phi(\boldsymbol{X}_i^s) - \boldsymbol{f}_\phi(\boldsymbol{X}_{\boldsymbol{pos}}^s)||^2 - ||\boldsymbol{f}_\phi(\boldsymbol{X}_i^s) - \boldsymbol{f}_\phi(\boldsymbol{X}_{\boldsymbol{neg}}^s)||^2 + \alpha, \boldsymbol{0}), \tag{9}$$

where $\boldsymbol{X}_{\boldsymbol{pos}}^s$ and $\boldsymbol{X}_{\boldsymbol{neg}}^s$ denote positive and negative samples, i.e., images belonging to common and different classes, respectively, and α is set as default, 0.5. The trained embedding function is then utilized to extract scene prototypes and initialize f_ϕ in the phase of learning the memory retrieval module. Besides, all the other setups are remained the same. We compare the performance of PM-Net using embedding functions

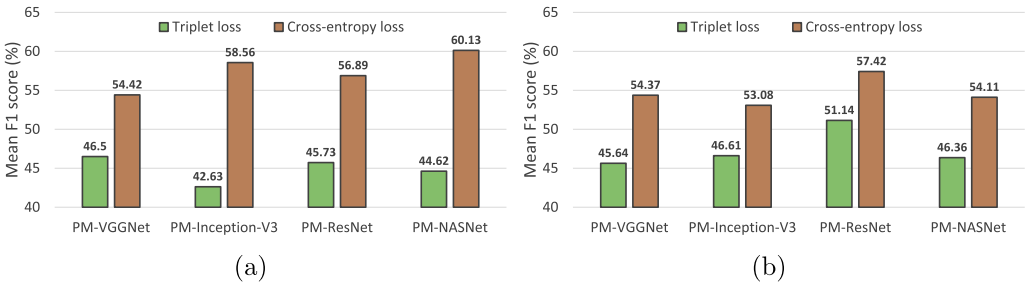


Fig. 11. Comparisons of different loss functions on (a) UCM2MAI and (b) AID2MAI, respectively. Green bars denote the performance of PM-Net using embedding functions trained by the triplet loss, and brown bars denote the performance of PM-Net with the cross-entropy loss as $\mathscr{L}$.

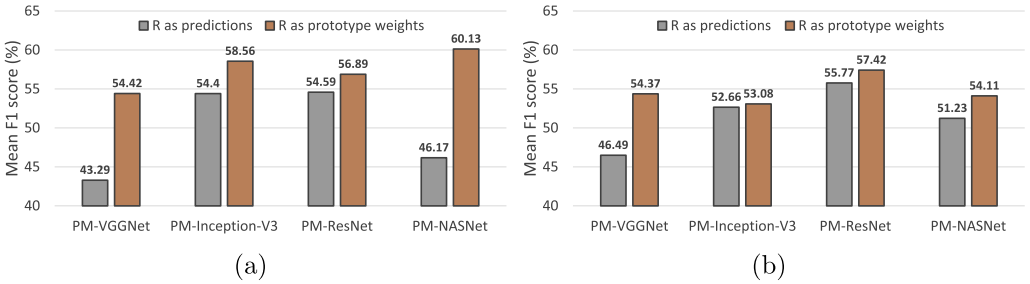


Fig. 12. Comparisons between taking relevance $R(\boldsymbol{X},\boldsymbol{M})$ as predictions and prototype weights on (a) UCM2MAI and (b) AID2MAI, respectively. Gray and brown bars represent the performance of PM-Net making predictions from relevances and aggregated scene prototypes, respectively.

