LONG-SHORT SKIP CONNECTIONS IN DEEP NEURAL NETWORKS FOR DSM REFINEMENT

Ksenia Bittner[1]*, Lukas Liebel[2], Marco Körner[2], Peter Reinartz[1]
[1] Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling, Germany
[2] Computer Vision Research Group, Chair of Remote Sensing Technology (LMF), Technical University of Munich (TUM), Munich, Germany

Commission II, WG II/4

KEY WORDS: Conditional generative adversarial networks (cGANs), balancing hyper-parameters, long-short skip connections, 3D scene refinement, building geometry

ABSTRACT:

Detailed *digital surface models (DSMs)* from space-borne sensors are the key to successful solutions for many remote sensing problems, like environmental disaster simulations, change detection in rural and urban areas, 3D urban modeling for city planning and management, *etc*. Traditional methodologies, *e.g.*, stereo matching, used to generate photogrammetric DSMs from stereo imagery, usually deliver low-quality results due to the matching errors in homogeneous areas or the lack of information when observing the scene under different viewing angles. This makes the tasks related to building reconstruction very challenging since in most cases it is difficult to recognize the type of roofs, especially if overlaid with trees. This work represents a continuation of research regarding the automatic optimization of building geometries in photogrammetric DSMs with half-meter resolution and introduces an improved *generative adversarial network (GAN)* architecture which allows to reconstruct complete and detailed building structures without neglecting even low-rise urban constructions. The generative part of the network is constructed in a way that it simultaneously processes height and intensity information, and combines short and long skip connections within one architecture. To improve different aspects of the surface, several loss terms are used, the contributions of which are automatically balanced during training. The obtained results demonstrate that the proposed methodology can achieve two goals without any manual intervention: improve the roof surfaces by making them more planar and also recognize and optimize even small residential buildings which are hard to detect.

1. INTRODUCTION

1.1 Problem Statement

For humans, it is usually an easy task to understand the realistic shape and appearance of real objects in an image due to accumulated experience and knowledge. For computer algorithms, on the other hand, it is not a trivial task to estimate the true 3D geometry from 2D object representations, since the information extracted from an image is very limited and usually many constraints or conditions are predefined. Therefore, it is common to combine the knowledge from different data sources to compensate for the lack of information in a single data source.

In remote sensing, for example, intensity and height information is usually paired to reconstruct building geometries which are one of the prominent objects on the ground surface. This is necessary since *digital surface models (DSMs)* generated from high-resolution satellite images with different viewing angles still feature noise, inconsistency, and sometimes non-realistic building appearances due to occlusions or errors of stereo matching algorithms. However, combining different data sources is not enough to solve the building reconstruction task since those objects vary a lot in shapes and sizes. Traditional methods incorporate different constraints, usually assuming only primitive forms, to get closer to the realistic building appearances. In this work, we propose a machine learning approach that can automatically eliminate the vegetation and refine 2.5D building geometries in elevation models after processing *pan-chromatic (PAN)* and DSM images. Elevation models with optimized building shapes can be further utilized as input for 3D city model generation. In continuation of our previous work (Bittner et al., 2019b), we introduce an improved version of the generative part of a *conditional generative adversarial network (cGAN)*-based architecture which can reconstruct better building roof structures not only of big residential and industrial buildings but also of low-rise ones. Moreover, the contributions of multiple loss terms constructing our objective function are automatically weighted, to escape their tedious manual tuning.

* Corresponding author

1.2 Related Work

There have already been several attempts to reconstruct building geometries as close as possible to their real appearances in the cities from remote sensing imagery. One way to approach the task of obtaining realistic building shapes is to improve the low-quality of DSMs, precisely, to reduce the number of outliers and inconsistencies resulting from low-textured or shadowed areas. (Felicísimo, 1994; Wang, 1998; López, 2000) propose to use statistical criteria to identify anomalous height values within neighboring pixels. Although those strategies provide more continuous surfaces, the major drawback is that they oversmooth the steepness of building walls. Some methodologies, after detecting uncorrelated points, suggest applying different interpolation techniques. Among a variety of interpolation strategies *Inverse Distance Weighting (IDW)* (Goovaerts et al., 1997), kriging (Anderson et al., 2005) and spline-based methods (Smith et al., 2005) are well known. The recent work of (Chen and Li, 2013) demonstrates that methods based on multi-quadric interpolation con-

structed as the objective functions can achieve better results. Despite interpolation-based methods accomplishments within rural areas, in case of complex urban landscapes, where discontinuities between the ground and building constructions are very strong and should be kept, the methods fail to preserve the sharpness of building edges.

With the appearance of deep learning techniques, it became possible to automatically solve regression problems and generate continuous sets of values. Apart from the computer vision field where many works have been done and a great success for generating depth images was achieved, in remote sensing, it is still under intensive research. Several studies have recently appeared at the same time. The work of (Mou and Zhu, 2018) investigates the DSMs generation from monocular aerial images using an end-to-end *Fully convolutional network (FCN)* with skip connections. The authors were able to reconstruct building shapes with a high degree of accuracy since aerial imagery has more detailed information about object boundary, but with an only relative height of buildings which does not face the reality. Like us, (Ghamisi and Yokoya, 2018) utilize *generative adversarial networks (GANs)* for DSMs generation, but as input data they use aerial images consisting of near-infrared, red and green bands. Although the approach can generate reasonable results similar to the training data, it fails to generalize over images with different spatial-spectral information. Our previous approaches (Bittner et al., 2018, 2019b) pursuit not only automatic height images creation from photogrammetric half-meter resolution satellite DSMs but also a simultaneous building shapes refinement on them involving cGANs. (Bittner et al., 2019a; Liebel et al., 2020) investigate mutli-task learning for improving the mutual information between different but correlated problems like roof type classification and building shape refinement. The work of (Bittner et al., 2019b) propose to incorporate several types of information, precisely intensity, and height, to gain both detailed roof ridge lines reconstruction and the compilation of building structure if they are badly represented in photogrammetric DSMs. Following those goals, in this work, we integrate short skip connections together with long skip connections in the improved architecture to increase the network confidence of detecting and reconstructing also low-rise buildings—the challenge which the previous network was not able to achieve. Moreover, since the objective function we aim to minimize consists of several terms, we make their balancing hyper-parameters learnable, to allow the system itself to decide which contribution is more valuable for better reconstruction. By incorporating the strategy of learnable weighting parameters for multiple terms, we aim to further improve the building forms and the planarity of their roof surfaces.

