Late or Earlier Information Fusion from Depth and Spectral Data? Large-Scale Digital Surface Model Refinement by Hybrid-cGAN

Ksenia Bittner, Peter Reinartz
German Aerospace Center (DLR)
Munich, Germany

ksenia.bittner@dlr.de
peter.reinartz@dlr.de

Marco Körner
Technical University of Munich (TUM)
Munich, Germany

marco.koerner@tum.de

Abstract

We present the workflow of a digital surface model (DSM) refinement methodology using a HYBRID-cGAN *where the generative part consists of two encoders and a common decoder which blends the spectral and height information within one network. The inputs to the Hybrid-cGAN are single-channel photogrammetric DSMs with continuous values and single-channel pan-chromatic (PAN) half-meter resolution satellite images. Experimental results demonstrate that the earlier information fusion from data with different physical meanings helps to propagate fine details and complete an inaccurate or missing 3D information about building forms. Moreover, it improves the building boundaries making them more rectilinear.*

1. Introduction

Digital surface models (DSMs) with detailed and realistic building shapes are beneficial data sources for many remote sensing applications, such as urban planning, cartographic analysis, environmental investigations, *etc*. Several ways for deriving 3D elevation morphology, but the DSMs generated from space-borne data using stereo image matching techniques nowadays have lower costs compared to other technologies, show relatively high spatial resolution and wide coverage which is important for large-scale remote sensing applications, such as disaster monitoring.

However, apart from many advantages, DSMs generated by image-based matching techniques show a reasonable amount of noise and outliers because of matching errors or occlusions due to densely located buildings or trees, which cover parts of the building constructions. To solve these problems, algorithms from computer vision have been analyzed and adapted to satellite imagery processing. In the literature, very few of the proposed approaches work towards photogrammetric DSM improvement of urban areas. In
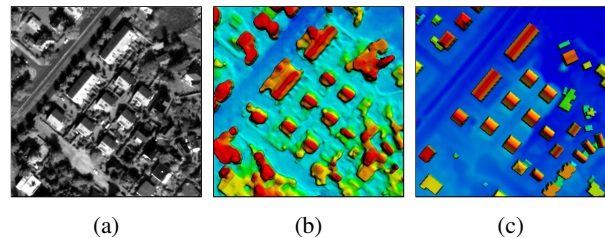

Figure 1: Sample of area from our dataset illustrated both inputs to the network (a) *pan-chromatic (PAN)* image and (b) photogrammetric DSM, and (c) the ground truth *level of detail (LoD)*-2-DSM. The DSM images are color-shaded for better visualization.

contrast, earlier methodologies investigate the DSMs refinement by applying filtering techniques, such as Gaussian [1], Kalman [2], geostatistical [3], or interpolation routines, including *inverse distance weighting (IDW)* and kriging [4]. Despite achieving smoother surfaces of roof planes, they negatively influence the steepness of the building walls. Later, some methodologies proposed to additionally utilize spectral images for DSM refinement tasks, as, for instance, they contain accurate information about object boundaries or texture. For instance, Krauß *et al.* [5] transfer segmentation information from stereo satellite imagery to the DSM and, from statistical analysis and spectral information, perform object detection and classification. As a further step, this information is used to refine the DSM. Sirmacek *et al.* [6] propose building shape refinement on DSMs by a multiple-step procedure. First, they extract possible building segments by thresholding the *normalized digital surface model (nDSM)*. Then, by considering Canny edge information from the spectral image, they fit rectangular boxes to detect building shapes. Finally, applying this information about detected rectangular boundaries, the building shapes are further enhanced. The drawback of this methodology is that they assign one single height value to each generated

building object. Moreover, it is limited to the detection and enhancement of rectangular buildings only.

