remote sensing

MDPI

Article

Multi-Task cGAN for Simultaneous Spaceborne DSM Refinement and Roof-Type Classification

Ksenia Bittner [1,*], Marco Körner [2], Friedrich Fraundorfer [3] and Peter Reinartz [1]

1 German Aerospace Center (DLR), Remote Sensing Technology Institute, Münchner Str. 20, 82234 Weßling, Germany; peter.reinartz@dlr.de
2 Technical University of Munich (TUM), Department of Civil, Geo and Environmental Engineering, Chair of Remote Sensing Technology, Arcisstraße 21, 80333 Munich, Germany; marco.koerner@tum.de
3 Institute for Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria; fraundorfer@icg.tugraz.at
* Correspondence: ksenia.bittner@dlr.de; Tel.: +49-8153-28-4285

Received: 12 April 2019; Accepted: 24 May 2019; Published: 28 May 2019

Abstract: Various deep learning applications benefit from multi-task learning with multiple regression and classification objectives by taking advantage of the similarities between individual tasks. This can result in improved learning efficiency and prediction accuracy for the task-specific models compared to separately trained models. In this paper, we make an observation of such influences for important remote sensing applications like elevation model generation and semantic segmentation tasks from the stereo half-meter resolution satellite *digital surface models (DSMs)*. Mainly, we aim to generate good-quality DSMs with complete, as well as accurate *level of detail (LoD)*2-like building forms and to assign an object class label to each pixel in the DSMs. For the label assignment task, we select the roof type classification problem to distinguish between flat, non-flat, and background pixels. To realize those tasks, we train a *conditional generative adversarial network (cGAN)* with an objective function based on least squares residuals and an auxiliary term based on normal vectors for further roof surface refinement. Besides, we investigate recently published deep learning architectures for both tasks and develop the final end-to-end network, which combines different models, as using them first separately, they provide the best results for their individual tasks.

Keywords: multi-task learning; conditional generative adversarial networks; digital surface model; 3D scene refinement; semantic segmentation; roof type classification; urban region; satellite imagery

1. Introduction

DSMs generated by traditional photogrammetry with multi-view stereo images from space have been proven to be reliable and cost-effective for larger regions and are especially suitable for remote areas and developing countries [1–3]. They provide elevation information, which is an attractive element for many geoscience application, such as city management, navigation, tourism, disaster analysis, virtual environment generation, etc. Although spaceborne DSMs show sub-meter spatial resolution and wide coverage, some unwanted blunders and spikes are occurring in automatically generated DSMs due to image data noise, shadows, hidden points and surfaces, temporal changes within the stereo image pairs, or matching errors [4,5]. This introduces difficulties for building detection and reconstruction. Hence, the development of methodologies able to improve DSMs automatically to a higher quality level with more accurate and complete building geometries is in demand.

However, not only surface topography contains useful data for achieving this goal. The 2D information in the form of pixel-wise semantic segmentation is also very crucial for many remote sensing applications, as it provides additional knowledge about object boundaries or categories to

which the object belongs. Mainly, building footprint extraction or roof type classification is one of the most challenging, but important problems. It is common to use DSMs as input data for classification tasks regarding buildings [6,7], as depth information provides geometrical silhouettes and allows a better understanding of building forms. Although a vast amount of attempts have already been made on accurate pixel-wise classification [8,9], it still remains a challenging task in practice due to the wide variety of building appearances.

In most cases, each task, e.g., depth image generation and pixel-wise image classification, is tackled independently, although they are closely connected. Solving those multiple tasks jointly can enhance the performance of each independent task, as well as speed up computation time. This observation leads to the advantages of *multi-task (MT)* learning. The approach of simultaneously improving the generalization performance of multiple outputs from a single input was applied to numerous machine learning techniques. As a promising concept for *convolutional neural networks (CNNs)*, MT learning has been proven to leverage a variety of problems successfully, like classification and semantic segmentation [10] or classification and object detection [11]. Due to the fact that different tasks may conflict, MT learning is regarded as the optimization of MT loss, which minimizes a linear combination of contributed single-task loss functions.

In this work, we aim to produce good-quality LoD2-like DSMs with realistic building geometries together with dense pixel-wise rooftop classification masks, defining multiple classes, like ground, flat, and non-flat roofs, derived from the given low-quality elevation models generated by spaceborne half-meter resolution stereo imagery. Moreover, the proposed methodology tackles the desired multiple outputs in a joint end-to-end CNN architecture. The CNN-based approaches have been confirmed to be state-of-the-art in the fields of image recognition [12,13] and generation [14–18]. Therefore, our network combines multiple *fully-convolutional networks (FCNs)* in parallel within one cGAN architecture to produce two outputs at once: a refined elevation model and a rooftop classification mask. Each FCN stream is responsible for the individual task, but benefits from the joint learning. As a result, such a knowledge transfer can help to improve generalization by sharing the domain information between complementary tasks. Additionally, the adversarial manner of training the final MT-cGAN model supports the results, so that the generated images become more similar to the real ones.

The remainder of this paper is organized as follows. The related work for pixel-wise image classification, depth image regression, as well as their simultaneous learning by traditional and deep learning-based methodologies is summarized in Section 2. The background of *generative adversarial networks (GANs)*, the objective functions used, and details of the proposed deep network architecture are described in Section 3. In Section 4, we introduce the dataset used and present implementation details and training setups. The experimental results obtained with different network architectures together with their qualitative and quantitative evaluation are shown and discussed in Section 5. Section 6 concludes the paper.

2. Related Work

2.1. Pixel-Wise Image Classification

Building roof type classification is an important step in model-based approaches for 3D building reconstruction. Previously introduced methodologies for roof type classification were often based on low-level features, like edges, line segments, and corners, grouped together to form building hypotheses. For instance, Mohajeri et al. [19] used a set of handcrafted features, such as the number of roof surfaces and the distribution of the binned slope angles, extracted from *light detection and ranging (LIDAR)* DSMs, to perform roof classification through a *support vector machine (SVM)*. Assouline et al. [20] experimented with raster features and geometric features from a DSM to label rooftops using a random forest classifier. Although classical machine learning-based algorithms show promising results, they invariably face computational complexity challenges caused by the high dimensionality of data sources [21].

With recent advances in the field of artificial neural networks, it became possible to learn image features automatically instead of extracting them by classical methods. Innovative architectures, such as CNNs, have demonstrated the ability to classify high dimensional data sources accurately and robustly and have become the state-of-the-art for image recognition tasks. Alidoost and Arefi [22] proposed an approach based on fine-tuning of a pre-trained CNN model, which accomplishes building segmentation, feature extraction, and building roof labeling to create an automatic recognition system for various types of buildings. The training is done on both spectral and depth images separately, and the final predicted roof shape is simply taken as the highest probability result between the two models. Partovi et al. [23] fine-tuned a CNN using patched satellite images showing only building roof-tops. The authors used high-level features extracted from the last layer of a fine-tuned CNN as inputs to a second-stage SVM classifier. A similar strategy was proposed by Axelsson et al. [24], who classified the patches of buildings from RGB aerial images into the two most common roof types, i.e., slope roofs and flat roofs, using transfer learning of a pre-trained CNN. Although patch-based classification with CNNs does not require feature pre-definition anymore, it needs an additional special training data preparation and, in most cases, does not contain the whole building as one element, but only parts of it. On the contrary, our goal is to have a mask that provides the entire and complete information about the categorized building segments.

2.2. Depth Image Regression

Despite the available techniques capable of generating large-scale elevation models from high-resolution satellite images, DSMs still feature many mismatches and noise due to occlusions by dense and complex building structures, perspective differences, or stereo matching errors during their generation. However, only very few of the existing approaches were dedicated to stereo DSM enhancement tasks. The pioneer methodologies investigated the DSMs refinement by applying filtering techniques, like geostatistical [25], Kalman [26], or Gaussian [27] filters, to remove inconsistency errors. The major drawback of filtering approaches is the smoothness effect, which dramatically influences the steepness of building walls. Other methods try to enhance DSMs by interpolation routines. Examples of popular interpolation methods include *inverse distance weighting (IDW)* and kriging interpolations [28], which reduce point densities, but maintain a satisfactory DSM estimation; spline-based methods [29], which produce smooth surfaces, but with fewer recognizable characteristics; and hybrid methods that combine linear and non-linear interpolation [30]. Although these approaches achieve some progress towards DSM refinement, they are not suitable for areas with strong surface discontinuities like urban areas, as this leads to losing sharpness in building edges [5]. In recent years, researchers have paid more attention not only to removing noise and inconsistency errors from DSMs, but also keeping the forms and shapes of building constructions more accurate. For instance, Sirmacek et al. [5] extracted potential building segments through thresholding the DSM and applying a box-fitting algorithm to detect 2D building shapes. Then, 3D building forms were refined by sharpening building walls using the information from the detected building shapes and smoothing the noise in building rooftops. Sirmacek et al. [31] considered Canny edge information from spectral images in the procedure of fitting a chain of active shape models to the input data to further improve the detection of complex buildings. However, to extract the 3D building information, only one single height value from the input DSM was assigned. Although those strategies of fitting the predefined shapes lead to more detailed and sharper elevation models production, they fail if the building structures are more complicated.

More recently, CNNs have been fast emerging and have become the method of choice for many tasks. A dominant amount of work was dedicated to spectral imagery exploration, allowing not only a single category label assignment to images [32,33], but dense per-pixel predictions like semantic segmentation [34,35] and super-resolution [36,37]. Much less attention is given to depth estimation tasks. A first attempt at applying CNNs for depth estimation was done by Eigen et al. [38] and followed by Eigen and Fergus [39], where the authors achieved this through the use of two and three CNNs in stages, respectively, by regressing a global depth structure at finer resolution from a single image.

Liu et al. [40] employed a *conditional random field (CRF)* to learn the unary and pairwise potentials in a unified CNN framework for modeling the local consistency in the output image.

In contrast to the computer vision field, height image generation from single input data has rarely been addressed in the remote sensing community so far. Mou and Zhu [41] tackled a problem of height prediction from a single monocular remote sensing image using an end-to-end fully-convolutional-deconvolutional network architecture, encompassing residual learning. We approach the problem differently and generate improved elevation models from initial low-quality multi-view stereo DSM, applying a different class of neural networks, GANs [42,43]. In this work, we extend the idea of height image quality refinement via joint end-to-end training of regression and pixel-wise classification tasks.