2. METHODOLOGY

2.1 Network Architecture

Our earlier network architecture (Bittner et al., 2019b) proved the effectiveness of the method to refine building geometries in DSMs. However, some notable issues, like roof surface unevenness or a lack of low-rise buildings reconstruction, remain in the prediction results. We revise this architecture with a more comprehensive GAN concept, which incorporates identity shortcut connections, also known as residual connections, together with long skip connections within the generative part of the network. The strength of residual connections is their ability to prevent the vanishing gradient problem which gives us the potential to train

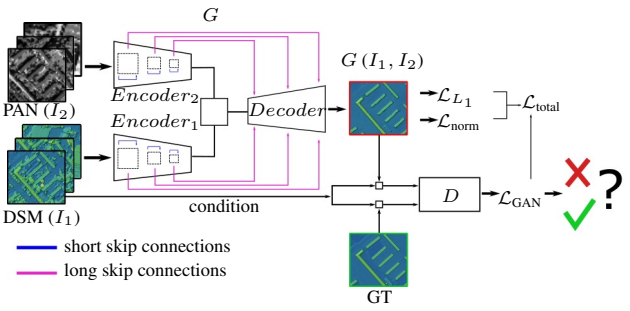


Figure 1. Schematic overview of the proposed architecture for building shape optimization in low-quality photogrammetric DSMs. The generative module G is represented by an encoder-decoder network which consists of two separate encoders, processing the depth and intensity information simultaneously, and a single decoder module.

deeper networks and, as a result, to get a better semantic understanding of the scene. The power of long skip connections is reflected in their ability to carry detailed information from lower layers to upper ones, improving the up-sampling results of the decoder lost during hierarchical down-sampling in the encoder.

The idea of a combination of long and short skip connections within one network structure is not new. (Drozdzal et al., 2016; Zhang et al., 2018) have already explored it for different semantic image segmentation tasks. Following their strategy, we investigate a so-called *ResNet-in-UNet* architecture based on a *residual network (ResNet)* as an encoder which codes back the process with five up-sampled decoder layers combined with five long skip connections resembling the U-shaped structure. This *UResNet* forms the generative model G of the basic cGAN structure introduced by (Isola et al., 2016) which builds the main approach for height image optimization in our work. The general concept of the DSM optimization approach proposed in this work is illustrated in Fig. 1.

Since we are additionally interested in height and intensity information fusion from photogrammetric DSM and PAN images, the generative part G of our GAN architecture consists of two identical encoders E_1 and E_2, each for one data modality, which are concatenated at the bottleneck of U-shape structure. The combined information propagates then through the common decoder. The long skip connections at identical stages of each encoder are concatenated together with up-sampled features at the certain stages of the decoder forming the input for the next block.

To find the most appropriate network structure for our task, we try various existing ResNet architectures as encoders: 18-layer, 34-layer, 50-layer, 101-layer, and 152-layer. The number of features of 18-layer and 34-layer ResNets used as long skip connections stay equal to the original number of features of a certain network stage. On the other hand, the number of features of 50-layer, 101-layer, and 152-layer ResNets in long skip connections (depicted in magenta color in Fig. 1) is reduced twice for each stream which carries spectral and height information, respectively, to keep the number of total parameters as small as possible for fitting the capacity of GPU and prevent the overfitting.

The discriminator D of the proposed GAN architecture consists of 5 convolutional layers with a *sigmoid* activation function $\sigma_{\text{sigm}}(z) = \frac{1}{1+e^{-z}}$ placed on the top layer. As a result, the output of the discriminator D represents the probability that the input sample either resembles the real distribution of data or the fake one.

2.2 Objective Function

One of the significant achievements in computer vision in 2014 was the introduction of GANs (Goodfellow et al., 2014)—generative models that can learn to mimic any distribution of data. The problem is framed through the adversarial manner of learning by training two sub-models pitting against each other: a *discriminative model* D which tries to determine whether input samples are real or produced by *generative model* G which in turn tries to fool the discriminator D by creating fake samples as close to reality as possible. Often, it is also required to generate fake samples with specific characteristics rather than a generic sample from an unknown noise distribution. (Mirza and Osindero, 2014) proposed cGANs which utilize some external information for restricting both the generator G in its output and the discriminator D in its expected input. In this way, a condition gives the control on modes of the data being generated.

Mathematically it is defined as following: The generator G aims to learn a mapping function from a latent vector $\boldsymbol{z} \sim p_z(\cdot)$ to data space $\boldsymbol{y}^* \sim p_{\text{real}}(\cdot)$ as $G(\boldsymbol{z}|\boldsymbol{x}) = \boldsymbol{y}$. The discriminator outputs a single scalar $D(\boldsymbol{y}|\boldsymbol{x}) \in [0, 1]$ representing the probability that $\boldsymbol{y}$ came from real data rather than generated. The generator G and the discriminator D are trained simultaneously following a two-player min-max 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 G adjusts the parameters to minimize $\log(1 - D(G(\boldsymbol{z}|\boldsymbol{x})|\boldsymbol{x}))$ and D attempts to minimize $\log D(\boldsymbol{y}|\boldsymbol{x})$.