Within recent years, the already well-studied branch of the deep learning family, *convolutional neural networks (CNNs)*, has also been applied to the remote sensing field. They achieve state-of-the-art results for image classification, object detection, or semantic segmentation tasks. However, most of these methodologies extract information from spectral imagery, while depth image processing is still not well explored, as it is not straightforward to work with continuous values. For example, Eigen *et al.* [7] employ cascade deep network which first performs a coarse global prediction from a single spectral image and refines the predictions locally afterwards. Liu *et al.* [8] join CNNs and *conditional random fields (CRFs)* in a unified framework while making use of superpixels for preserving sharp edges. In case of height prediction problems, only a couple of works made attempts to improve DSMs. Our earlier approaches [9, 10] propose to refine building shapes to a high level of details from photogrammetric half-meter resolution satellite DSMs using *generative adversarial network (GAN)*-based techniques. Mainly, they investigate *conditional generative adversarial networks (cGANs)* with two objective functions—*i.e.* the negative log-likelihood and least square residuals—to generate accurate *light detection and ranging (LIDAR)*-like and LoD-2-like DSMs with enhanced 3D building shapes directly from noisy input data. Moreover, the experiment performed on a completely new dataset, belonging to different geographical areas, showed the network's potential to generalize well to different cities with complex constructions without many difficulties. As it is common in the field of remote sensing to fuse data with different modalities to complement missing knowledge, in our followed work [11] we introduce a cGAN-based network which merges depth and spectral information within an end-to-end framework. Fusing data from separate networks—which are fed with PAN images and DSMs—is performed at a *later* stage right before producing the final output.

Following up to the aforementioned approach, in this work, we investigate the fusion of spectral (Fig. 1a) and height (Fig. 1b) information at an *earlier* stage within an end-to-end Hybrid-cGAN network to further improve not only the small and simple residential buildings, but also complex industrial ones. Besides, we add an auxiliary *normal vector loss* term to the objective function enforcing the model to produce more planar and flat roof surfaces, similar to the desired LoD2-DSM (Fig. 1c) artificially produced from *city geography markup language (CityGML)* data.

2. Methodology

2.1. Objective function

The advent of GAN-based domain adaptation neural networks, introduced by Goodfellow *et al.* [12], lead to great performance gains in generating realistic, but entirely artificial images. These GANs realize an adversarial manner of learning by training a pair of networks in a competing way: A *generator* $G(\boldsymbol{z}|\boldsymbol{x}) = \boldsymbol{y}$ produces a fake image $\boldsymbol{y}$ from a latent vector $\boldsymbol{z} \sim p_z(\cdot)$ drawn from any distribution $p_z(\cdot)$—*e.g.* a uniform distribution $p_z(\cdot) = \mathrm{Unif}(\cdot; a, b)$—which looks like a real one $\boldsymbol{y}^* \sim p_{\mathrm{real}}(\cdot)$. Adversarially, a *discriminator* $D(\boldsymbol{y}|\boldsymbol{x}) \in [0, 1]$ tries to decide whether a presented image $\boldsymbol{y}$ is a real or a generated fake one. Frequently, some external information in the form of a source image $\boldsymbol{x}$ is additionally used as an input to restrict both the generator in its output and the discriminator in its expected input. The objective function for such conditional GANs can be expressed through a two-player minimax game

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

where $\mathbb{E}[\cdot]$ denotes the expectation value. To overcome the problem of instability during training cGAN when its objective function is based on the negative log-likelihood, we use a least square loss L_2 instead which yields the *conditional least square generative adversarial network (cLSGAN)* objective function

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

In order to obtain DSMs in which the buildings feature sharply defined ridgelines and steep walls, we utilize the L_1 distance

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

which prevents blurring effects.

As our major goal is to improve roof surfaces by making them flatter and looking closer to realistic ones, we integrate a *normal vector loss* term

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

into the learning process, as proposed by Hu *et al.* [13]. This normal vector loss measures the angle between the set of surface normals $\mathcal{N}^{\mathrm{p}} = \{\boldsymbol{n}_1^{\mathrm{p}}, \dots, \boldsymbol{n}_m^{\mathrm{p}}\}$ of an estimated DSM and the set of surface normals $\mathcal{N}^{\mathrm{t}} = \{\boldsymbol{n}_1^{\mathrm{t}}, \dots, \boldsymbol{n}_m^{\mathrm{t}}\}$ of the target LoD2-DSM. Adding those three terms together leads to our final objective