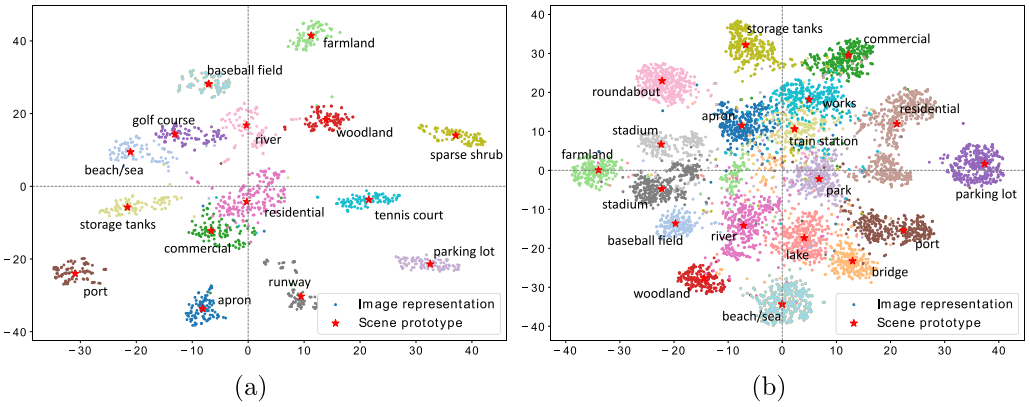


Fig. 13. T-SNE visualization of image representations and scene prototypes learned by VGGNet on (a) UCM and (b) AID datasets, respectively. Dots in the same color represent features of images belonging to the same scene, and stars denote scene prototypes.

trained through different loss functions in Fig. 11. It can be seen that training embedding functions with the triplet loss leads to decrements of the network performance. This can be attributed to that limited numbers of positive and negative samples in each batch can lead to local optimum. More specifically, the size of training batches is 32, and the number of scenes are 16 and 20 in UCM2MAI and AID2MAI, respectively. Thus, it is high probably that only a certain number of scenes are included in one batch, and comprehensively modeling relations between embeddings of samples from all scenes is infeasible. This also illustrates the larger performance decay on UCM2MAI compared to AID2MAI.

4.4.2. *The effectiveness of our multi-head attention-based memory retrieval module*

As a key component of the proposed PM-Net, the multi-head attention-based memory retrieval module is designed to retrieve scene prototypes from the external memory, and we evaluate its effectiveness by comparing PM-Net with Mem-N2N. As shown in Table 3, PM-Net outperforms Mem-N2N with variant embedding functions. Specifically, PM-VGGNet increases the mean F_1 and F_2 scores by 2.26% and 0.23%, respectively, compared to Mem-N2N-VGGNet. While taking ResNet as the embedding function, the improvement can reach 2.58% in the mean F_1 score. Besides, the highest increments of mean F_1 and F_2 scores, 4.96% and 6.52, are achieved by PM-NASNet. These observations demonstrate that our memory retrieval module plays a key role in inferring multiple aerial scenes. An explanation could be that compared to the memory reader in Mem-N2N, our module comprise multiple heads, and each of them focuses on encoding a specific relevance between the query image and variant scene prototypes. In this case, more comprehensive scene-related memories can be used for inferring multiple scene labels. Moreover, we analyze the influence of the number of heads in the memory retrieval module. Fig. 8 shows mean F_1 scores achieved by PM-Net with variant head numbers on both UCM2MAI and AID2MAI. We can observe that the network performance is first boosted with an increasing number of heads and then decreased gradually when the number exceeds 20.

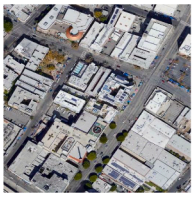
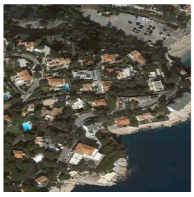
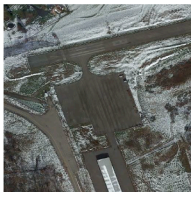
Moreover, we also conduct experiments on directly utilizing relevances for inferring multiple scene labels. Specifically, we set the number of heads to 1 and replace the softmax activation in Eq. 4 with the sigmoid function. Relevances between the query image and scene prototypes can then be interpreted as the existence of each scene. We compare it with our memory retrieval module on variant backbones, and results are shown in Fig. 12. We can see that utilizing relevances $R(\boldsymbol{X},\boldsymbol{M})$ as weights for aggregating scene prototypes leads to higher network performance.