In our earlier work (Bittner et al., 2018), it has been already examined that changing the negative log-likelihood in Eq. (1) to a least square loss L_2 leading to a *conditional least square generative adversarial network (cLSGAN)*

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

which manages to overcome the problem with instability during the training.

To adjust the objective function for achieving the main goal of this work—to generate a height image with detailed building geometries in it—the loss function in Eq. (2) is extended with two additional terms: commonly used L_1 *distance*

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

which is responsible for building ridgelines to be as sharp as possible, and *normal vector loss* term from (Bittner et al., 2019a)

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

which is responsible for the planarety of roof polygons. This normal vector loss represents the angle between the set of surface normals $\mathcal{N}^{\text{p}} = \{\boldsymbol{n}_1^{\text{p}}, \ldots, \boldsymbol{n}_m^{\text{p}}\}$ of an generated DSM and the set of surface normals $\mathcal{N}^{\text{t}} = \{\boldsymbol{n}_1^{\text{t}}, \ldots, \boldsymbol{n}_m^{\text{t}}\}$ of the reference height image. The smaller the angle, the more straightforward and plain the estimated roof surfaces.

The final objective function combines three above-described loss terms

$$\mathcal{L}_{\text{total}} = w_{\text{cLSGAN}}\mathcal{L}_{\text{cLSGAN}}(G, D) + w_{L_1}\mathcal{L}_{L_1}(G) + w_{\text{normal}}\mathcal{L}_{\text{normal}}(\mathcal{N}^{\text{t}}, \mathcal{N}^{\text{p}}), \quad (5)$$

where parameters $0 \leq \boldsymbol{w_l} \in \mathbb{R}$ are playing the role of balancing hyper-parameters between individual losses $l \in \mathcal{L}$. Below we introduce the way of finding a balance between individual losses when they play together as a team towards objective function minimization.

2.3 Finding A Balance Between Multiple Losses

The wish to automatically adjust the influence of various loss terms on the final objective function is not new. For example, (Kendall et al., 2018; Liebel and Körner, 2018) introduced the approach for balancing multiple losses for different tasks during training. Precisely, the authors proposed to learn those balancing hyper-parameters along with model parameters to find a good trade-off between multiple outputs.

Several experiments have been already done by us for obtaining multiple remote sensing tasks from deep network architecture applying this methodology (Liebel et al., 2020). Based on these experiments, we have decided to incorporate the learning of balancing hyper-parameters $\boldsymbol{w_l}$ in Eq. (5) instead of their manually tuning as it was done before by (Bittner et al., 2019a,b). Mainly, together with learning usual network parameters $\boldsymbol{\theta}$ with each iteration during the training, the balancing hyper-parameters $\boldsymbol{w_l}$ are also inserted in the optimization process with

$$w_l = \begin{cases} 0.5 \cdot \exp(-\log(\sigma_l^2)) & \text{for } \mathcal{L}_{L_1} \text{ and } \mathcal{L}_{\text{normal}} \\ \exp(-\log(\sigma_l^2)) & \text{for } \mathcal{L}_{\text{cLSGAN}} \end{cases}$$

However, in situations like $\sigma_l^2 < 1$, the loss can yield negative values. To avoid it, the regularization term $\mathcal{R}_l = 0.5 \cdot \log(\sigma_l^2)$ should be inserted to Eq. (5) as it was proposed by (Kendall et al., 2018) and supported by further investigation of (Liebel and Körner, 2018). As a result, the final loss of this work is formulated as

$$\mathcal{L}_{\text{total}} = \sum_l w_l \cdot \mathcal{L}_l + \mathcal{R}_l, \quad (6)$$

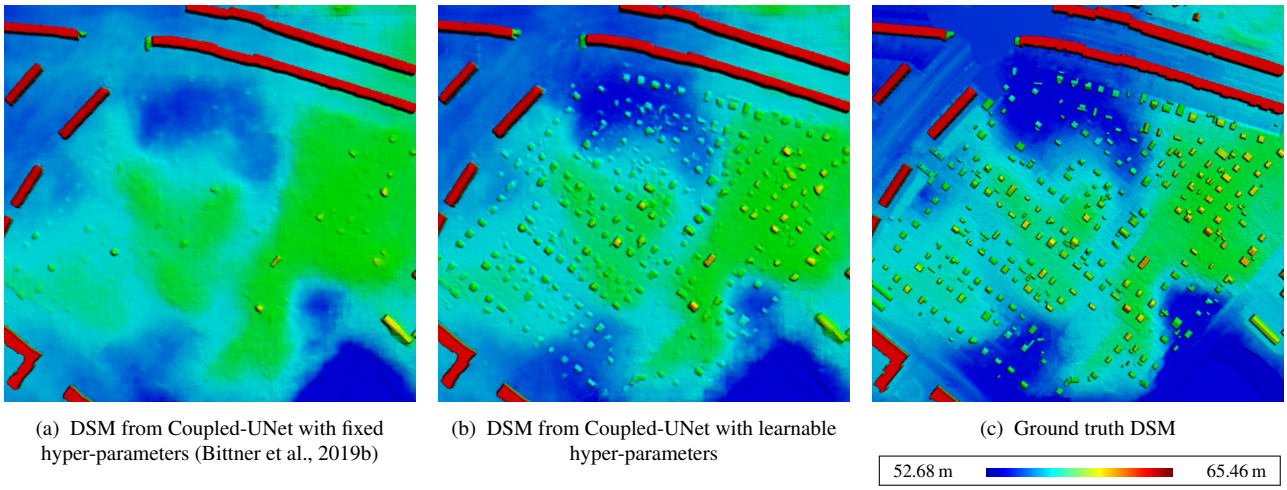
combining three individual losses $l \in \mathcal{L}$ for the most effective task loss minimization.