$$G^{\star} = \arg\min_G \max_D \mathcal{L}_{\mathrm{cGAN}}(G, D) + \lambda \mathcal{L}_{L_1}(G)$$
$$+ \gamma \mathcal{L}_{\mathrm{normal}}(\mathcal{N}^{\mathrm{t}}, \mathcal{N}^{\mathrm{p}}), \quad (5)$$
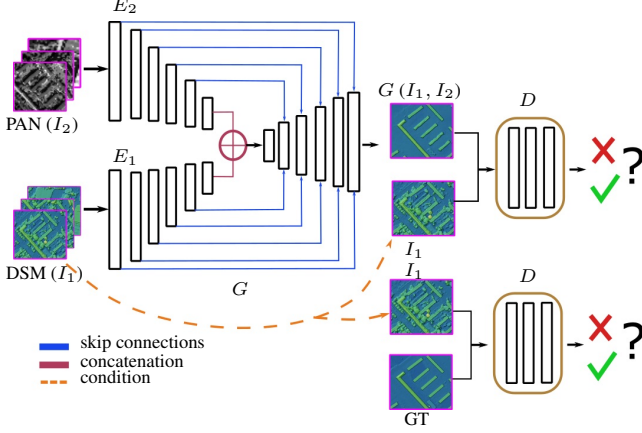
Figure 2: Schematic overview of the proposed architecture for the building shape refinement on photogrammetric DSMs by Hybrid-cGAN using both depth and spectral information.

where G intents to minimize the objective function $\mathcal{L}_{cLSGAN}(G, D)$ against D that aims to maximize it. The parameters $0 \leq \lambda \in \mathbb{R}$ and $0 \leq \gamma \in \mathbb{R}$ are the balancing hyper-parameters.

2.2. Network architecture

We have already made an attempt to investigate the LoD2-like DSM generation with enhanced building shapes from a single noisy photogrammetric DSM using cGAN architecture with an objective function based on least square residuals [10]. In this work, we refer to it as a single-stream cGAN. As each photogrammetric DSM is a product obtained from panchromatic image pairs, it seems reasonable to integrate depth and spectral data into one single network, as the latter provides sharper information about building silhouettes, which allows not only a better reconstruction of missing building parts but also the refinement of building outlines. We fuse two separate but identical *UNet*-type networks at the later end within the G part of a cGAN, where the first stream is fed with the PAN image and the second stream with the stereo DSM, creating a so-called WNet architecture [11].

In this paper, we examine the potential of an earlier fusion of data from different modalities, as it could potentially even better blend together the depth and spectral information. Mainly, the generator G of the proposed Hybrid-cGAN network consists of two encoders E_1 and E_2, concatenated at the top layer, and a common decoder, which integrates information from two different modalities and generates an LoD2-like DSM with refined building shapes. The inputs to E_1 are the single-channel orthorectified PAN images, while E_2 receives the single-channel photogrammetric DSMs with continuous values. Since intensity and depth

information have different physical meanings, it is unlikely to make sense to jointly propagate them right from the beginning. It seems reasonable to separate them first and allow the network to learn the most valuable features from each modality itself. The generator G is constructed based on an U-form network proposed by Isola *et al.* [14]. As a result, in our case it has 14 skip connections from both the spectral stream and the depth stream allowing the decoder to learn back detailed features that were lost by pooling in the encoders. In particular, the encoder and decoder consist of 8 convolutional layers each, followed by a *leaky rectified linear unit (LReLU)* [15] activation function

$$\sigma_{\text{LReLU}}(z) = \begin{cases} z, & \text{if } z > 0 \\ az, & \text{otherwise} \end{cases}, a \in \mathbb{R}^+,$$

and *batch normalization (BN)* in case of the encoder, and a *rectified linear unit (ReLU)* activation function

$$\sigma_{\text{ReLU}}(z) = \begin{cases} z, & \text{if } z > 0 \\ 0, & \text{otherwise} \end{cases}$$

and BN in case of the decoder. On top of the generator network G, the $\tanh$ activation function $\sigma_{\text{tanh}}(z) = \tanh(z)$ is applied.

The discriminator network D is a binary classification network constructed with 5 convolutional layers, followed by LReLU activation function and a BN layer. It has a *sigmoid* activation function $\sigma_{\text{sigm}}(z) = \frac{1}{1+e^{-z}}$ at the top layer to output the likelihood that the input image belongs either to class 1 ("real") or class 0 ("generated"). A schematic representation of the proposed architecture is depicted in Fig. 2.