4.4.3. *The benefit of exploiting single-scene training samples*

Let's start with the conclusion: exploiting single-scene images significantly contributes to our task. To analyze its benefit, we mainly compare CNNs* and CNNs. It can be observed that even with identical network architectures, the performance of CNN is superior to that of CNN*. More specifically, VGGNet achieves the highest improvement of the mean F_1 scores, 19.26%, in comparison with VGGNet*. NASNet shows higher performance in all metrics compered to ResNet*, while other CNNs perform poorly in only the mean example-based precision with respect to their corresponding CNNs*. Besides, we visualize features of single-scene images learned by VGGNet on UCM and AID datasets via t-SNE, respectively. As shown in Fig. 13, extracted features are discriminative and separable in the embedding space, which demonstrates the effectiveness of learning the embedding function on single-scene aerial image datasets. To summarize, except for learning scene prototypes, single-scene training samples can also benefit multi-label scene interpretation by pretraining CNNs which are further utilized to initialize the embedding function.

We exhibit several example predictions of PM-ResNet trained on UCM2MAI in Table 4. False positives are marked as red, while false negatives are in blue. As shown in the forth example at the top row, we

Table 4
Example images and predictions on UCM2MAI.

Sample Multi-scene Aerial Images from MAI Dataset				
Ground Truths	farmland, woodland, residential	commercial, parking lot, residential	woodland, farmland	commercial, beach, parking lot, residential
Predictions	farmland, woodland, residential	commercial, parking lot, residential	woodland, farmland	commercial, beach, parking lot, residential
Sample Multi-scene Aerial Images from MAI Dataset				
Ground Truths	farmland, parking lot, residential	baseball field, parking lot, residential, tennis court	beach, parking lot, woodland, residential, sea	apron, runway
Predictions	farmland, parking lot, residential	baseball field, parking lot, residential, tennis court	commercial, beach, parking lot, woodland, residential, sea	apron, residential, runway, parking lot

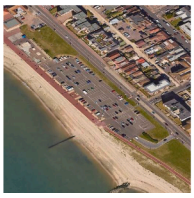
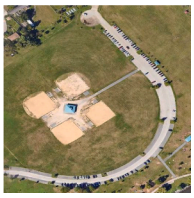
Blue predictions are false negatives, while red predictions indicate false positives.

Table 5
Numerical results on AID2MAI (%).

Model	m. F_1	m. F_2	m. p_e	m. r_e	m. p_l	m. r_l
VGGNet* (Simonyan and Zisserman, 2014)	41.57	36.36	64.02	34.04	25.98	12.80
VGGNet (Simonyan and Zisserman, 2014)	48.30	50.80	48.53	**54.19**	32.89	**44.75**
Mem-N2N-VGGNet (Sukhbaatar et al., 2015)	45.92	43.17	56.16	42.22	23.10	18.76
K-Branch CNN (Sumbul and Demir, 2019)	47.67	43.88	63.84	42.37	26.53	16.15
proposed PM-VGGNet	**54.37**	**51.44**	**65.69**	50.39	**48.06**	22.40
Inception-V3* (Szegedy et al., 2015)	45.92	40.76	66.17	38.43	39.56	14.71
Inception-V3 (Szegedy et al., 2015)	51.81	49.44	62.91	48.93	45.26	**36.32**
Mem-N2N-Inception-V3 (Sukhbaatar et al., 2015)	52.13	**53.83**	52.53	**56.21**	33.33	29.05
proposed PM-Inception-V3	**53.08**	49.26	**69.42**	47.85	**48.20**	24.65
ResNet* (He et al., 2016)	50.06	46.88	64.32	45.98	39.48	22.34
ResNet (He et al., 2016)	54.74	52.76	65.54	52.62	47.54	**40.23**
Mem-N2N-ResNet (Sukhbaatar et al., 2015)	53.26	**60.41**	46.15	**68.07**	23.75	30.21
proposed PM-ResNet	**57.42**	54.34	**70.62**	53.33	**55.34**	29.55
NASNet* (Zoph and Le, 2017)	47.53	42.93	**65.57**	40.94	34.79	16.42
NASNet (Zoph and Le, 2017)	53.08	50.68	64.33	50.17	**46.68**	**37.43**
Mem-N2N-NASNet (Sukhbaatar et al., 2015)	39.27	40.72	38.52	42.38	20.03	20.41
proposed PM-NASNet	**54.11**	**52.39**	64.03	**52.30**	43.16	33.99