3. STUDY AREA AND EXPERIMENTAL SETUP

3.1 Datasets

We experimented on three datasets consisting of PAN images showing close to nadir view and photogrammetric DSMs both with *ground sampling distance (GSD)* of 0.5 m.

The *first dataset* shows 410 km^2 of Berlin city, Germany, and was used for training (353 km^2), validation (6 km^2) and testing (50 km^2) of the proposed model. The photogrammetric DSM was generated using *semi-global matching (SGM)* (Hirschmuller, 2008) utilizing six PAN images acquired by the WorldView-1 satellite on two different days, following the workflow of (d'Angelo and Reinartz, 2011). Those DSMs and one of six PAN images were inputs to our model. To perform a learning process, we generated a ground truth by interpolating height information on building roof polygons utilizing the constructed them



(a) DSM from Coupled-UNet with fixed hyper-parameters (Bittner et al., 2019b)

(b) DSM from Coupled-UNet with learnable hyper-parameters

(c) Ground truth DSM

Figure 2. Visual analysis of DSMs, generated by Coupled-UNet cLSGAN with (a) fixed balancing hyper-parameters (Bittner et al., 2019b) and (b) learnable balancing hyper-parameters in comparison to (c) referenced surface model over selected urban area. DSM images are color-shaded for better visualization.

points from *city geography markup language (CityGML)* data freely available on Berlin3D portal [1]. We refer to it as *level of detail (LoD)*2-DSM. More detailed information about LoD2-DSM generation is given by (Bittner et al., 2018).

The *second dataset* shows 3.8 km^2 of Munich city, Germany, and was used to investigate the model's generative capacity over a different urban landscape. The input photogrammetric DSM was generated from three pairs of PAN images acquired from different viewing angles by a space-borne sensor distinctive to the Berlin dataset. Precisely, PAN images from the WorldView-2 satellite with a GSD of 0.5 m were processed with SGM algorithm (Hirschmuller, 2008) to obtain the photogrammetric DSM, following the same procedure of (d'Angelo and Reinartz, 2011). For evaluation, the ground truth LoD2-DSM simulated, similar as before, from CityGML data provided by *Bavarian Agency for Digitisation, High-Speed Internet and Surveying* was used.

Since Berlin and Munich are both European cities and their infrastructures have similarities, we tested the generalization capacity of the proposed model on a *third dataset* showing 2.7 km^2 of Istanbul city, Turkey. PAN images with a resolution of 0.5 m and a derived photogrammetric DSM come from the same space-borne sensor as the Munich dataset—the WorldView-2 satellite. There exists no CityGML data for this area. Therefore, we solely use it for visual evaluation.

3.2 Implementation Details

Our implementation is a *PyTorch*-based extension of the GAN architecture developed by (Isola et al., 2016). The training was performed on 21 480 samples of 256×256 px. The samples were augmented not only by horizontal and vertical flipping but also tiled from the original image with a random overlap every epoch to give the model a clue about building geometries which happened to be on the patch border in previous epochs.

The Coupled-UNet and Coupled-UResNet networks, used in this paper for comparison were trained with minibatch *stochastic gradient descent (SGD)* using the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of $\alpha = 0.0002$ which was dropped by a factor of 10 after 100 epochs and momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The starting values of balancing hyper-parameters in Eq. (6) were equally initialized with 0.3, making w_{cLSGAN} unchangeable and w_{L_1}, w_{normal} learnable, as proposed by (Liebel et al., 2020). The network was trained on a single NVIDIA TITAN X (PASCAL) GPU with 12 GB of memory.

[1] `http://www.businesslocationcenter.de/downloadportal`

4. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we investigate the influence of learnable balancing hyper-parameters between multiple loss terms, explore the effect of combination of short and long skip connections within one network architecture, and inspecte the planarity of roof surface as their smoothness is one of our main concerns. Moreover, the proposed model is tested for its generalization capability for different urban aerials, distinctive from the training dataset.

In cases, where ground truth data was available, the obtained results was quantitatively evaluated using the *root mean squared error (RMSE)*

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

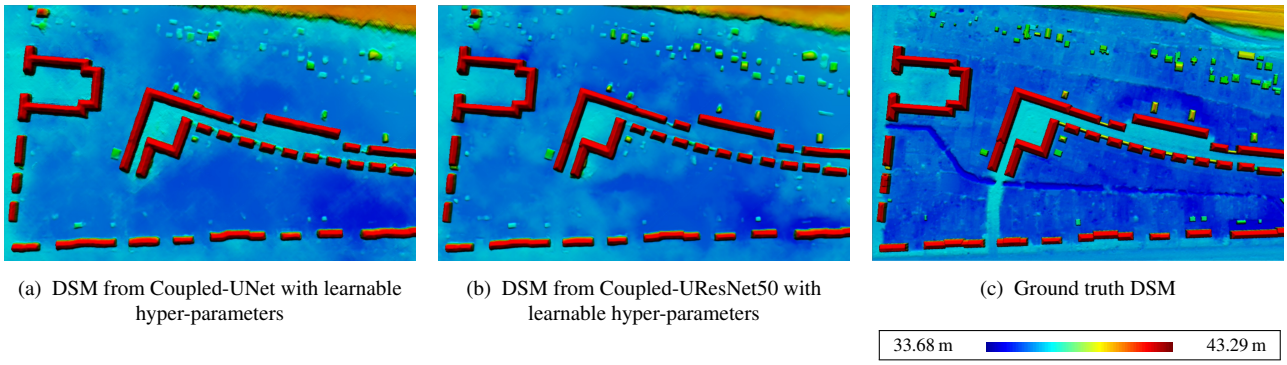
and the *normalized median absolute deviation (NMAD)*

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

where $\boldsymbol{h} = (h_j)_j$ and $\hat{\boldsymbol{h}} = (\hat{h}_j)_j, 1 \leq j \leq n$, denote the observed and the predicted heights, respectively, height errors are defined as Δh_j, and the median error is $m_{\Delta \boldsymbol{h}}$. In cases, when data errors are normally distributed, the constant 1.4826 in the NMAD metric is comparable to the standard deviation. NMAD metric is assumed to be more robust to outliers in the dataset (Höhle and Höhle, 2009).