2.3. Multi-Task Learning

Image analysis tasks, whether classification, semantic segmentation, or regression, are related to each other and can feature some things in common. As a result, one task can help to learn other tasks. Multi-task learning has been used successfully across many applications of machine learning, from natural language processing [44] and speech recognition [45] to computer vision [13]. Eigen and Fergus [39] introduced a network based on a multi-scale CNN for simultaneous prediction of depth, surface normals, and semantic labels from a single image. CNNs were applied at three different scales where the output of the smaller scale network was fed as input to the larger one. Kokkinos [46] proposed the *UberNet* architecture, which relied on diverse training datasets and simultaneously handled low-, mid-, and high-level vision tasks. Liebel and Körner [47] showed that even seemingly unrelated auxiliary tasks, like time or weather prediction, can improve semantic segmentation performance.

Recently, the remote sensing community also expanded their capability to learn alternative tasks together with common image classification. Marmanis et al. [48] predicted object boundaries jointly with the land cover task. This helps to sharpen classification maps, but has a high computational load, as the boundaries, separately detected from *color-infrared (CIR)* and height data, are added as an extra channel to each data source for further image classification task training. The work of Vakalopoulou et al. [49] introduced a model that learns jointly the registration between the images, the land use classification of each input, and a change detection map with a CRF. This is done by fusing boundary priors with the classification scores from a two-layer CNN architecture under a single energy formulation. To have a system that is able to predict reasonably accurate DSMs automatically would be very valuable. Srivastava et al. [50] proposed, to our knowledge, the first deep learning-based methodologies to predict semantic segmentation maps, as well as *normalized digital surface models (nDSMs)* from a single monocular image. The authors used a joint loss function for CNN training, which is a linear combination of a dense image classification loss and a regression loss responsible for DSM error minimization. The model is trained by alternating over two losses. Similar to Srivastava et al. [50], we investigate the prediction of semantic segmentation maps for a roof classification task and simultaneous generation of refined DSM with improved building forms out of a single photogrammetric depth image within an end-to-end neural network. Our contributions are:

- We efficiently adapt the cGAN architecture developed by Isola et al. [18] for multi-task learning.
- The proposed framework generates images with continuous values representing elevation models with enhanced building geometries and, at the same time, images with discrete values depicting the label information meaning to which class out of three (flat roof, non-flat roof, and background) every single pixel belongs.
- We investigate the potential of different network architectures for each task and select the combination of models that provides the best results for both pixel-wise classification and depth map generation. We show that joint training of multiple tasks within the end-to-end framework is beneficial. Moreover, the obtained roof classification information can be used later in a post-processing step for the final building modeling task.

- We investigate the potential of using a normal vector loss, which is included as an additional term to the objective function with least squares, thereby gaining more accurate and planar roof structures.

3. Methodology

3.1. Building Shape Improvements and Roof Type Understanding Model

Multi-task learning introduces the problem of optimizing the neural network model with respect to multiple objectives as it requires modeling the trade-off between competing tasks [51]. In this work, we performed two tasks within one model: building shape improvement and roof surface classification. The building shape refinement problem can be considered as a generative task, which has been recently successfully solved by GANs in other applications. GANs were firstly introduced by Goodfellow et al. [15] in 2014 and represent the type of machine learning technique that trains a pair of networks, namely a generator G and a discriminator D, in an adversarial manner to compete against each other. Moreover, Mirza and Osindero [52] proposed to condition the model by side information, as it restricts both the generator G in its output and the discriminator D in its expected input. The resulting cGANs allow the generation of fake images $\boldsymbol{y}$ similar to some known input image $\boldsymbol{x}$. This idea suits our task, as we aim to generate LoD2-like height images with better-quality building shapes, but with an appearance similar to the given DSMs from stereo satellite imagery.

To understand building geometry and semantics, we introduce architectures that learn to predict the pixel level depth and semantic classes from an input image. The network architecture presented by Isola et al. [18] was adapted for our purpose. Within the whole study, the G part of the network was modified from experiment to experiment, while the D part stayed unchanged.

The D part was a binary classification network constructed with five convolutional layers in our case. A *sigmoid* activation function $\sigma_{\text{sigm}}(z) = \frac{1}{1+e^{-z}}$ was placed on the top layer of discriminator D, as it was supposed to output the probability that the input image belonged either to lass 1 ("real") or Class 0 ("synthesized"). The input to discriminator D was a concatenation of conditional information, mainly given by a DSM, with either a generated fake DSM or a ground-truth DSM. The D part of architecture stayed unchanged for all the following networks.

3.1.1. One Generator, Two Outputs

As a continuation of our previous work [43], we first considered the G part of the cGAN consisting of a one-stream *UNet* [53] with two outputs producing a single-channel depth image with continuous values and a three-channel roof class probability map. The *hyperbolic tangent* activation function $\sigma_{\text{tanh}}(z) = \tanh(z)$ was applied to the top layer of the generator G responsible for depth image generation and $softmax(z) = \frac{e^{z_k}}{\sum_j e^{z_j}}$ normalization in the training phase on top of the layer producing class probability maps. The encoder part of U-Net progressively down-sampled the given low-quality DSM through eight layers and coded back the process with eight up-sampled decoder layers. To recover important details that are lost in down-sampling in the encoder, seven skip connections were added to the network. The detailed description of the UNet architecture adapted from Bittner et al. [43] is depicted in Figure 1.

Following the same strategy, we investigated a pre-trained *residual network (ResNet)* [54] consisting of 34 layers as a basis for the encoder part of the G network. Architectures based on ResNets have already achieved state-of-the-art results and were successful in several competitions for image recognition tasks. One of the problems ResNets solve is an effect known as the vanishing gradient. When the network is too deep, the gradients from where the loss function is calculated can easily shrink to zero after several applications of the chain rule. This can lead to the problem that the weights never update their values, and therefore, no learning is being performed. With ResNets, the gradients can flow directly through identity shortcut connections backward from later layers to initial filters.

To complete the decoder part of the *G* network, we up-sampled the feature maps that had been down-sampled by ResNet34 to obtain the resulting two outputs of the same size as the input image.

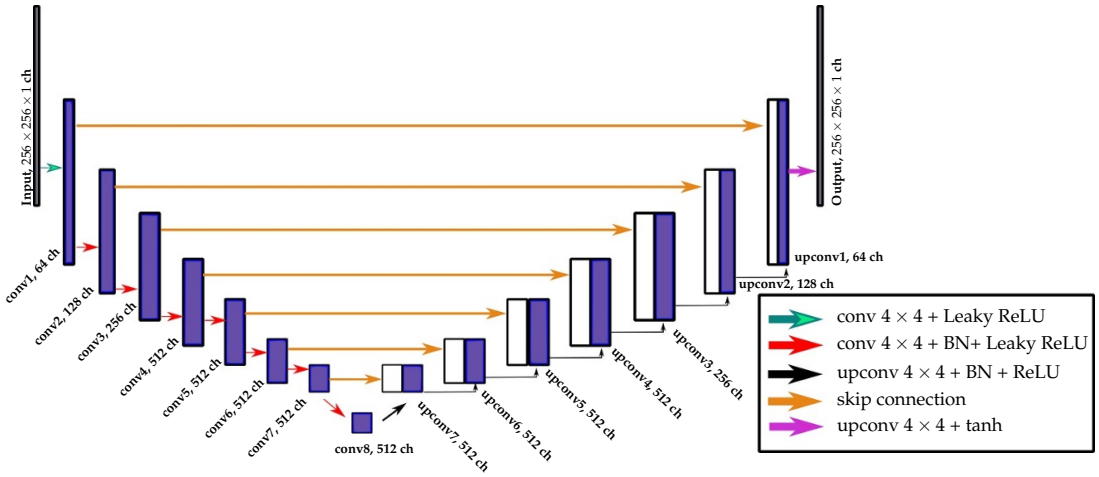


Figure 1. Schematic overview of the *UNet* architecture used adapted from Bittner et al. [43]. Each convolution operation has a kernel of size 4 × 4 with stride 2. For up-sampling, the transposed convolution operations with kernels of size 4 × 4 and stride 2 are used. The Leaky ReLU activation function in the encoder part of the network has a negative slope of 0.2.

Finally, we adapted the recently published *DeepLabv3+* [55] architecture to our multiple output *G* network. We used a re-implementation of this architecture with a pre-trained ResNet101. It utilized a ResNet as a feature extractor to provide rich semantic information and used *à trous spatial pyramid pooling (ASPP)* [56] to preserve the spatial resolution at different rates. Thus, it provided the possibility to refine the segmentation results, especially along object boundaries. The advantage of using à trous convolutions is that they allow one to expand the receptive field of filters to incorporate a larger context without increasing the number of weights. As a result, it offers an efficient mechanism to control the field-of-view and finds the best trade-off between accurate localization (small field-of-view) and context relation (large field-of-view).

The schematic representations of the proposed architectures are depicted in Figure 2a.

3.1.2. Two Generators, Two Outputs

Depth regression and semantic segmentation representations are not the same, but can complement each other. Therefore, training only one common *G* network for different problems is critical, especially because depth regression is a more complicated task and can badly influence the final building outline results, while the segmentation has to follow the pattern of the building structure. Moreover, the intermediate features of 3D representations are different from 2D. We aim at improving the roof forms and building shapes, specifically along the building borders, by integrating the information about building outlines and roof types into the system. We propose a cGAN model with two generators G_1 and G_2 responsible for better-quality DSM generation and building roof type pixel-wise classification mask production, respectively. At the beginning of the proposed architecture, two generators G_1 and G_2 are joined through two 1 × 1 convolutional layers with 8 and 32 channels, respectively. Coupling two tasks into a single model ensures that the model agrees between the independent task outputs while reducing computation time [57]. The schematic representations of the proposed architectures are depicted in Figure 2b.