CNN, Mem-N2N, and PM-Net are initialized with parameters pretrained on the AID dataset.

Table 6
Example images and predictions on AID2MAI.

Sample multi-scene aerial images from MAI dataset				
Ground Truths	bridge, river, commercial, parking lot, residential	beach, commercial, parking lot, residential, sea	bridge, farmland, river, woodland	baseball field, parking lot, park, residential
Predictions	bridge, river, commercial, parking lot, residential	beach, commercial, parking lot, residential, sea	bridge, farmland, river, woodland	baseball field, parking lot, park, residential
Sample multi-scene aerial images from MAI dataset				
Ground Truths	beach, commercial, parking lot, residential, sea	bridge, woodland, river, storage tanks	baseball field, commercial, parking lot, park, residential, soccer field	baseball field, parking lot, soccer field
Predictions	beach, commercial, parking lot, residential, sea, woodland	bridge, woodland, farmland, river, storage tanks, parking lot	baseball field, commercial, woodland, parking lot, park, soccer field, residential	baseball field, commercial, woodland, parking lot, park, soccer field, residential

see that PM-Net can accurately perceive aerial scenes even in complex contexts, but unseen scene appearance (i.e. apron and runway in snow) can influence its prediction.

4.5. Results on AID2MAI

Table 5 reports numerical results on the AID2MAI configuration. It can be seen that the performance of PM-Net is superior to all competitors in the mean F_1 score. Compared to Mem-N2N-VGGNet, the proposed PM-VGGNet increases the mean F_1 and F_2 scores by 6.70% and 7.56%, respectively, while improvements reach 6.07% and 0.64% in comparison with VGGNet. PM-ResNet achieves the best mean F_1 score and example-based precision, 57.42% and 70.62, respectively. With NASNet as the backbone, exploiting the proposed memory retrieval module contributes to increments of 1.03% and 1.71% in mean F_1 and F_2 scores compared to directly learning NASNet on a small number of multi-scene samples.

We present some example predictions of PM-ResNet in Table 6. As shown in the top row, PM-ResNet learned with a limited number of annotated multi-scene images can accurately identify various aerial scenes even image contextual information is complicated. The bottom row shows some inaccurate predictions. It can be observed that although bridge and parking lot account for relatively small areas in last two examples at the top row, the proposed PM-Net can successfully detect them. Similar observations can also be found in the first and third example at the bottom row that residential and parking lot are recognized by our network, even they are located at the corner. In conclusion, quantitative results illustrate the effectiveness of our network in learning to perform unconstrained multi-scene classification, and example predictions further demonstrate it.

5. Conclusion

In this paper, we propose a novel multi-scene recognition network, namely PM-Net, to tackle both the problem of aerial scene classification in the wild and scarce training samples. To be more specific, our network