4.1 Inserting Learnable Balancing Hyper-Parameters

We have started our experiments with integrating learnable balancing hyper-parameters into the already appeared Coupled-UNet architecture proposed by (Bittner et al., 2019b) to confirm its



(a) DSM from Coupled-UNet with learnable hyper-parameters

(b) DSM from Coupled-UResNet50 with learnable hyper-parameters

(c) Ground truth DSM

Figure 3. Visual analysis of DSMs, generated by (a) Coupled-UNet cLSGAN with fixed balancing hyper-parameters (Bittner et al., 2019b) and (b) Coupled-UResNet50 cLSGAN with learnable balancing hyper-parameters in comparison to (c) referenced surface model over selected urban area (first row) and the zoomed version of highlighted area for corresponding models (second row). DSM images are color-shaded for better visualization.

applicability even for a single desired output. Samples of generated DSMs from the Coupled-UNet trained with fixed balancing hyper-parameters and Coupled-UNet with learnable balancing hyper-parameters are depicted in Fig. 2 together with the corresponding ground truth sample. It can easily be observed that the proposed learning strategy for the same model can extract even low-rise densely located buildings. This observation is also confirmed by quantitative evaluation over the whole test area using the chosen metrics. The results related to this experiment are presented in the first two lines of Table 1. The RMSE value for Coupled-UNet with learnable balancing hyper-parameters was dropped down by 0.36 m revealing the increase of accuracy by extracting additional building constructions. The NMAD value also decreased when using the proposed strategy meaning the reduction of overall outliers in the generated DSM.

4.2 Long-Short Skip Connections Within One Network

Encouraged by the improved results from the above experiment, the decision was made to elaborate more on the network architecture and its construction blocks. Mainly, in the next experiments, the backbone structure of the Coupled-UNet was changed to existing ResNet architectures with different levels of extension forming the proposed Coupled-UResNet model. The whole settings were kept similar between the training experiments to be able to compare networks performance. The quantitative results are summarized in Table 1. Comparing evaluation metrics for ResNet-based architectures with different levels of extension, it can be noticed that going deeper and involving more residual blocks helped to improve the RMSE and NMAD values. However, a too deep network with a huge number of parameters can lead to overfitting. Good examples are the Coupled-UResNet models with 101- and 152-layer ResNet as an encoder. The metrics started growing up again after reaching its minimum by the network with 50-layer ResNet among all investigated networks. Hence, we choose the so-called Coupled-UResNet50 model, which single-stream encoder is based on three-layers deep residual blocks forming a regular 50-layer ResNet, as it provides the best results in this experiment.

Table 1. Quantitative results for RMSE and NMAD metrics evaluated over selected test area of Berlin city.

Method	Error	
	RMSE (m)	NMAD (m)
Coupled-UNet (Bittner et al., 2019b)	2.82	0.56
Coupled-UNet modified	2.46	0.53
Coupled-UResNet 18	4.08	0.99
Coupled-UResNet 34	3.92	0.68
Coupled-UResNet 50	**2.41**	**0.48**
Coupled-UResNet 101	2.45	0.49
Coupled-UResNet 152	2.46	0.51

The RMSE error of around 2.4 m between the ground truth and the generated surface model can lead to doubts about good performance of the proposed model. The explanation lays in the time difference between the reference data and the given photogrammetric DSM which we use as input to the model. There can be cases when in one data source the buildings exist and in the other not (due to new buildings construction or their destruction), and vice versa. The model cannot generate new building construction if no initial building structure is located in the processing area. This also reveals, that our model does not hallucinate new structures in surface models. The NMAD metric indicates that the proposed model can generate more continuous DSM surfaces without sudden deviation of values between the neighboring pixels if its values seeks to zero. Therefore, we can say that the proposed Coupled-UResNet50 model produces the most consistent ground surface among others used in the experiment since its NMAD values is the smallest.

But there is still one open question. How the superiority of Coupled-UResNet50 over Coupled-UNet can be distinguished? To demonstrate the visual differences, the samples of DSMs generated by Coupled-UNet and Coupled-UResNet50 are depicted in Fig. 3. At first sight, the shown areas between Coupled-UNet and Coupled-UResNet50 look very similar. The amount of big buildings is the same and coincides with the ground truth. The shape of big buildings is also very resembled. However, it can be observed that, although the Coupled-UNet cLSGAN with learnable balancing hyper-parameters can already extract more low-rise residential buildings in comparison to the Coupled-UNet cLSGAN introduced by (Bittner et al., 2019b), the Coupled-UResNet can reconstruct even more of them (upper part area in Fig. 3b).

4.3 Roofs planarity

To make quantitative evaluation of roofs planarity, we evaluate the flatness and orientation of 3D planes $\boldsymbol{\pi}_k^p$ fitted to predicted roof surface points $\boldsymbol{P}_{k;m,n}^p$ in comparison to 3D planes $\boldsymbol{\pi}_k^t$ fitted to the ground truth roof surface points $\boldsymbol{P}_{k;m,n}^t$, as proposed

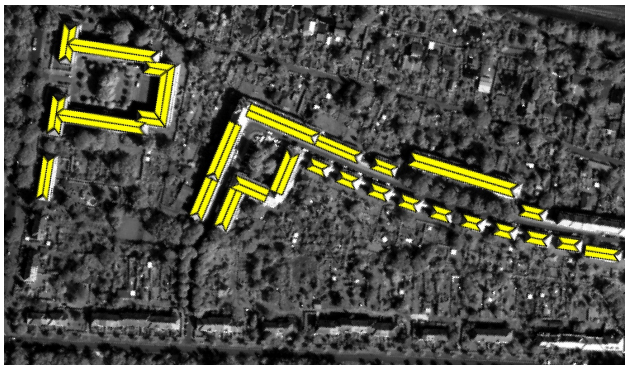


Figure 4. Selected polygons for planarity metrics evaluation overlaid on the pan-chromatic image.