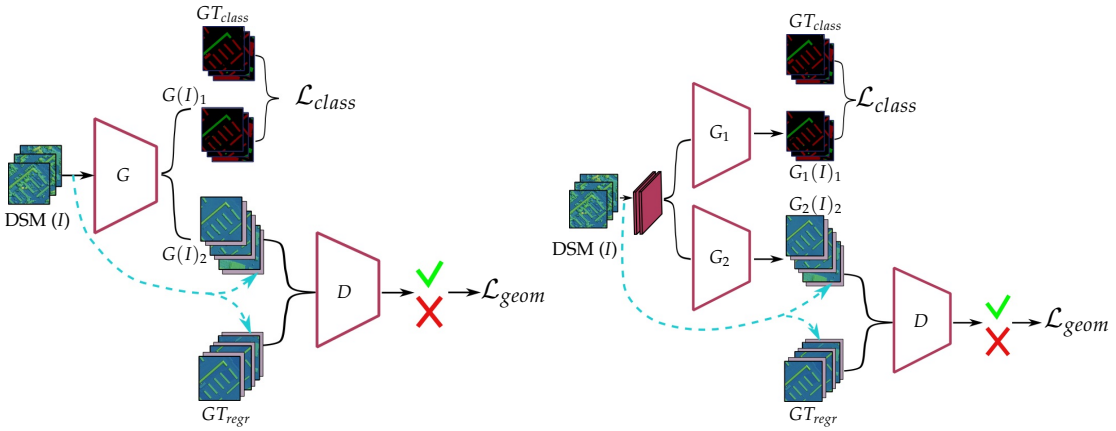


Figure 2. Schematic overview of two investigated architectures with (**a**) a one-stream generator G and (**b**) a two-stream generator G_1 and G_2 for simultaneous building shape refinement $G(I)_2$ and roof classification map $G(I)_1$ generation. The input to both networks is a single photogrammetric *digital surface model (DSM)* (I). The discriminator D is identical for both models. The ground truth for the regression task (GT_{regr}) is represented by *level of detail (LoD)*2-DSM generated from *City Geography Markup Language (CityGML)* data. The ground truth for classification task (GT_{class}) is obtained from the orientation of the computed slope for each pixel. Each architecture is a *conditional generative adversarial network (cGAN)* network which conditions (- - - -) the model on side information such as input photogrammetric DSM. It is concatenated with either generated depth image $G(I)_2$ or ground truth (GT_{regr}) as an additional channel (■) before going to the D network. Although the multi-task problems $G(I)_2$ and $G(I)_1$ of the two-stream network are depicted as independent networks, in reality, they are connected through the two 1×1 convolutional layers (■) with 8 and 32 channels, respectively. As a result, the joint loss function, which sums losses responsible for geometry reconstruction ($\mathcal{L}_{geom}$) and classification ($\mathcal{L}_{class}$), propagates back through the task-dependent layers, as well as the shared ones.

In this work, we investigate the constellations of (a) G_1: UNet and G_2: ResNet and (b) G_1: UNet and G_2: DeepLabv3+, from which the description is given in Section 3.1.1. The reason for these combinations is that the UNet, in general, produces better 3D building representations but ResNet, and based on it DeepLabv3+, gives more accurate and complete semantic segmentation maps.

3.2. Loss Function

In general, the training of cGANs can be described as a two-player minimax game:

$$\min_G \max_D \mathcal{L}_{\text{cGAN}}(G, D) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p_{\text{real}}(\boldsymbol{y})}[\log D(\boldsymbol{y}|\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z}|\boldsymbol{x})|\boldsymbol{x}))], \quad (1)$$

where $\mathbb{E}[\cdot]$ denotes the expectation value. The $G(\boldsymbol{z}|\boldsymbol{x})$ learns to fool the D by synthesizing real-looking images from a latent vector $\boldsymbol{z} \sim p_z(\cdot)$ drawn from a distribution $p_z(\cdot)$ by mapping them to an element $\boldsymbol{y} \sim p_{\text{real}}(\cdot)$ sampled from p_{real}. Differently, $D(\boldsymbol{y}|\boldsymbol{x})$ is realized as a binary classification network that attempts to differentiate between the generated data $\boldsymbol{y}$ (Class 0) and the real sample $\boldsymbol{y}^*$ (Class 1). The similarity of the generated image to the data we are interested in imitating is controlled by the information from the input image $\boldsymbol{x}$ sent to $G(\boldsymbol{z}|\boldsymbol{x})$, as well as to $D(\boldsymbol{y}|\boldsymbol{x})$.

We have already investigated that the technique proposed by Mao et al. [58] of replacing the negative log likelihood in Equation (1) by a least squares loss L_2 leads to higher stability of the training and the generation of more accurate results [43]. This transformation leads to a *conditional least squares generative adversarial network (cLSGAN)* objective function:

$$\min_G \max_D \mathcal{L}_{\text{cLSGAN}}(G, D) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p_{\text{real}}(\boldsymbol{y})}[(D(\boldsymbol{y}|\boldsymbol{x}) - 1)^2] + \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}\sim p_z(\boldsymbol{z})}[D(G(\boldsymbol{z}|\boldsymbol{x})|\boldsymbol{x})^2]. \quad (2)$$

Moreover, to help the generator generate the synthesized image close to the given ground-truth, it is typical to combine the *generative adversarial network (GAN)* objective with losses, such as L_1 or L_2 distances. In this work, we are interested in creating images that feature 3D building shapes with sharp ridge lines, as well as steep walls. Therefore, the L_1 distance:

$$\mathcal{L}_{L_1}(G) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p_{\text{real}}(\boldsymbol{y}),\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\|\boldsymbol{y} - G(\boldsymbol{z}|\boldsymbol{x})|\boldsymbol{x}\|_1], \tag{3}$$

is used to encourage less blurring.

To further refine the surface of roof planes, we consider a normal vector loss:

$$\mathcal{L}_{\text{normal}}(\mathcal{N}^{\text{t}},\mathcal{N}^{\text{p}}) = \frac{1}{m}\sum_{i=1}^{m}\left(1 - \frac{\left\langle \boldsymbol{n}_i^{\text{t}}, \boldsymbol{n}_i^{\text{p}} \right\rangle}{\left\|\boldsymbol{n}_i^{\text{t}}\right\| \left\|\boldsymbol{n}_i^{\text{p}}\right\|}\right), \tag{4}$$

as proposed by Hu et al. [59], for the training to measure the angle between the normal to the surface of an estimated DSM with respect to a target DSM. Here, $\mathcal{N}^{\text{t}} = \{\boldsymbol{n}_1^{\text{t}}, \ldots, \boldsymbol{n}_m^{\text{t}}\}$ and $\mathcal{N}^{\text{p}} = \{\boldsymbol{n}_1^{\text{p}}, \ldots, \boldsymbol{n}_m^{\text{p}}\}$ are normal vectors of the target and predicted DSMs, respectively, and $\langle \cdot, \cdot \rangle$ denotes the scalar product of two vectors. The loss is computed only within the building segments using an available binary building mask.

To learn pixel-wise roof type class probabilities, we use the cross-entropy loss function:

$$\mathcal{L}_{\text{CE}}(\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{p}) = -\sum_i t_i \log p(x_i) \tag{5}$$

paired with $softmax(z) = \frac{e^{z_k}}{\sum_j e^{z_j}}$ normalization applied to the neural network outputs z_k before the cross-entropy loss computation. Here, $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is the set of input examples in the training dataset and $\boldsymbol{t} = \{t_1, \ldots, t_n\}$ is the corresponding set of ground-truth values for the input examples.

To train our multi-task jointly, we sum in a weighted linear manner the losses of each individual task together with losses, responsible for image appearance refinement. This leads to our final combined objective function:

$$G^\star = \arg\min_G \max_D \underbrace{\mathcal{L}_{\text{cLSGAN}}(G,D) + \lambda \cdot \mathcal{L}_{L_1}(G) + \beta \cdot \mathcal{L}_{\text{normal}}}_{\mathcal{L}_{geom}} + \underbrace{\gamma \cdot \mathcal{L}_{\text{CE}}(G)}_{\mathcal{L}_{class}}, \tag{6}$$

where $\lambda, \beta, \gamma \in \mathbb{R}$ are the balancing hyper-parameters.

4. Study Area and Model Settings

4.1. Dataset

We performed experiments on a DSM with a resolution of 0.5 m generated from six panchromatic Worldview-1 images using *semi-global matching (SGM)* [60]. The panchromatic images were acquired on two different days. These data were considered as the input for our model. It shows the city of Berlin, Germany, with a total area of 410 km^2. As the ground-truth for the DSM generation task, a so-called LoD2-DSM, produced from the CityGML data model, was used. The data are freely available from the download portal Berlin 3D (http://www.businesslocationcenter.de/downloadportal). Following the procedure from our previous work [43] for LoD2-DSM creation, we selected only roof polygons from the CityGML data model and triangulated them using the algorithm introduced by Shewchuk [61] based on Delaunay triangulation [62]. For raster height image generation, we calculated a unique height value of pixels that lay inside each triangle using barycentric interpolation and filled outside pixels with a *digital terrain model (DTM)*. To generate the ground-truth for the pixel-wise classification task, we used the obtained LoD2-DSM to compute the slope for each pixel within the whole image as the maximum rate of change of elevation between that pixel and its surroundings. Then, the aspect

was defined as the orientation of the computed slope, which was measured clockwise in degrees from 0–360, where 0 is north-facing, 90 is east-facing, 180 is south-facing, and 270 is west-facing. The area that did not correspond to buildings was set to Class 0, background, and 90 degrees to Class 1, flat roofs, and the rest of the degree values were set to Class 2, sloped roofs.