All along the training phase, the two networks G and D are trained at the same time by alternating one gradient descent step of D and one gradient descent step of G. During the inference process, only the trained generator model G of the Hybrid-cGAN network is involved.

3. Study Area and Experiments

3.1. Dataset

Experiments have been carried out on data showing the city of Berlin, Germany, covering a total area of $410 \, \text{km}^2$. It consists of half-meter resolution orthorectified PAN images showing the closest nadir view and photogrammetric DSMs generated with *semi-global matching (SGM)* [16] from six panchromatic Worldview-1 images acquired at two different days. As ground truth, the artificial LoD2-DSM, generated with a resolution of 0.5 px from a CityGML data model together with an available *digital terrain model (DTM)*, was used. The CityGML data model is freely available from the download portal *Berlin 3D* [1]. We followed the same LoD2-

---

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

DSM creation procedure as in Bittner *et al.* [10].

3.2. Implementation and Training Details

The proposed Hybrid-cGAN network was realized using the *PyTorch* python package based on the implementation published by Isola *et al.* [14]. The prepared training dataset covers $353\,\text{km}^2$ and consists of $21\,480$ pairs of patches of size $256 \times 256\,\text{px}$ obtained by tiling the given satellite image on the fly with random overlap in both horizontal and vertical directions. This procedure provides the network the possibility to learn building shapes, which, during one epoch, may be located on the patch border and, as a result, are only partially introduced to the network. For the validation phase, an area covering $6\,\text{km}^2$ was used for tuning the hyper-parameters. The Hybrid-cGAN network and others, used in this paper for comparison, were trained on mini-batch *stochastic gradient descent (SGD)* using the Adam optimizer [17] with an initial learning rate of $\alpha = 0.0002$ and momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We set the weighting hyper-parameters $\lambda = 1000$ and $\gamma = 10$ after performing training and examining the resulting generated images from the validation dataset. The complete 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.

Different to the training phase, where two networks G and D were trained at the same time by alternating one gradient descent step of D and one gradient descent step of G, in inference phase only the trained generator model G was involved. Stitching the predicted LoD2-like height images—entirely unseen during training and validation phases—with a fixed overlap of half the patch size in both horizontal and vertical directions, we generated the final full image covering an area of $50\,\text{km}^2$.

4. Results and Discussion

Selected test samples of the DSMs generated by the single-stream cGAN model [10], the two-stream WNet-cGAN model [11], and our proposed Hybrid-cGAN model are illustrated in Figs. 3, 5 and 6. One can notice that small buildings in all generated DSMs show more rectilinear borders and are not merged with adjacent trees, as present in the input DSMs (*c.f*. Figs. 3c, 5b and 6b). The integration of spectral information into the model obviously benefits the building reconstruction process. First of all, the number of reconstructed buildings is increased. For instance, the magenta arrows in Fig. 3b highlight the areas in the DSM generated by the single-stream cGAN model, where the model was not able to reconstruct individual buildings, as opposed to the WNet-cGAN network and our proposed Hybrid-cGAN network. Second, the roof ridge lines are more distinguishable and rectilinear. This statement can be also confirmed by exemplary investigating the profiles of

the two buildings highlighted in Fig. 3f. From Fig. 4 we can notice that the Hybrid-cGAN network was able to reconstruct much finer building shapes more similar to the ground truth. Moreover, the surfaces of roof planes are smoother or even flat in many cases affirming the influence of the normal vector loss. The profiles also demonstrate the strength of all networks to separate the buildings from adjacent vegetation. From the demonstrated results we can further conclude that most of the generated building shapes followed the correct pattern and feature improved roof forms. However, this statement is true for many residential and not big industrial buildings. How does the network behave in case of big and complicated building structures? The single depth-stream cGAN model [10] and the two-stream WNet-cGAN model [11] only partially extract such buildings. In case of spectral information integration, although at the late fusion, it helps to improve the silhouette of the buildings as well as the detailed constructions on the rooftops (*c.f*. Figs. 5e and 6e) but still misses or has incompleted inside parts of structures. Moreover, as the input photogrammetric DSMs contain noise and many outliers, they propagate along the height image reconstruction and influence the results, as indicated by the dark blue areas in Figs. 5d and 5e, and Figs. 6d and 6e. On the other hand, the proposed Hybrid-cGAN network was able to not only eliminate those artifacts, but also to reconstruct the complete building structures at any single detail. Besides, although the building rooftops seem to be entirely flat in the ground truth data (*c.f*. Figs. 5c and 6c)—which is not the case in reality—, such cases do not confuse the model during the training phase and make it capable to preserve detailed 3D information from input photogrammetric DSM. Those observations prove that the introduced Hybrid-cGAN architecture may successfully blend together the spectral and height information. The earlier combination of both modalities forces the network to accommodate the information even better.