consists of three key elements: 1) a prototype learning module for encoding prototype representations of variant aerial scenes, 2) a prototype-inhabiting external memory for storing high-level scene prototypes, and 3) a multi-head attention-based memory retrieval module for retrieving associated scene prototypes from the external memory for recognizing multiple scenes in a query aerial image. For the purpose of facilitating the progress as well as evaluating our method, we propose a new dataset, MAI dataset, and experiment with two dataset configurations, UCM2MAI and AID2MAI, based on two single-scene aerial image datasets, UCM and AID. In scene prototype learning, we train the embedding function on most of single-scene images as we aim to simulate the real-life scenario, where massive single-scene samples can be collected at low cost by resorting to OSM data. To learn memory retrieval, our network is fine-tuned on only around 100 training samples from the MAI dataset. Experimental results on both UCM2MAI and AID2MAI illustrate that learning and memorizing scene prototypes with our PM-Net can significantly improve the classification accuracy. The best performance is achieved by employing ResNet as the embedding function, and the best mean F_1 score reaches nearly 0.6. We hope that our work can open a new door for further researches in a more complicated and challenging task, multi-scene interpretation in single images. Looking into the future, we intend to apply the proposed network to the recovery of weakly supervised scenes.

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.

Acknowledgements

This work 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 Framework of Helmholtz AI [Grant No.: 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" and 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 No.: 01DD20001)

References

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

Bi, Q., Qin, K., Li, Z., Zhang, H., Xu, K., Xia, G., 2020. A multiple-instance densely-connected ConvNet for aerial scene classification. IEEE Trans. Image Process. 29, 4911–4926.

Byju, A., Sumbul, G., Demir, B., Bruzzone, L., 2000. Remote sensing image scene classification with deep neural networks in JPEG 2000 compressed domain, arXiv: 2006.11529.

Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: Benchmark and state of the art. Proc. IEEE 105 (10), 1865–1883.

Cheng, G., Xie, X., Han, J., Guo, L., Xia, G., 2020. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 3735–3756.

Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R., 2020. Meshed-memory transformer for image captioning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F., 2009. ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. North American Chapter of the Association for Computational Linguistics.

Dozat, T., 2015. Incorporating Nesterov momentum into Adam, http://cs229.stanford.edu/proj2015/054_report.pdf, online.

Guerriero, S., Caputo, B., M. T., 2018. DeepNCM: Deep nearest class mean classifiers.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Hua, Y., Mou, L., Zhu, X.X., 2019. Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification. ISPRS J. Photogram. Remote Sens. 149, 188–199.

Hua, Y., Mou, L., Zhu, X.X., 2020. Relation network for multilabel aerial image classification. IEEE Trans. Geosci. Remote Sensing.

Huang, L., Huang, Y., Ouyang, W., Wang, L., 2020. Relational prototypical network for weakly supervised temporal action localization. In: AAAI Conference on Artificial Intelligence.

Hu, D., Li, X., Mou, L., Jin, P., Chen, D., Jing, L., Zhu, X.X., Dou, D., 2020. Cross-task transfer for multimodal aerial scene recognition, arXiv:2005.08449.

Jin, P., Xia, G., Hu, F., Lu, Q., Zhang, L., 2018. AID++: An updated version of aid on scene classification. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS).

Khan, N., Chaudhuri, U., Banerjee, B., Chaudhuri, S., 2019. Graph convolutional network for multi-label VHR remote sensing scene recognition. Neurocomputing 357, 36–46.

Koda, S., Zeggada, A., Melgani, F., Nishii, R., 2018. Spatial and structured SVM for multilabel image classification. IEEE Trans. Geosci. Remote Sens. 56 (10), 5948–5960.

Lai, Z., Lu, E., Xie, W., 2020. MAST: A memory-augmented self-supervised tracker. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Lee, W., Kim, S., Lee, Y., Lee, H., Choi, M., 2017. Deep neural networks for wild fire detection with unmanned aerial vehicle. In: IEEE International Conference on Consumer Electronics (ICCE).

Li, Q., Mou, L., Liu, Q., Wang, Y., Zhu, X.X., 2018. HSF-Net: Multi-scale deep feature embedding for ship detection in optical remote sensing imagery. IEEE Trans. Geosci. Remote Sens.

Li, Q., Qiu, C., Ma, L., Schmitt, M., Zhu, X.X., 2020a. Mapping the land cover of Africa at 10 m resolution from multi-source remote sensing data with Google Earth engine. Remote Sensing 12 (4), 602.

Li, Q., Shi, Y., Huang, X., Zhu, X.X., 2020b. Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF). IEEE Trans. Geosci. Remote Sens.