4.2. Implementation Details and Training

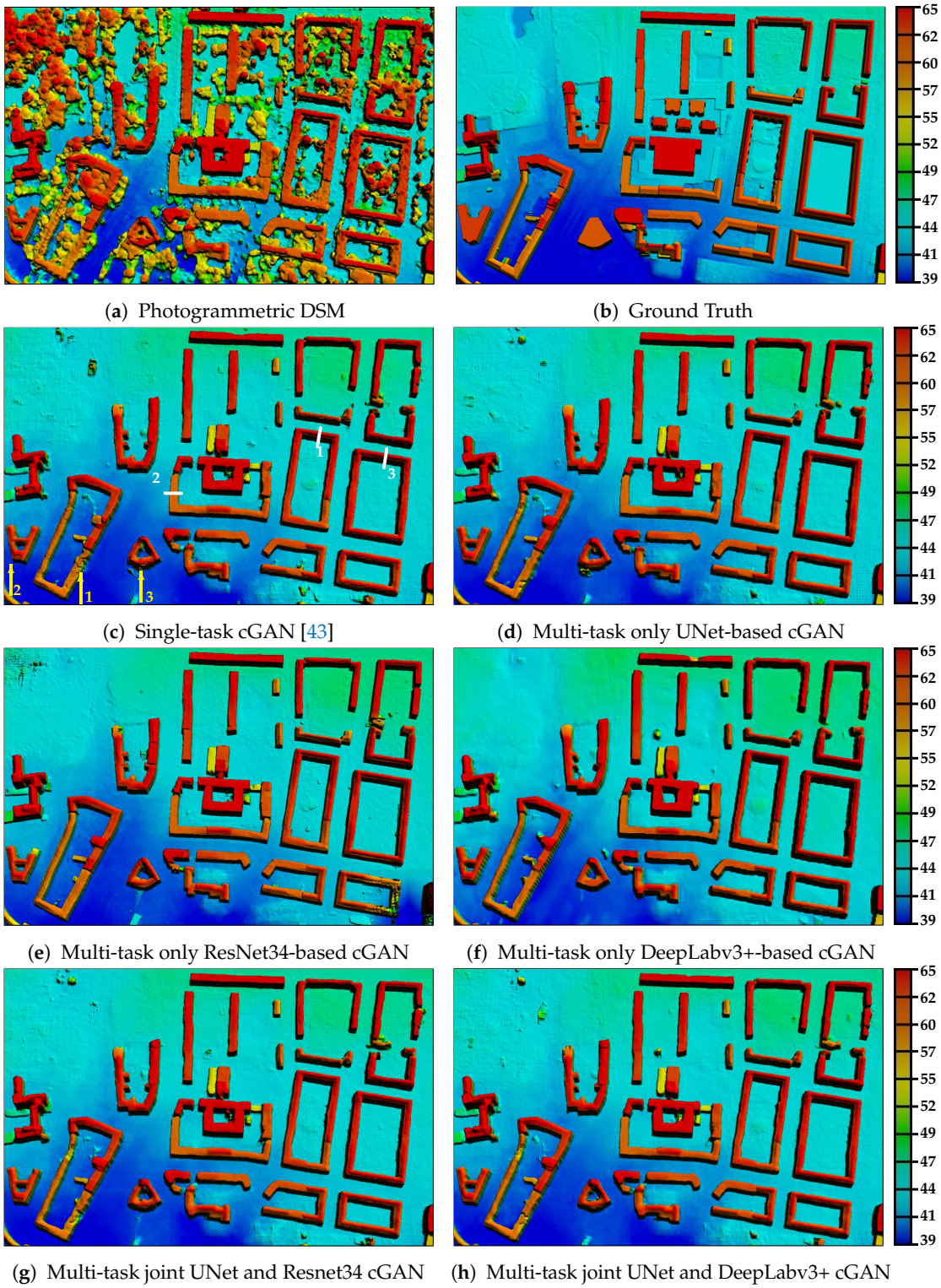
Our multi-task cGAN implementation was developed on the *PyTorch* Python package based on the *pix2pix* software introduced by Isola et al. [18]. The required training data consisted of satellite images tiled on-the-fly into 21,480 pairs of patches of size 256 × 256 px covering an area of 353 km^2. The selected patch size was appropriate to capture the buildings of different sizes together with their surroundings. For data augmentation, we used the procedure of random cropping of the tiles up to one tile size. This means that in one epoch, the network may observe, for example, only some parts of a building within one patch for one cropping and in the next epoch the whole building. This strategy makes the model more robust to input perturbations. The training of all proposed networks within the scope of this paper was done by minibatch *stochastic gradient descent (SGD)* using the Adam optimizer [63] with an initial learning rate of $\alpha = 0.0002$ and momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The total number of epochs was set to 200 with a batch size of 5 on a single NVIDIA TITAN X (PASCAL) GPU with 12 GB of memory. To estimate prediction errors for model selection, a held-out validation set covering 6 km^2 was used.

4.3. Inference Phase

During the inference phase, only the trained generators G_1, responsible for DSM refinement, and G_2, responsible for pixel-wise building roof type classification map generation, were engaged. The test area covered 50 km^2 and was unseen while training the model. The final prediction image of the test area was generated by stitching the patches of size 256 × 256 px with a fixed overlap of 128 px in the horizontal and vertical directions.

5. Results and Discussion

In this work, we performed several experiments on simultaneous depth image generation with good-quality building shapes and pixel-wise building roof type classification map extraction. First of all, comparing a single-task result with multi-task results in Figure 3, we can see that the integration of the semantic segmentation task, even for the single-stream network, already improved the results, although we directly produced two outputs from the single-stream network illustrated in Figure 2b. One of the examples highlighted by the number "1" in Figure 3c did not have the complete structure by using a single output network from Bittner et al. [43]. On the other hand, the results obtained by multi-task networks (see Figure 3d–h) were able to further improve its shape. Moreover, it can be noticed that A-shaped Building "2" and Building "3", highlighted in Figure 3c, were also more accurate and very close to the corresponding building in the ground truth.



(**a**) Photogrammetric DSM (**b**) Ground Truth

(**c**) Single-task cGAN [43] (**d**) Multi-task only UNet-based cGAN

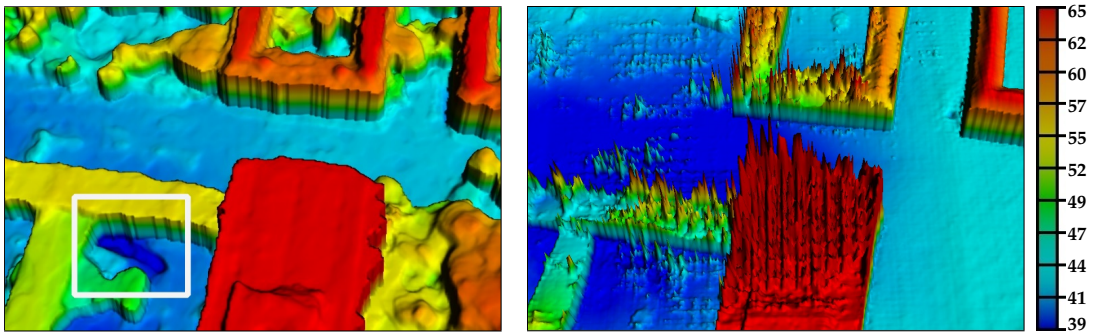
(**e**) Multi-task only ResNet34-based cGAN (**f**) Multi-task only DeepLabv3+-based cGAN

(**g**) Multi-task joint UNet and Resnet34 cGAN (**h**) Multi-task joint UNet and DeepLabv3+ cGAN

Figure 3. Visual comparison of DSMs over selected urban areas, generated by a cGAN with least squares residuals using (**c**) the one-stream generator network for a single task [43], (**d**) the one-stream generator based on the UNet network for multiple tasks, (**e**) the one-stream generator based on ResNet34 network for multiple tasks, (**f**) the one-stream generator based on the DeepLabv3+ network for multiple tasks, (**g**) the two-stream generator network with jointly trained UNet and ResNet34 architectures for multiple tasks, and (**h**) the two-stream generator network with jointly trained UNet and DeepLab architectures for multiple tasks. The DSMs images are color-shaded for better visualization.

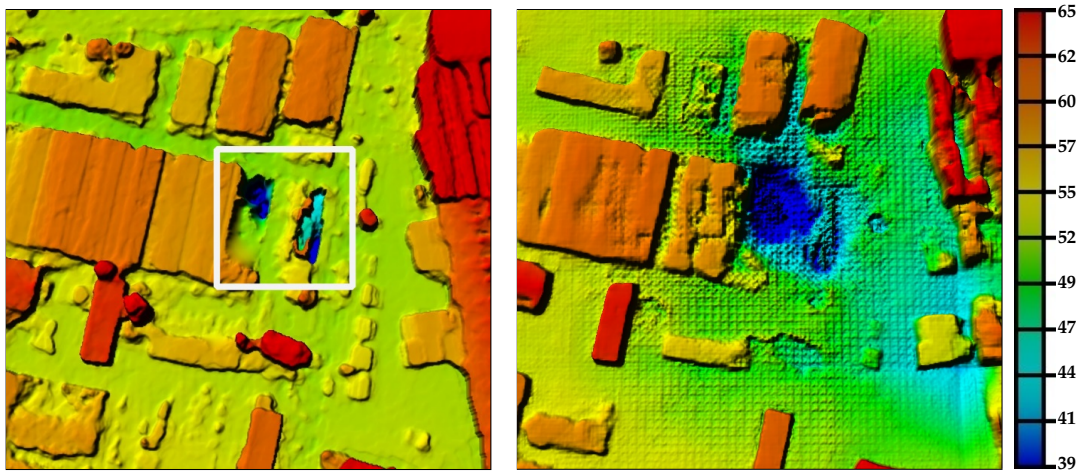
The ResNet34-based network showed, at first sight, better results (see Figure 3e). The outlines of the buildings were more rectilinear, and the ridge lines were more distinguishable. However, one can notice the inability of the model to reconstruct the building in the lower-right corner correctly. This is due to the incorrect height information presented in the input photogrammetric DSM. In the detailed view illustrated in Figure 4a, the highlighted area depicts a recess in the ground. This incorrectly reconstructed part of photogrammetric DSM happened usually due to occlusion of this area by the building walls or trees. As a result, while reconstructing this area with the ResNet34-based network, this error propagated within network layers, as the receptive field grew, and influenced the reconstruction of the whole patch. This can be identified by the dark blue area in Figure 4b. The same phenomenon happened with other architectures as well, but with less strength. The examples are depicted in Figure 5. All generated results underwent the failure influence highlighted in Figure 5a. However, the UNet in Figure 5b and the DeepLabv3+ in Figure 5d generated better results compared to the ResNet34 in Figure 5c. The propagation of incorrect values is less intensive and wide and, in the case of DeepLabv3+, even narrower. The explanation could be the presence of long skip connections in both cases compared to ResNet34, where only local residual connections were built. Those long skip connections sent more detailed information from earlier layers to the top ones in the decoder, helping to compensate the incorrect reconstruction propagated from the failure region. Besides, there were not many such failure regions within the given photogrammetric DSM. Therefore, the networks also had not the possibility to learn how to manage this problem properly. A further study needs to be dedicated to solving the problem.