In order to evaluate the generated elevation models quantitatively, we utilized error metrics commonly used in the relevant literature [18–21], namely, the *root mean squared error (RMSE)*

$$\varepsilon_{\text{RMSE}}(\boldsymbol{h}, \hat{\boldsymbol{h}}) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\hat{h}_j - h_j)^2}, \tag{6}$$

the *normalized median absolute deviation (NMAD)*

$$\varepsilon_{\text{NMAD}}(\boldsymbol{h}, \hat{\boldsymbol{h}}) = 1.4826 \cdot \underset{j}{\text{median}}(|\Delta h_j - m_{\Delta\boldsymbol{h}}|), \tag{7}$$

and the *mean absolute error (MAE)*

$$\varepsilon_{\text{MAE}}(\boldsymbol{h}, \hat{\boldsymbol{h}}) = \frac{1}{n}\sum_{j=1}^{n}|\hat{h}_j - h_j|, \tag{8}$$

(a) Pan-chromatic image

(b) Single depth-stream DSM [10]

(c) Photogrammetric DSM

(d) WNet-cGAN [11]

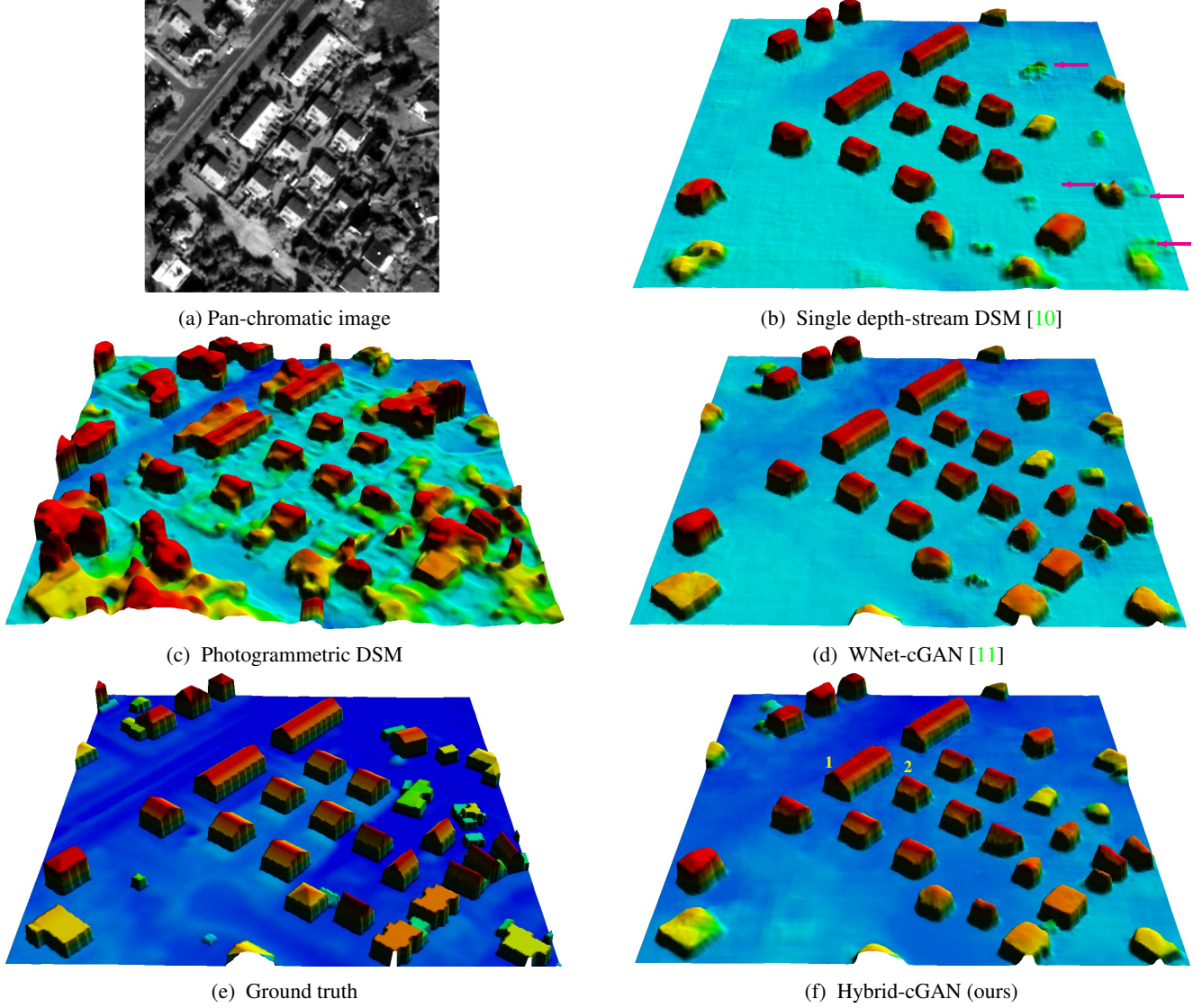(e) Ground truth

(f) Hybrid-cGAN (ours)

Figure 3: Visual analysis of DSMs, generated by (c) a standard photogrammetric method, (b) the single-stream cGAN model [10], (d) the two-stream WNet-cGAN [11], and (f) our proposed Hybrid-cGAN architecture. The DSM images are color-coded for better visualization.