by (Koch et al., 2019). Mainly, the flatness error

$$\varepsilon_{flat}(G(I_1, I_2) \odot \mathcal{P}) = \frac{1}{k} \sum_{\boldsymbol{P}_{k;m,n} \in \mathcal{P}_k} \frac{1}{|\mathcal{P}_k|} d(\boldsymbol{\pi}_k, \boldsymbol{P}_{k;m,n}) \tag{9}$$

is computed over a predicted depth image $G(I_1, I_2)$ masked with binary image $\mathcal{P}$ containing a certain number of planes and represents a standard deviation of averaged distances between the predicated 3D points and fitted 3D plane to them. The orientation error

$$\varepsilon_{orie}(G(I_1, I_2) \odot \mathcal{P}) = \arccos(\boldsymbol{n}_i^{\mathrm{t}} \cdot \check{n}_i^{\mathrm{p}}), \tag{10}$$

is defined as the angle difference between the normal vectors of 3D planes fitted to the predicted surface points and the given ground truth points.

For the scene depicted in Fig. 3b we select some roof polygons from CityGML for which appropriate 3D planes can be estimated. The evaluated flatness and orientation metrics for the Coupled-UResNet model (Bittner et al., 2019b), Coupled-UResNet and Coupled-UResNet50 models with learnable hyper parameters are presented in Table 2. Obtained quantitative results support the previous study and show that the model, precisely Coupled-UResNet50, which includes both the short and long skip connections and trained to minimize the multi-term loss with learnable hyper-parameters can predict the improved building roof planes among the currently developed models. Moreover, one can see that the flatness and orientation of building roof surfaces in comparison to initial photogrammetric DSM is improved more than twice. This leads to the conclusion that the goal of DSM optimization applying the neural networks is achieved.

4.4 Model Generalization

To investigate how well the developed Coupled-UResNet50 model can generalize to diverse urban landscapes, different from the training dataset, we perform a building shape refinement task over city areas of Munich, Germany, and Istanbul, Turkey.

Table 2. Quantitative results for flatness and orientation metrics evaluated over selected test area of Berlin city.

Method	Error	
	ε_{flat}(cm)	$\varepsilon_{orie}(^\circ)$
Photogrammetric DSM	68.59	14.11
Coupled-UNet (Bittner et al., 2019b)	37.19	7.14
Coupled-UNet modified	28.03	5.76
Coupled-UResNet 50	**25.53**	**5.51**

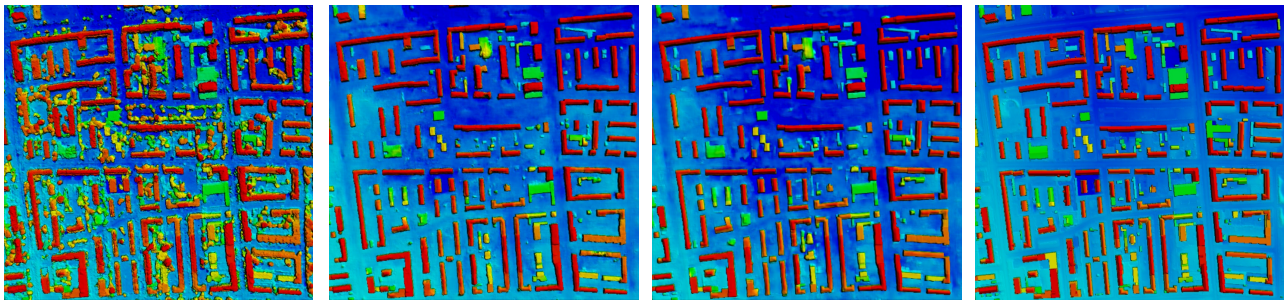
Munich dataset: Even though Munich and Berlin belong to the same country and their building appearances can have a similar style, it is impossible to meet identical building constructions. Since there is a time difference of several years between our satellite images and CityGML data, many inconsistencies in terms of newly built or demolished city constructions exist over the available scene (Liebel et al., 2020). Therefore, we select a smaller area depicted in Fig. 5 which shows no changes in between the acquisition of the photogrammetric DSM and the city model to perform a quantitative evaluation.

Similar to the previous study, the RMSE drops down even stronger between Coupled-UNet and Coupled-UResNet50 in the advantage of the last. This indicates that the proposed model is more general. The next example with a detailed view can demonstrate it better. An area shown in Fig. 6 also establishes the strong outperformance of the proposed Coupled-UResNet50 over the Coupled-UNet (Bittner et al., 2019b), as reconstructed building shapes are more complete and demonstrate close resemblance to the ground truth. It is worth to mention that, in general, the input photogrammetric DSM is very challenging for the model. First of all, it comes from a different sensor. Second, this dataset has more and bigger areas that were filled by interpolation in a post-processing step after the stereo-matching procedure. This influences the appearance of buildings and makes them have slightly distinctive characteristics in comparison to the ones learned by models from Berlin dataset. However, from the demonstrated results we can say that the proposed Coupled-UResNet50 is more generic and resistant to the bad quality of input data. Besides, PAN images enormously influence the realistic reconstruction of building appearances, since from spectral information one can identify the ridge lines for such roof types, like hip, gable, mansard, *etc*. This statement was earlier investigated by (Bittner et al., 2019b).