Examining the profiles in Figure 6g–i of three selected buildings (highlighted with white color in Figure 3c), we can see that the roof plane reconstruction results were far from acceptable compared to the ones produced by UNet-based architectures. We can conclude that for such complicated tasks as the 3D reconstruction of tiny objects from satellite images, compared to the big size of objects in media images commonly used in the computer vision field, the skip connections were needed. The skip connections carried detailed and fine information from lower layers, which added to the up-sampled feature maps and helped to refine the final output [34].

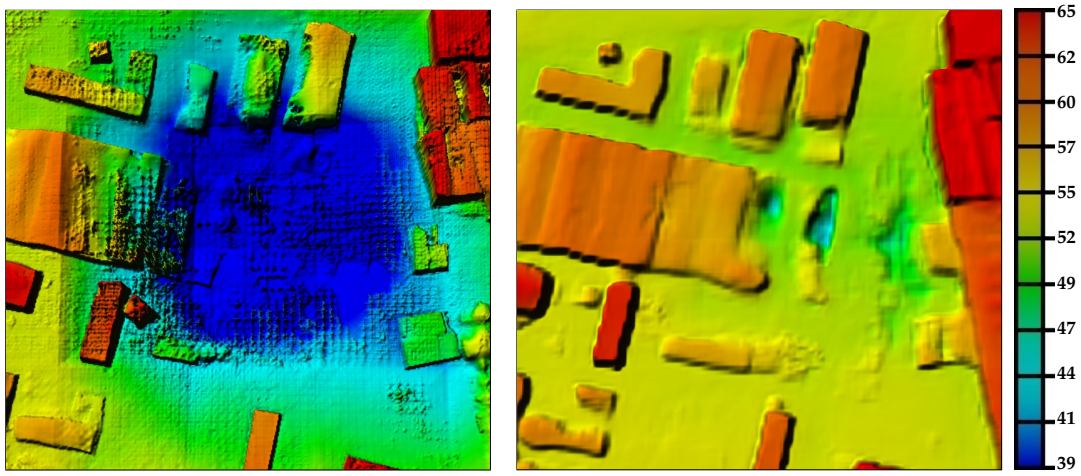


(**a**) Photogrammetric DSM (**b**) Multi-task only ResNet34-based cGAN

Figure 4. A detailed demonstration of a failure case on the generated LoD2-like DSM obtained by the ResNet34-based network Figure 2a architecture. (**a**) depicts the input photogrammetric DSM, and (**b**) shows the resulted ResNet34-based DSM from Figure 3e.



(**a**) Photogrammetric DSM (**b**) Multi-task only UNet-based cGAN



(**c**) Multi-task only ResNet34-based cGAN (**d**) Multi-task only DeepLabv3+-based cGAN

Figure 5. A detailed demonstration of a failure case example on generated LoD2-like DSMs obtained by the UNet-, ResNet34-, and DeepLabv3+-based network Figure 2a architectures. (**a**) depicts the input photogrammetric DSM with the area highlighting the presented incorrect height information and its influence on the reconstructed LoD2-like DSMs from (**b**) multi-task only UNet-based cGAN, (**c**) multi-task only ResNet-based cGAN, and (**d**) multi-task only DeepLabv3+-based cGAN. The area that undergoes the influence is presented as a darker blue shade around the location where the failure is originated in (**a**).

Investigating Figure 3f, we can say that the model based on the DeepLabv3+ architecture was able to generate depth maps with reasonable building forms regarding building boundaries. However, one also can notice that the walls of the buildings looked different compared to the results obtained from the rest of the models. Going deeper and examining the profiles of selected buildings in Figure 6j–l, we can see that these buildings had a smooth transition from roof to ground, similar to a Gaussian form, compared to the DSMs generated by other networks. This effect can also be seen in Figure 7f. The reason behind this is the last bilinear up-sampling layer with a factor of four in the decoder, which smoothed the results and was not able to learn such a complicated task like elevation model generation.

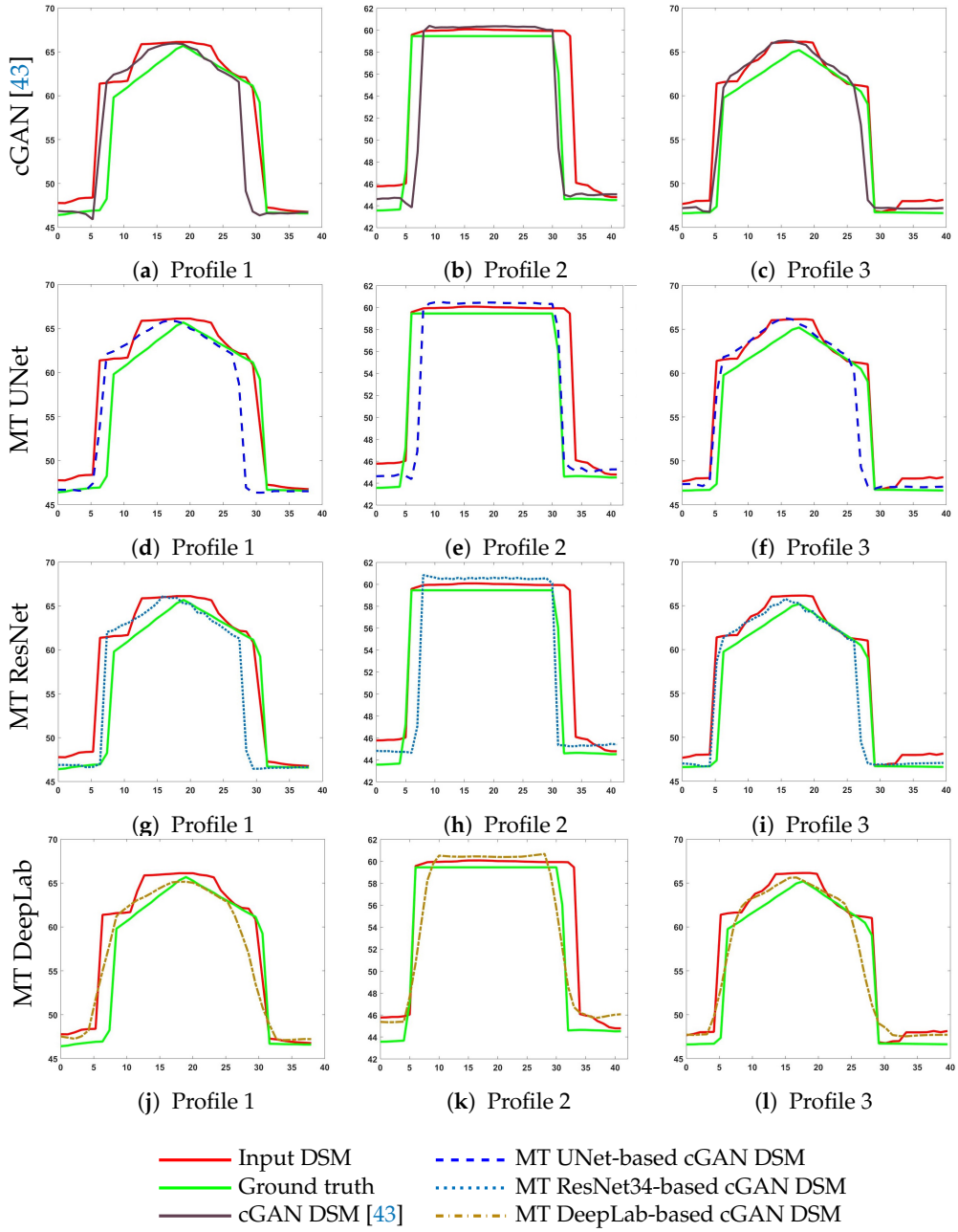


Figure 6. Illustration of the profiles for three selected buildings (cf. Figure 3c) from DSMs generated by (**a**–**c**) the cGAN model [43], (**d**–**f**) the multi-task only UNet-based cGAN, (**g**–**i**) the multi-task only ResNet34-based cGAN, and (**j**–**l**) the multi-task only DeepLabv3+-based cGAN. The results from the second, third, and fourth lines are generated by a one-generator, two-output network, depicted in Figure 2a. MT, multi-task.

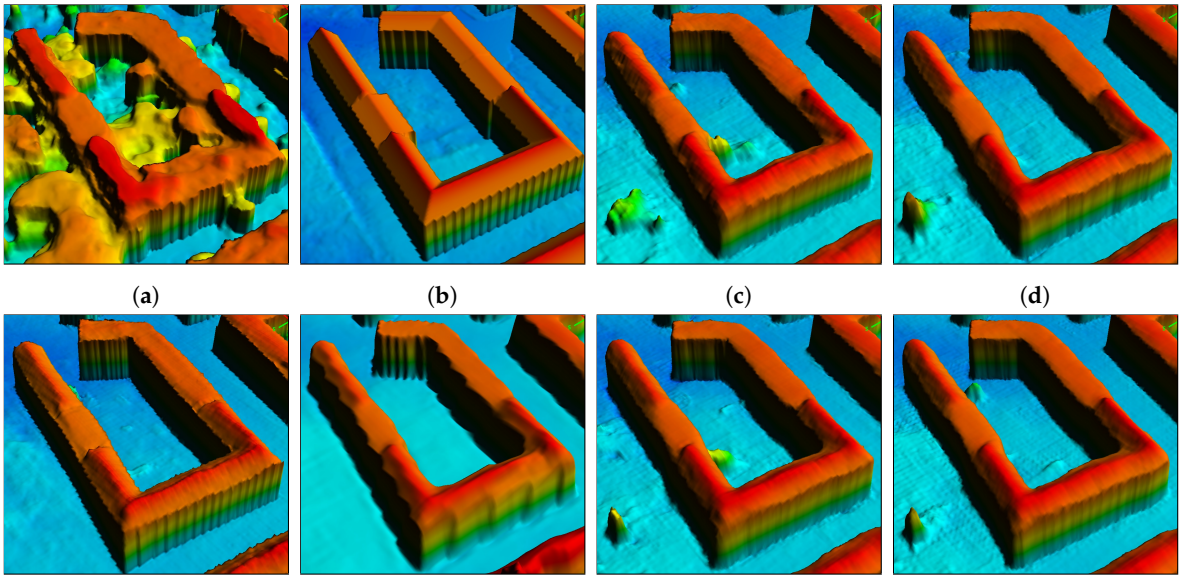


Figure 7. Comparison of the generalization over DSMs from (**c**) the one-stream generator network for a single task [43], (**d**) the one-stream generator based on the UNet network for multiple tasks, (**e**) the one-stream generator based on the ResNet34 network for multiple tasks, (**f**) the one-stream generator based on the DeepLabv3+ network for multiple tasks, (**g**) the two-stream generator network jointly trained UNet and ResNet34 architectures for multiple tasks, and (**h**) the two-stream generator network jointly trained UNet and DeepLabv3+ architectures for multiple tasks. (**a**) illustrates the input photogrammetric DSM, and (**b**) demonstrates the ground truth data.