where $\boldsymbol{h} = (h_j)_j$ and $\hat{\boldsymbol{h}} = (\hat{h}_j)_j, 1 \leq j \leq n$, denote the actually observed and the predicted heights, respectively, height errors are defined as Δh_j, and the median error is $m_{\Delta \boldsymbol{h}}$. The constant 1.4826, included in the NMAD metric, is comparable to the standard deviation if the data errors are distributed normally. This estimator can be considered more robust to outliers in the dataset [18].

To exclude the influence of time acquisition difference between the photogrammetric DSMs and the CityGML data model—carrying the risk of absence or the appearance of new buildings—, we manually checked the evaluation regions in this regard and carried out evaluation on regions showing buildings in both the photogrammetric DSM as well as in the reference LoD2-DSM.

The obtained results for the three test area samples are presented in Tables 1 to 3. In comparison to the other methods, the DSMs created by the single-stream model [10] showed inferior results in terms of $\varepsilon_{\mathrm{RMSE}}$ for all three areas. As we have already pointed out before in Fig. 3b, the model was not able to reconstruct some buildings at all or only partially reconstructed them, exemplary highlighted in Figs. 5d and 6d.

With the intensity information integrated into the learning process, the RMSE error $\varepsilon_{\mathrm{RMSE}}$ decreased and, in case

(a) Single depth-stream DSM [10]   (b) WNet-cGAN [11]   (c) Hybrid-cGAN (ours)