Istanbul dataset: The generalization results of Coupled-UNet and Coupled-UResNet50 are presented in Fig. 7. From the inputs PAN and DSM, depicted in Fig. 7a and Fig. 7b respectively one can see that the building structures are very different from the European style; the houses are lower in height and their placement is very dense. However, both networks were able to generalize over this area. Under close investigation, it can be recognized that the Coupled-UResNet50 was able to generate better building shapes in terms of smooth roof planes, distinctive roof types (independent from their complexity) and more complete structure, which supports the observation of previous studies.

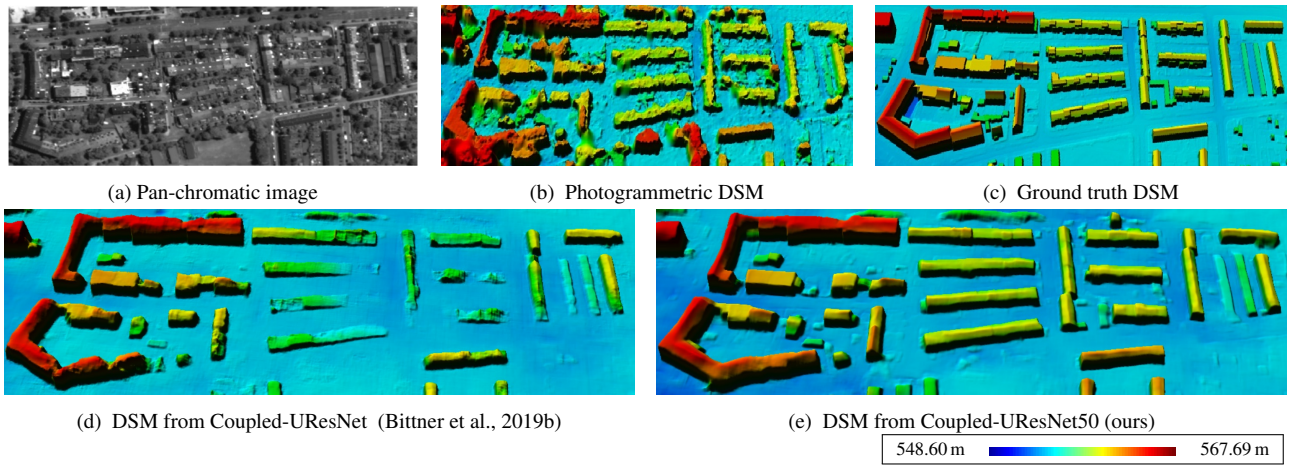
5. CONCLUSION

In this paper, the influence of long skip connection combined with short skip connections within the generative module of *conditional generative adversarial networks (cGANs)* towards the improvement of building geometries in photogrammetric *digital surface models (DSMs)* was investigated. Since photogrammetric DSMs are the product of processing stereo imagery, in our case from space-borne sensors, they are usually of low-quality: partially or badly reconstructed building, strong discontinuities, *etc*. Therefore, the intensity information from the closest to the nadir view *pan-chromatic (PAN)* images was involved for a better understanding of building borders and roof ridge lines by the developed system. Moreover, instead of manual tuning of hyper-parameters related to the weighting of contributions of different



(a) Photogrammetric DSM (b) Coupled-UNet (Bittner et al., 2019b), RMSE: 3.16 (c) Coupled-UResNet50 (ours), RMSE: 2.86 (d) Ground truth

Figure 5. Visual analysis of DSMs, generated by (a) Coupled-UNet cLSGAN with fixed balancing hyper-parameters (Bittner et al., 2019b) and (c) by the proposed Coupled-UResNet50 in comparison to (d) referenced surface model over Munich area. (a) depicts the initial photogrammetric DSM. DSM images are color-shaded for better visualization.



(a) Pan-chromatic image (b) Photogrammetric DSM (c) Ground truth DSM

(d) DSM from Coupled-UResNet (Bittner et al., 2019b) (e) DSM from Coupled-UResNet50 (ours)

Figure 6. Detailed visual analysis of DSMs, generated by (d) Coupled-UNet cLSGAN with fixed balancing hyper-parameters (Bittner et al., 2019b) and (e) the proposed Coupled-UResNet50 in comparison to (c) referenced surface model over selected Munich area. (a) and (b) depict the initial PAN and DSM images, respectively. DSM images are color-shaded for better visualization.

loss terms constructing the main objective function for the training, their automatic learning by the system was implemented.

The obtained results demonstrate that the proposed Coupled-UResNet50 model can extract more buildings, including challenging low-rise constructions, improves the surfaces of building roofs making them smoother and better co-inside with ground truth data and generalize to different urban areas, even very distinctive from the training dataset. A 3D visualization of predicted elevation models illustrates their resemblance to the real urban scenes making them useful for different remote sensing applications.

References

Anderson, E., Thompson, J. and Austin, R., 2005. Lidar density and linear interpolator effects on elevation estimates. *International Journal of Remote Sensing* 26(18), pp. 3889–3900.

Bittner, K., d'Angelo, P., Körner, M. and Reinartz, P., 2018. Dsm-to-lod2: Spaceborne stereo digital surface model refinement. *Remote Sensing* 10(12), pp. 1926.

Bittner, K., Körner, M., Fraundorfer, F. and Reinartz, P., 2019a. Multi-task cgan for simultaneous spaceborne dsm refinement and roof-type classification. *Remote Sensing* 11(11), pp. 1262.