Adding the normal vector loss function to the model helps in further refining the roof surfaces, making the roof planes flatter and more realistic. This beneficial effect can be seen comparing the profiles of selected buildings from the cGAN DSM [43] with profiles from the presented multi-task cGAN models. Moreover, the zoom-in view in Figure 7 confirms this statement, as the roof planes look smoother, but at the same time keep the right form and sharp ridge lines. This is reasonable, as we push the models to learn roof surface representations where the normal vectors, which belong to the same plane, look in the same direction and close to the ground truth.

From visual comparison between the ground truth and results from all proposed networks, one can also conclude that the networks did not generate new buildings where no buildings were placed on the low-quality input image, but rather tried to improve the available ones, even though during the training, some inconsistency between existing and no longer existingbuildings occurred. A good example can be seen in Figure 3b, where five small buildings in the middle of the scene were located. One can notice that in Figure 3a, only two different buildings are existing. This problem is not unique and can be faced in many cases during the training due to the time difference between the given photogrammetric DSM and the ground truth LoD2-DSM. However, our models were robust to such differences and did not produce "ghost" constructions.

The results of the roof classification task, simultaneously obtained in parallel with good-quality LoD2-like DSM generation, are presented in Figures 8 and 9. In the first region in Figure 8, we can see that masks, generated by ResNet34 and DeepLabv3+, provided better results for building boundaries, as well as for class separation compared to UNet-based results. Analyzing the ResNet34 and DeepLabv3+ results, one can observe that the DeepLabv3+ network provided more accurate classification. Good examples are Buildings "1", "2", "3", "4" depicted in Figure 8f. The ResNet34 model was able only to classify partially the building roof correctly, while the DeepLabv3+ set the right classes to them. Moreover, investigating the buildings highlighted as "2" and "3" in Figure 8f, one can conclude that the separate training of task-specific problems from some point in the network

positively influenced the final results compared to the multi-task network, which had a common body for several task-specific outputs (see Figure 8b–d). A larger difference showed up when investigating the residential area with small single-family houses. Looking at Figure 9, we can see that more small buildings were extracted by the DeepLabv3+ network, compared to other networks. This observation confirms that the combination of short and long skip connections together with the à trous convolution at multiple scales within the DeepLabv3+ architecture positively influenced the small objects' extraction.

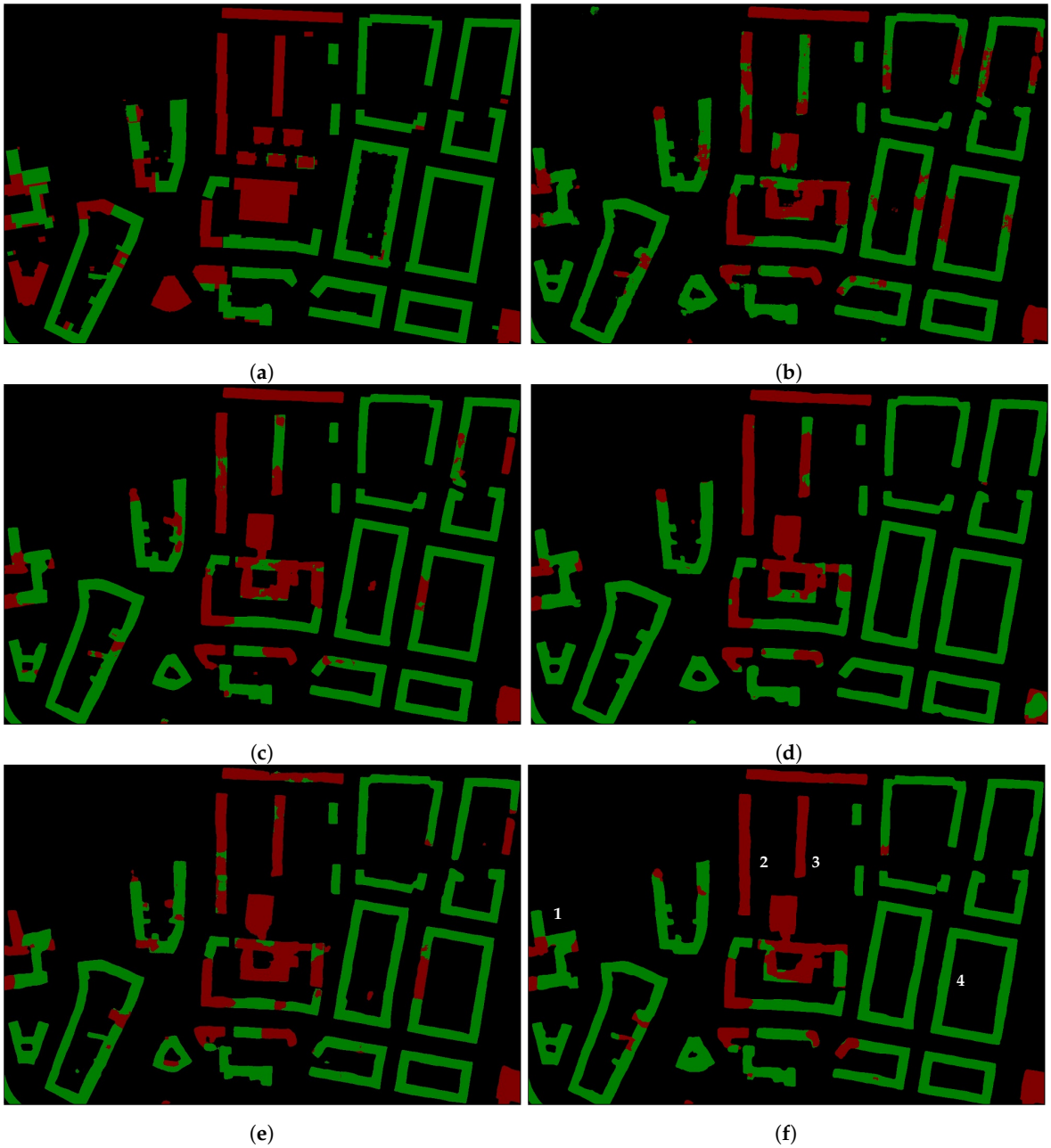


Figure 8. Visual comparison of roof classification maps over selected urban areas, generated by cGAN with least squares residuals using (**b**) the one-stream generator based on the UNet network for multiple tasks, (**c**) the one-stream generator based on the ResNet34 network for multiple tasks, (**d**) the one-stream generator based on the DeepLabv3+ network for multiple tasks, (**e**) the two-stream generator network jointly trained UNet and ResNet34 architectures for multiple tasks, (**f**) the two-stream generator network jointly trained UNet and DeepLab architectures for multiple tasks. (**a**) illustrates the ground truth label mask.

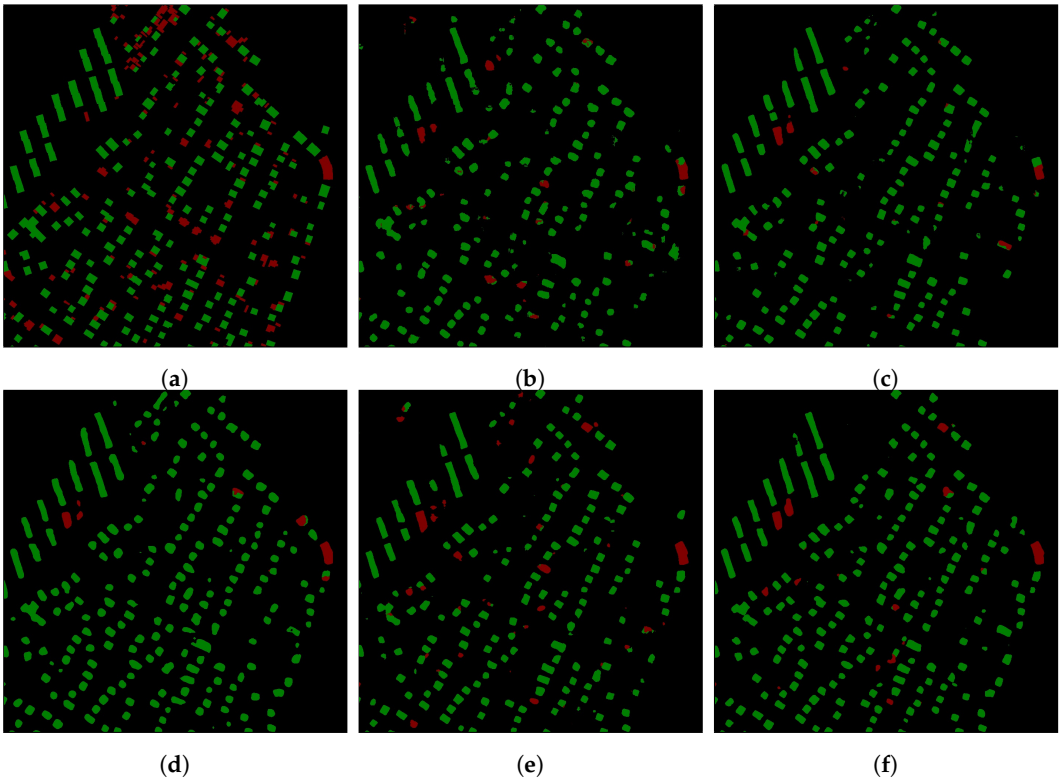


Figure 9. Visual comparison of roof classification maps over selected urban areas, generated by cGAN with least squares residuals using (**b**) the one-stream generator based on the UNet network for multiple tasks, (**c**) the one-stream generator based on the ResNet34 network for multiple tasks, (**d**) the one-stream generator based on the DeepLabv3+ network for multiple tasks, (**e**) the two-stream generator network jointly trained UNet and ResNet34 architectures for multiple tasks, and (**f**) the two-stream generator network jointly trained UNet and DeepLab architectures for multiple tasks. (**a**) depicts ground truth label mask.