■ input DSM   ■ Ground truth   ■ Single depth-stream DSM [10]
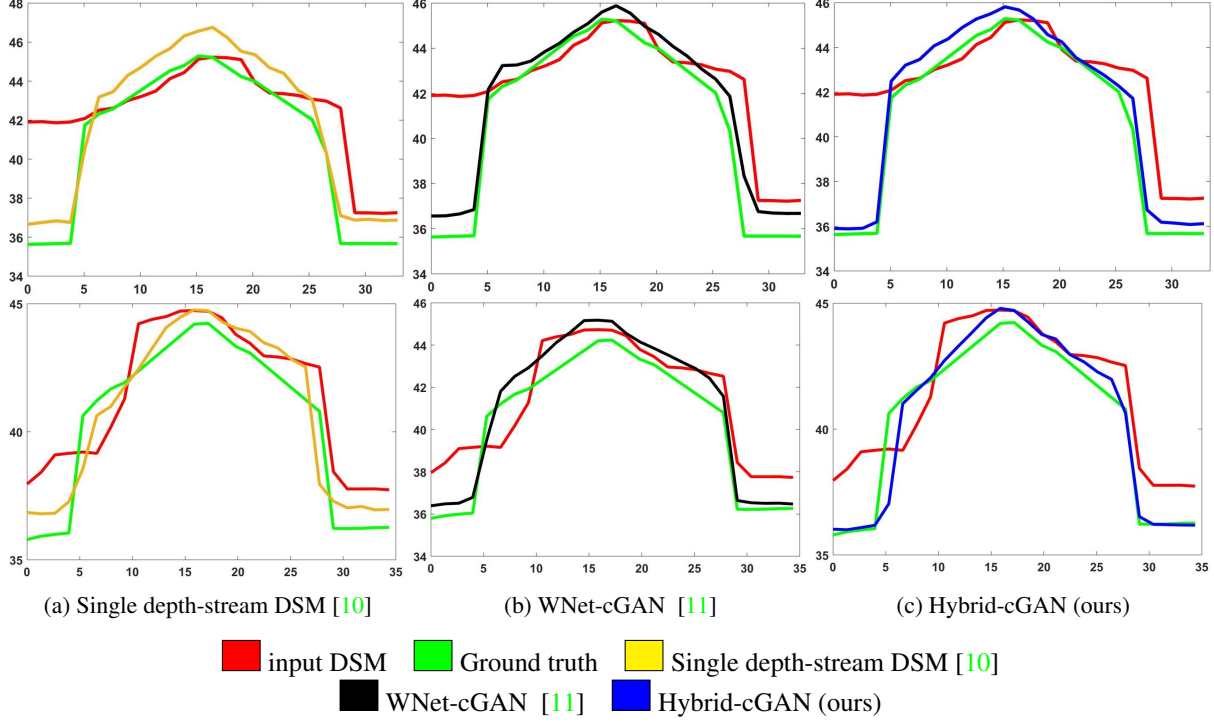■ WNet-cGAN [11]   ■ Hybrid-cGAN (ours)

Figure 4: Visual investigation of profiles for two selected buildings from the DSMs generated by (a) the single-stream cGAN model [10], (b) the two-stream WNet-cGAN model [11], and (c) the proposed Hybrid-cGAN architecture. The first row shows profiles of building "1" and the second row depicts the profiles of building "2" (*c.f.* Fig. 3f).

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

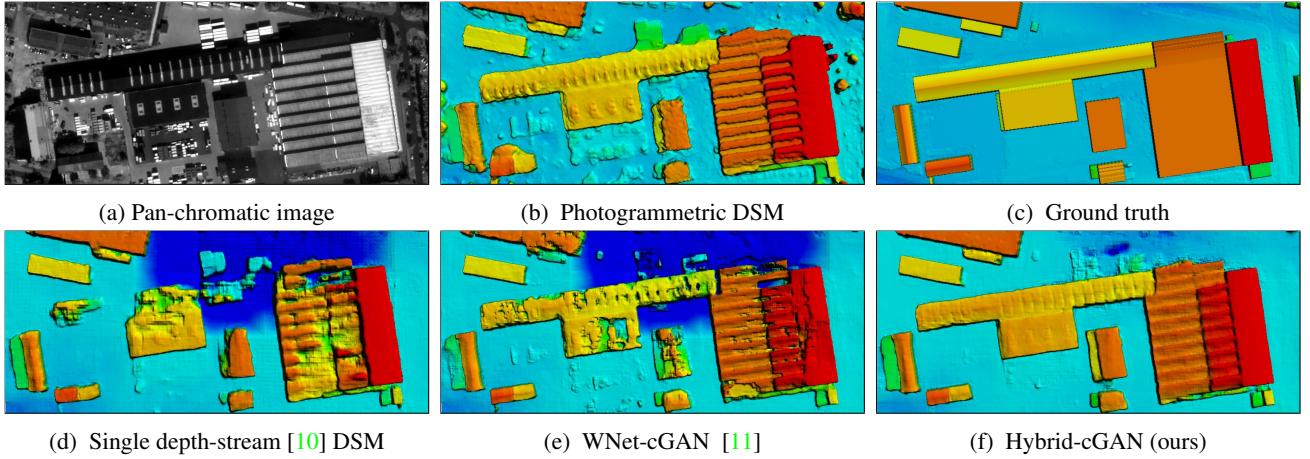(d) Single depth-stream [10] DSM   (e) WNet-cGAN [11]   (f) Hybrid-cGAN (ours)