Bittner, K., Reinartz, P. and Körner, M., 2019b. Late or earlier information fusion from depth and spectral data? large-scale digital surface model refinement by hybrid-cgan. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

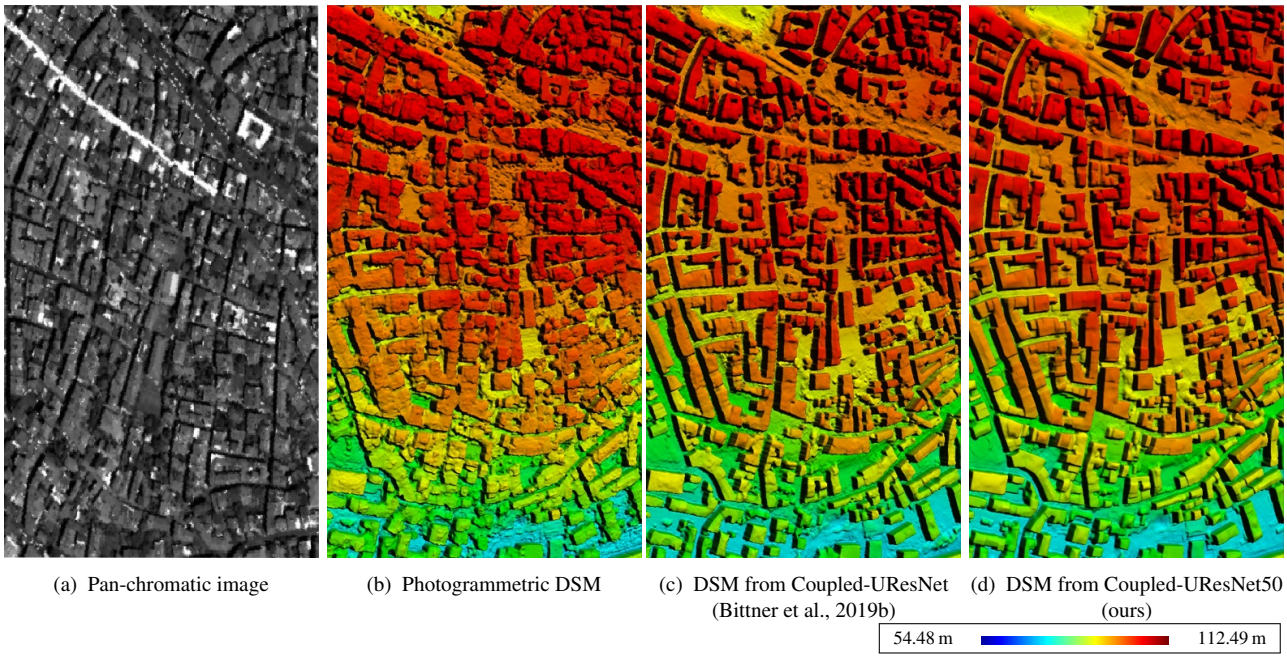
Chen, C. and Li, Y., 2013. A robust multiquadric method for digital elevation model construction. *Mathematical Geosciences* 45(3), pp. 297–319.

d'Angelo, P. and Reinartz, P., 2011. Semiglobal matching results on the isprs stereo matching benchmark. *ISPRS Hannover Workshop* 38(4/W19), pp. 79–84.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S. and Pal, C., 2016. The importance of skip connections in biomedical image segmentation. In: *Deep Learning and Data Labeling for Medical Applications*, Springer, pp. 179–187.

Felicísimo, A. M., 1994. Parametric statistical method for error detection in digital elevation models. *ISPRS Journal of Photogrammetry and Remote Sensing* 49(4), pp. 29–33.

Ghamisi, P. and Yokoya, N., 2018. Img2dsm: Height simulation from single imagery using conditional generative adversarial net. *IEEE Geoscience and Remote Sensing Letters* 15(5), pp. 794–798.



(a) Pan-chromatic image (b) Photogrammetric DSM (c) DSM from Coupled-UResNet (Bittner et al., 2019b) (d) DSM from Coupled-UResNet50 (ours)

Figure 7. Visual analysis of DSMs, generated by (c) Coupled-UNet cLSGAN with fixed balancing hyper-parameters (Bittner et al., 2019b) and (d) the proposed Coupled-UResNet50 over Istanbul city, Turkey. (a) and (b) depict the initial PAN and DSM images, respectively. DSM images are color-shaded for better visualization.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 2672–2680.

Goovaerts, P. et al., 1997. *Geostatistics for natural resources evaluation*. Oxford University Press on Demand.

Hirschmuller, H., 2008. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence* 30(2), pp. 328–341.

Höhle, J. and Höhle, M., 2009. Accuracy assessment of digital elevation models by means of robust statistical methods. *ISPRS Journal of Photogrammetry and Remote Sensing* 64(4), pp. 398–406.

Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A. A., 2016. Image-to-image translation with conditional adversarial networks. *arxiv*.

Kendall, A., Gal, Y. and Cipolla, R., 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: *Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491.

Kingma, D. P. and Ba, J., 2014. Adam: A method for stochastic optimization.

Koch, T., Liebel, L., Fraundorfer, F. and Körner, M., 2019. Evaluation of cnn-based single-image depth estimation methods. In: L. Leal-Taixé and S. Roth (eds), *Proceedings of the European Conference on Computer Vision Workshops (ECCV-WS)*, Springer International Publishing, pp. 331–348.

Liebel, L. and Körner, M., 2018. Auxiliary tasks in multi-task learning. *arXiv preprint arXiv:1805.06334*.

Liebel, L., Bittner, K. and Körner, M., 2020. A generalized multi-task learning approach to stereo dsm filtering in urban areas. to appear.

López, C., 2000. Improving the elevation accuracy of digital elevation models: a comparison of some error detection procedures. *Transactions in GIS* 4(1), pp. 43–64.

Mirza, M. and Osindero, S., 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Mou, L. and Zhu, X. X., 2018. Im2height: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network. *Arxiv Prepr. Arxiv:1802.10249*.

Smith, S., Holland, D. and Longley, P., 2005. Quantifying interpolation errors in urban airborne laser scanning models. *Geographical Analysis* 37(2), pp. 200–224.

Wang, P., 1998. Applying two dimensional kalman filtering for digital terrain modelling. *Proceedings of International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences* pp. 649–656.

Zhang, Z., Liu, Q. and Wang, Y., 2018. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters* 15(5), pp. 749–753.