Besides visual inspections of refined depth images and pixel-wise classification maps, we investigated the following error metrics commonly used in relevant publications [64–67]. Namely, for depth map evaluation, we used the *root mean squared error (RMSE)*:

$$\varepsilon_{\mathrm{RMSE}}(\boldsymbol{h}, \hat{\boldsymbol{h}}) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\hat{h}_j - h_j)^2}, \tag{7}$$

the *normalized median absolute deviation (NMAD)*

$$\varepsilon_{\mathrm{NMAD}}(\boldsymbol{h}, \hat{\boldsymbol{h}}) = 1.4826 \cdot \underset{j}{\mathrm{median}}(|\Delta h_j - m_{\Delta \boldsymbol{h}}|), \tag{8}$$

and the *mean absolute error (MAE)*

$$\varepsilon_{\mathrm{MAE}}(\boldsymbol{h}, \hat{\boldsymbol{h}}) = \frac{1}{n}\sum_{j=1}^{n}|\hat{h}_j - h_j|, \tag{9}$$

where $\boldsymbol{h} = (h_j)_j$ and $\hat{\boldsymbol{h}} = (\hat{h}_j)_j, 1 \leq j \leq n$, define the actually observed and the predicted heights, respectively, height errors are denoted as Δh_j, and the median of height errors is $m_{\Delta \boldsymbol{h}}$. The constant 1.4826, introduced in the NMAD metric, is proportional to the standard deviation if the data errors are distributed normally and, as a result, more robust to outliers in the dataset [64]. Moreover, to

exclude the influence of time acquisition difference between the input photogrammetric DSMs and the reference CityGML data model, leading to the absence or the appearance of new buildings, we manually checked the evaluation region shown in Figure 3 in this regard and evaluated the metrics on buildings showing in both the photogrammetric DSM, as well as in the reference LoD2-DSM.

For the semantic segmentation task evaluation, we used the common *intersection over union (IoU)*:

$$IoU = \frac{target \cap prediction}{target \cup prediction} \tag{10}$$

the *Precision*

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

the *Recall*

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

and the *F1-score*

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{13}$$

The IoU metric measures how much overlap is existing between two regions. It is calculated for each class separately and then averaged over all classes to provide a global statistic. Precision answers the question about the correct proportion of positive identifications. Recall shows how well all the positives are found. For the semantic segmentation task, we made an evaluation over the whole test area.

The evaluation results for depth map generation and roof classification tasks are presented in Tables 1 and 2, respectively. In terms of $\varepsilon_{\mathrm{RMSE}}$, the DeepLab-based cGAN network showed the best result. As we have already seen from the profiles in Figure 6j–l, the surfaces of the roofs presented the smoothest results compared to other profiles. Moreover, from the semantic segmentation results shown in Table 2, the DeepLab-based network showed the best results in terms of classification and outlines of buildings. As a result, those facts were reflected in the RMSE error. However, due to the over-smoothed effect of building walls, we did not consider the DeepLab model for DSMs generation tasks, but used it for the roof classification problem.

Table 1. Quantitative results for the RMSE, NMAD, and MAE metrics evaluated on 12 selected buildings existing for both input photogrammetric DSM and ground truth LoD2-DSM of the area depicted in Figure 3.

Method	Error		
	RMSE (m)	**NMAD (m)**	**MAE (m)**
cGAN [43]	3.29	0.88	1.78
only UNet based	3.20	0.91	1.71
only ResNet34 based	3.23	0.96	1.71
only DeepLabv3+ based	2.51	1.07	1.51
joint UNet and ResNet34	3.21	0.89	1.72
joint UNet and DeepLabv3+	3.12	0.90	1.69

Regarding $\varepsilon_{\mathrm{NMAD}}$ metric, one can notice that the evaluation result had the lowest value of 0.88 m and was the same for both the cGAN [43] and the joint UNet and the DeepLabv3+ network, although the $\varepsilon_{\mathrm{RMSE}}$ was different. This can be due to selecting only the median height error value, which did

not reflect the true error distribution. It can be clearly observed from the qualitative results shown in Figures 3h, 7h and 10f that the proposed joint UNet and DeepLabv3+ network can offer significantly improved roof surface qualities with noticeably smoother planes than that provided by the other tested methods. This observation suggests the need to develop new evaluation metrics that can assess the roof surface planarity better than the existing metrics.

In general, from both qualitative and quantitative evaluation, we can conclude that an improved building boundary reconstruction together with correct class label assignment were positively influencing the whole elevation model generation. This is also visually confirmed by investigating the profiles from joint UNet and DeepLab network illustrated in Figure 10d–f.

Table 2. Quantitative results for the IoU, F1-score, precision, and recall metrics evaluated on the test area covering 50 km^2.

Method	Error			
	IoU (%)	**F1-Score (%)**	**Precision (%)**	**Recall (%)**
only UNet based	59.78	72.07	77.05	48.43
only ResNet34 based	61.05	73.28	79.55	51.64
only DeepLabv3+ based	62.73	74.83	78.59	52.18
joint UNet and ResNet	61.54	73.73	79.28	51.80
joint UNet and DeepLabv3+	**64.44**	**76.34**	**80.03**	**55.2**

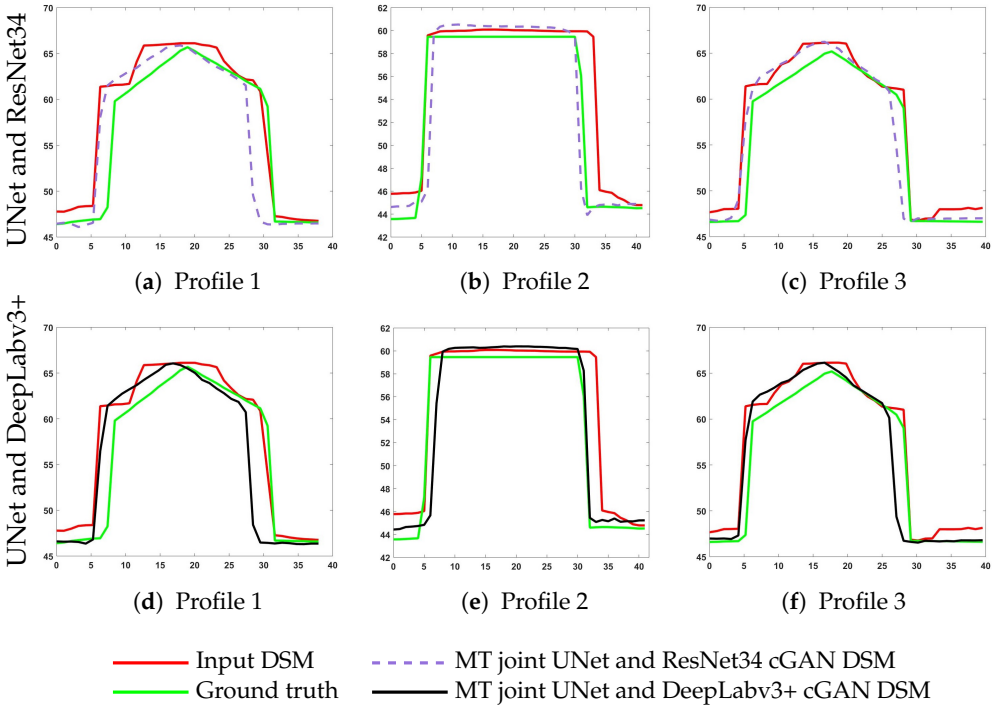


Figure 10. Illustration of the profiles for three selected buildings (cf. Figure 3c) from DSMs generated by (**a**–**c**) the two-stream generator network jointly trained UNet and ResNet34 architectures for multiple tasks and (**d**–**f**) the two-stream generator network jointly trained UNet and DeepLabv3+ architectures for multiple tasks. The results are generated by the two-generator, two-output network depicted in Figure 2b.

6. Conclusions

Good-quality DSMs with improved and complete building shapes is a very valuable data source for many remote sensing applications. Although the existing photogrammetric methodologies are

able to generate large-scale high-resolution DSMs from stereo imagery, they still exhibit irregularities and noise, due to problems like occlusion by trees, neighboring constructions, atmospheric effects, shadows, or matching errors. This leads into noisy or partly missing building structures and, as a result, requires a refinement procedure. Many attempts were made to refine 3D objects on DSMs, like buildings, by fitting models from libraries or reconstructing 3D information from detected 2D building footprints using prior knowledge about the type of roofs. However, these algorithms are not robust to the huge variety of existing building forms.

In this work, we proposed a methodology that was able to refine a huge variety of building shapes in photogrammetric DSMs automatically and, at the same time, produced improved roof classification maps. We developed a deep learning approach that utilized DSMs and consisted of two separated generators intended for multiple tasks. Although both generators were connected only at the beginning, they were able to contribute to each other for the reconstruction through joint learning and, as a result, produced more accurate results. Mainly, the height information helped to distinguish buildings among various terrestrial targets in an image much better, and roof classification maps, in turn, helped to refine building boundaries, making them more rectangular. Additionally, it is valuable information for roof forms' reconstruction as the network was able to determine this knowledge and use it for better learning. Besides, the additional vector normal loss function penalized irregularities on roof surfaces, leading to smoother roof shapes, which were closer to the ground truth.

The main limitation of the proposed network was an experimental manual tuning of balancing hyper-parameters responsible for weighting the losses of individual tasks. In the future, we aim to derive a multi-task loss function that can learn to balance various regression and classification losses automatically. Moreover, we are planning to combine two generators G_1 and G_2 of the network by sharing even more hidden layers between all tasks, while keeping only several task-specific output layers. This will overcome the problem of a big number of training parameters, due to separate architectures for each task, and should better influence the generation of both tasks, as complementary information will be learned jointly at the beginning. Besides, we want to investigate the advantages of combining short skip connection as in the ResNet network and long skip connections as in the UNet network within one single architecture to improve both 3D reconstruction and roof classification map generation.