Lin, J., Mou, L., Yu, T., Zhu, X.X., Wang, Z.J., 2020. Dual adversarial network for unsupervised ground/satellite-to-aerial scene adaptation. In: ACM International Conference on Multimedia (ACMMM).

Liu, E., Mercado III, E., Church, B., Orduña, I., 2008. The easy-to-hard effect in human (homo sapiens) and rat (rattus norvegicus) auditory identification. J. Comp. Psychol. 122 (2), 132.

Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28 (2), 129–137.

Long, Y., Xia, G., Li, S., Yang, W., Yang, M., Zhu, X.X., Zhang, L., Li, D., 2020. DiRS: On creating benchmark datasets for remote sensing image interpretation, arXiv: 2006.12485.

Lucchesi, S., Giardino, M., Perotti, L., 2013. Applications of high-resolution images and DTMs for detailed geomorphological analysis of mountain and plain areas of NW Italy. Eur. J. Remote Sens. 46 (1), 216–233.

Marcos, D., Volpi, M., Kellenberger, B., Tuia, D., 2018. Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models. ISPRS J. Photogramm. and Remote Sens. 145, 96–107.

Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. ISPRS J. Photogramm. Remote Sens. 135, 158–172.

McLaren, I., Suret, M., 2000. Transfer along a continuum: Differentiation or association. In: Annual Conference of the Cognitive Science Society.

Miller, A., Fisch, A., Dodge, J., Karimi, A., Bordes, A., Weston, J., 2016. Key-value memory networks for directly reading documents. In: Empirical Methods in Natural Language Processing (EMNLP).

Mou, L., Zhu, X.X., 2016. Spatiotemporal scene interpretation of space videos via deep neural network and tracklet analysis. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS).

Mou, L., Zhu, X.X., 2018. IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network, arXiv:1802.10249.

Mou, L., Zhu, X.X., 2018. RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images, arXiv:1805.02091.

Mou, L., Zhu, X.X., 2018c. Vehicle instance segmentation from aerial image and video using a multi-task learning residual fully convolutional network. IEEE Trans. Geosci. Remote Sens. 56 (11), 6699–6711.

Murray, J., Marcos, D., Tuia, D., In, Zoom, 2019. Zoom Out: Injecting scale invariance into landuse classification CNNs. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS).

National Research Council, 2000. How people learn: Brain, mind, experience, and school: Expanded edition.

Niazmardi, S., Demir, B., Bruzzone, L., Safari, A., Homayouni, S., 2017. Multiple kernel learning for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. 56 (3), 1425–1443.

Park, H., Noh, J., Ham, B., 2020. Learning memory-guided normality for anomaly detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Qiu, C., Mou, L., Schmitt, M., Zhu, X.X., 2019. Local climate zone-based urban land cover classification from multi-seasonal Sentinel-2 images with a recurrent residual network. ISPRS J. Photogramm. Remote Sens. 154, 151–162.