Author Contributions: Conceptualization, K.B., F.F., M.K. and P.R.; methodology, K.B., F.F. and M.K.; software, K.B.; validation, K.B.; formal analysis, K.B., F.F.; investigation, K.B. and F.F.; writing, original draft preparation, K.B.; writing, review and editing, F.F., M.K. and P.R.; visualization, K.B.; supervision, F.F., M.K. and P.R.

Funding: This research was funded by the German Academic Exchange Service (DAAD:DLR/DAAD Research Fellowship No. 57186656) for Ksenia Bittner.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hoja, D.; Reinartz, P.; Lehner, M. DSM generation from high resolution satellite imagery using additional information contained in existing DSM. In Proceedings of the High-Resolution Earth Imaging for Geospatial Information, Hannover, Germany, 17–20 May 2005; pp. 1–6.
2. Eckert, S.; Hollands, T. Comparison of automatic DSM generation modules by processing IKONOS stereo data of an urban area. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2010**, *3*, 162–167. [CrossRef]
3. Wohlfeil, J.; Hirschmüller, H.; Piltz, B.; Börner, A.; Suppa, M. Fully automated generation of accurate digital surface models with sub-meter resolution from satellite imagery. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *XXXIX-B3*, 75–80. [CrossRef]
4. Xu, F.; Woodhouse, N.; Xu, Z.; Marr, D.; Yang, X.; Wang, Y. Blunder elimination techniques in adaptive automatic terrain extraction. *ISPRS J.* **2008**, *29*, 21.
5. Sirmacek, B.; d'Angelo, P.; Krauss, T.; Reinartz, P. Enhancing urban digital elevation models using automated computer vision techniques. In Proceedings of the ISPRS Commission VII Symposium, Vienna, Austria, 5–7 July 2010.

6. Brédif, M.; Tournaire, O.; Vallet, B.; Champion, N. Extracting polygonal building footprints from digital surface models: A fully-automatic global optimization framework. *ISPRS J. Photogramm. Remote Sens.* **2013**, *77*, 57–65. [CrossRef]
7. Davydova, K.; Cui, S.; Reinartz, P. Building footprint extraction from digital surface models using neural networks. In *Image and Signal Processing for Remote Sensing XXII*; International Society for Optics and Photonics: Edinburgh, UK, 2016; Volume 10004, p. 100040J.
8. Arefi, H.; Alizadeh Naeini, A.; Ghafouri, A. Building Extraction Using Surface Model Classification. In Proceedings of the GIS Ostrava 2013—Geoinformatics for City Transformation, Ostrava, Czech Republic, 21–23 January 2013.
9. Bittner, K.; Cui, S.; Reinartz, P. Building Extraction from Remote Sensing Data using Fully Convolutional Networks. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *42*, 481–486. [CrossRef]
10. Liao, Y.; Kodagoda, S.; Wang, Y.; Shi, L.; Liu, Y. Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; pp. 2318–2325.
11. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229.
12. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
13. Girshick, R. Fast r-cnn. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
14. Gregor, K.; Danihelka, I.; Mnih, A.; Blundell, C.; Wierstra, D. Deep autoregressive networks. *arXiv* **2013**, arXiv:1310.8499.
15. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*; Terry Sejnowski: San Diego, CA, USA, 2014; pp. 2672–2680.
16. Gregor, K.; Danihelka, I.; Graves, A.; Rezende, D.J.; Wierstra, D. Draw: A recurrent neural network for image generation. *arXiv* **2015**, arXiv:1502.04623.
17. Oord, A.V.D.; Kalchbrenner, N.; Kavukcuoglu, K. Pixel recurrent neural networks. *arXiv* **2016**, arXiv:1601.06759.
18. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. *arXiv* **2016**, arXiv:1611.07004.
19. Mohajeri, N.; Assouline, D.; Guiboud, B.; Bill, A.; Gudmundsson, A.; Scartezzini, J.L. A city-scale roof shape classification using machine learning for solar energy applications. *Renew. Energy* **2018**, *121*, 81–93. [CrossRef]
20. Assouline, D.; Mohajeri, N.; Scartezzini, J.L. Building rooftop classification using random forests for large-scale PV deployment. In Proceedings of the Earth Resources and Environmental Remote Sensing/GIS Applications VIII, Warsaw, Poland, 5 October 2017; Volume 10428, p. 1042806.
21. Castagno, J.D.; Atkins, E.M. Automatic Classification of Roof Shapes for Multicopter Emergency Landing Site Selection. *arXiv* **2018**, arXiv:1802.06274.
22. Alidoost, F.; Arefi, H. Knowledge based 3D building model recognition using convolutional neural networks from lidar and areal imageries. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *41*, 833–840. [CrossRef]
23. Partovi, T.; Fraundorfer, F.; Azimi, S.; Marmanis, D.; Reinartz, P. Roof Type Selection based on patch-based classsification using deep learning for high Resolution Satellite Imagery. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *42*, 653–657. [CrossRef]
24. Axelsson, M.; Soderman, U.; Berg, A.; Lithen, T. Roof Type Classification Using Deep Convolutional Neural Networks on Low Resolution Photogrammetric Point Clouds From Aerial Imagery. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 1293–1297.
25. Felicísimo, A.M. Parametric statistical method for error detection in digital elevation models. *ISPRS J. Photogramm. Remote Sens.* **1994**, *49*, 29–33. [CrossRef]

26. Wang, P. Applying two dimensional Kalman filtering for digital terrain modelling. *Proc. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **1998**, *32*, 649–656.
27. Walker, J.P.; Willgoose, G.R. A comparative study of Australian cartometric and photogrammetric digital elevation model accuracy. *Photogramm. Eng. Remote Sens.* **2006**, *72*, 771–779. [CrossRef]
28. Anderson, E.; Thompson, J.; Austin, R. LIDAR density and linear interpolator effects on elevation estimates. *Int. J. Remote Sens.* **2005**, *26*, 3889–3900. [CrossRef]
29. Smith, S.; Holland, D.; Longley, P. Quantifying interpolation errors in urban airborne laser scanning models. *Geograph. Anal.* **2005**, *37*, 200–224. [CrossRef]
30. Shi, W.; Tian, Y. A hybrid interpolation method for the refinement of a regular grid digital elevation model. *Int. J. Geogr. Inf. Sci.* **2006**, *20*, 53–67. [CrossRef]
31. Sirmacek, B.; d'Angelo, P.; Reinartz, P. Detecting complex building shapes in panchromatic satellite images for digital elevation model enhancement. In Proceedings of the ISPRS Workshop on Modeling of Optical Airborne and Space Borne Sensors, Istanbul, Turkey, 11–13 October 2010.
32. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; Terry Sejnowski: San Diego, CA, USA, 2012; pp. 1097–1105.
33. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
34. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
35. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
36. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [CrossRef]
37. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
38. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*; Terry Sejnowski: San Diego, CA, USA, 2014; pp. 2366–2374.
39. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2650–2658.
40. Liu, F.; Shen, C.; Lin, G. Deep convolutional neural fields for depth estimation from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5162–5170.
41. Mou, L.; Zhu, X.X. IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network. *arXiv* **2018**, arXiv:1802.10249.
42. Bittner, K.; Korner, M. Automatic large-scale 3d building shape refinement using conditional generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1887–1889.
43. Bittner, K.; d'Angelo, P.; Körner, M.; Reinartz, P. DSM-to-LoD2: Spaceborne Stereo Digital Surface Model Refinement. *Remote Sens.* **2018**, *10*, 1926. [CrossRef]
44. Collobert, R.; Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 160–167.
45. Deng, L.; Hinton, G.; Kingsbury, B. New types of deep neural network learning for speech recognition and related applications: An overview. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 8599–8603.

46. Kokkinos, I. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6129–6138.
47. Liebel, L.; Körner, M. Auxiliary tasks in multi-task learning. *arXiv* **2018**, arXiv:1805.06334.
48. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]
49. Vakalopoulou, M.; Platias, C.; Papadomanolaki, M.; Paragios, N.; Karantzalos, K. Simultaneous registration, segmentation and change detection from multisensor, multitemporal satellite image pairs. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1827–1830.
50. Srivastava, S.; Volpi, M.; Tuia, D. Joint height estimation and semantic labeling of monocular aerial images with CNNs. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 5173–5176.
51. Sener, O.; Koltun, V. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*; Terry Sejnowski: San Diego, CA, USA, 2018; pp. 527–538.
52. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
53. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Switzerland, 2015; pp. 234–241.
54. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
55. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv* **2018**, arXiv:1802.02611.
56. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
57. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7482–7491.
58. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.
59. Hu, J.; Ozay, M.; Zhang, Y.; Okatani, T. Revisiting Single Image Depth Estimation: Toward Higher Resolution Maps with Accurate Object Boundaries. *arXiv* **2018**, arXiv:1803.08673.
60. d'Angelo, P.; Reinartz, P. Semiglobal matching results on the ISPRS stereo matching benchmark. In Proceedings of the High-Resolution Earth Imaging for Geospatial Information, Hannover, Germany, 14–17 June 2011; pp. 79–84.
61. Shewchuk, J.R. Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator. In *Workshop on Applied Computational Geometry*; Springer: Berlin/Heidelberg, Germany, 1996; pp. 203–222.
62. Delaunay, B. Sur la sphere vide. *Otdelenie Matematicheskii i Estestvennyka Nauk* **1934**, *7*, 1–2.
63. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
64. Höhle, J.; Höhle, M. Accuracy assessment of digital elevation models by means of robust statistical methods. *ISPRS J. Photogramm. Remote Sens.* **2009**, *64*, 398–406. [CrossRef]
65. Zhang, J.; Zhu, T.; Tang, Y.; Zhang, W. Geostatistical approaches to refinement of digital elevation data. *Geo-Spat. Inf. Sci.* **2014**, *17*, 181–189. [CrossRef]
66. Elaksher, A.F.; Bethel, J. Refinement of digital elevation models in urban areas using breaklines via a multi-photo least squares matching algorithm. *J. Terr. Obs.* **2010**, *2*, 7.
67. Hobi, M.L.; Ginzler, C. Accuracy assessment of digital surface models based on WorldView-2 and ADS80 stereo remote sensing data. *Sensors* **2012**, *12*, 6347–6368. [CrossRef]