Qiu, C., Schmitt, M., Geiß, C., Chen, T., Zhu, X.X., 2020. A framework for large-scale mapping of human settlement extent from Sentinel-2 images via fully convolutional neural networks. ISPRS J. Photogramm. Remote Sens. 163, 152–170.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving language understanding by generative pre-training.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language models are unsupervised multitask learners.
Ru, L., Du, B., Wu, C., 2020. Multi-temporal scene classification and scene change detection with correlation based fusion, arXiv:2006.02176.
Schroff, F., Kalenichenko, D., Philbin, J., 2015. FaceNet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Shi, Z., Fang, H., Tai, Y., Tang, C., 2019. DAWN: Dual augmented memory network for unsupervised video object tracking, arXiv:1908.00777.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556.
Snell, J., Swersky, K., Zemel, R., 2017. Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems (NeurIPS).
Sukhbaatar, S., Szlam, A., Weston, J., Fergus, R., 2015. End-to-end memory networks. In: Advances in Neural Information Processing Systems (NeurIPS).
Sumbul, G., Demir, B., 2019. A novel multi-attention driven system for multi-label remote sensing image classification. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS).
Sumbul, G., Charfuelan, M., Demir, B., Markl, V., 2019. BigEarthNet: A large-scale benchmark archive for remote sensing image understanding. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS).
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Tang, H., Li, Y., Han, X., Huang, Q., Xie, W., 2019. A spatial–spectral prototypical network for hyperspectral remote sensing image. IEEE Geosci. Remote Sens. Lett. 17 (1), 167–171.
Tsoumakas, G., Vlahavas, I., 2007. Random K-labelsets: An ensemble method for multilabel classification. In: European Conference on Machine Learning (ECML).
Tuia, D., Marcos, D., Camps-Valls, G., 2016. Multi-temporal and multi-source remote sensing image classification by nonlinear relative normalization. ISPRS J. Photogramm. Remote Sens. 120, 1–12.
Van Rijsbergen, C.J., 1979. Information Retrieval.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS).
Vetrivel, A., Gerke, M., Kerle, N., Nex, F., Vosselman, G., 2018. Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. ISPRS J. Photogramm. Remote Sens. 140, 45–59.
Wang, X., Xiong, X., Ning, C., 2019. Multi-label remote sensing scene classification using multi-bag integration. IEEE Access 7, 120399–120410.
Wen, D., Huang, X., Liu, H., Liao, W., Zhang, L., 2017. Semantic classification of urban trees using very high resolution satellite imagery. IEEE J. Sel. Top. Appl. Earth Obser. Remote Sens. 10 (4), 1413–1424.
Weng, Q., Mao, Z., Lin, J., Liao, X., 2018. Land-use scene classification based on a CNN using a constrained extreme learning machine. Int. J. Remote Sens. 1–19.
Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A., 2020. Transformers: State-of-the-art natural language processing. In: Empirical Methods in Natural Language Processing: System Demonstrations.
Wu, X., Zhou, Z., 2016. sA unified view of multi-label performance measures, arXiv: 1609.00288.
Xia, G., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. IEEE Trans. Geosci. Remote Sens.
Xu, Y., Du, B., Zhang, L., 2020. Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses. IEEE Trans. Geosci. Remote Sensing.
Yang, Y., Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification, in. In: International Conference on Advances in Geographic Information Systems (SIGSPATIAL).
Yang, H., Zhang, X., Yin, F., Liu, C., 2018. Robust classification with convolutional prototype learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Zarco-Tejada, P., Diaz-Varela, R., Angileri, V., Loudjani, P., 2014. Tree height quantification using very high resolution imagery acquired from an unmanned aerial vehicle (UAV) and automatic 3D photo-reconstruction methods. Eur. J. Agron. 55, 89–99.
Zegeye, B., Demir, B., 2018. A novel active learning technique for multi-label remote sensing image scene classification. In: Image and Signal Processing for Remote Sensing.
Zeggada, A., Melgani, F., Bazi, Y., 2017. A deep learning approach to UAV image multilabeling. IEEE Geosci. Remote Sens. Lett. 14 (5), 694–698.
Zepeda-Mendoza, M., Resendis-Antonio, O., 2013., Hierarchical Agglomerative Clustering.
Zhang, C., Yue, J., Qin, Q., 2020. Global prototypical network for few-shot hyperspectral image classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 4748–4759.
Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosci. Remote Sensing Mag. 5 (4), 8–36.
Zhu, Q., Zhong, Y., Zhang, L., Li, D., 2018. Adaptive deep sparse semantic modeling framework for high spatial resolution image scene classification. IEEE Trans. Geosci. Remote Sens. 56 (10), 6180–6195.
Zhu, Q., Sun, X., Zhong, Y., Zhang, L., 2019. High-resolution remote sensing image scene understanding: A review. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS).
Zoph, B., Le, Q., 2017. Neural architecture search with reinforcement learning. In: International Conference on Learning Representations (ICLR).