Figure 5: Visual analysis of DSMs, generated by (d) the single-stream cGAN model [10], (e) the two-stream WNet-cGAN [11], and (f) our proposed Hybrid-cGAN architecture. The DSM images are color-shaded for better visualization.

of our proposed Hybrid-cGAN , achieved the lowest value, even smaller than for the input photogrammetric DSM. This observation provides evidence that the proposed model improves the noisy and inaccurate photogrammetric DSMs to a high level of details. Considering the other two metrics, Hybrid-cGAN also outperformed the competing models, except for the third test area. The single-stream DSM [10]

achieved the lowest NMAD error with $\varepsilon_{\mathrm{NMAD}} = 0.45$. This seems reasonable, as the buildings show flat roofs in the ground truth, which is not the case for neither the input photogrammetric DSM, nor for the DSMs generated by the models with spectral information integrated. As the NMAD metric is sensible to outliers, it shows higher values for the results with more detailed roofs.

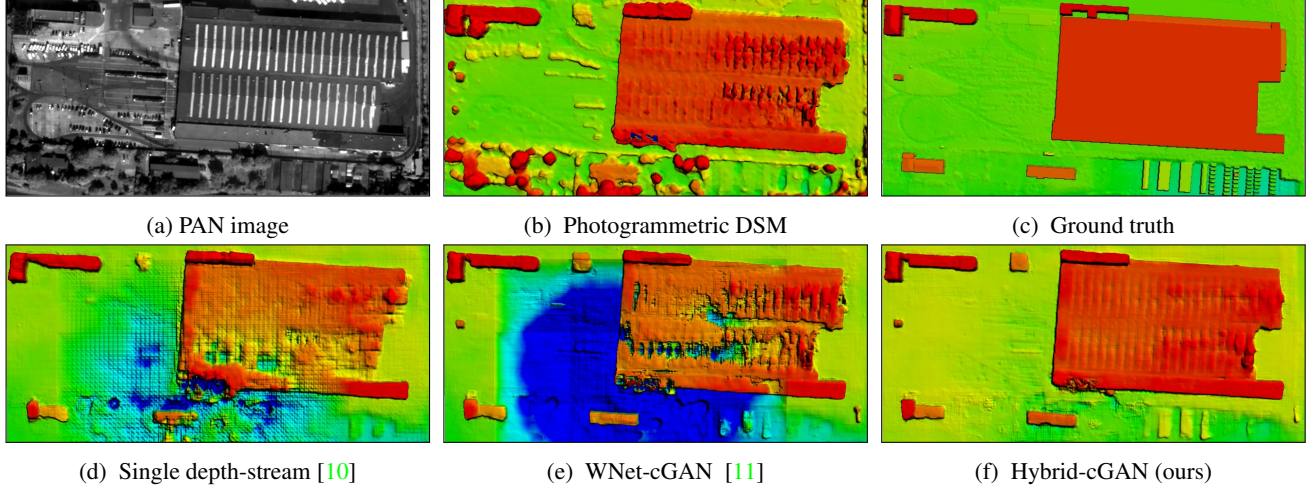|		
(a) PAN image	(b) Photogrammetric DSM	(c) Ground truth
(d) Single depth-stream [10]	(e) WNet-cGAN [11]	(f) Hybrid-cGAN (ours)

Figure 6: Visual analysis of an DSM, generated by (d) the single-stream cGAN model [10], (e) the two-stream WNet-cGAN model [11], and (f) our proposed Hybrid-cGAN architecture. The DSM images are color-coded for better visualization.

Table 1: Quantitative results for RMSE, NMAD, MAE metrics evaluated on 17 selected buildings existed on both input photogrammetric DSM and ground truth LoD2-DSM of the first area depicted in Fig. 3.

Method	Error		
	RMSE (m)	NMAD (m)	MAE (m)
photogrammetric DSM	1.66	1.01	1.23
single-stream cGAN [10]	2.28	1.09	1.86
WNet-cGAN [11]	1.63	0.72	1.22
Hybrid-cGAN (ours)	**1.52**	**0.62**	**0.96**

Table 2: Quantitative results for RMSE, NMAD, MAE metrics evaluated on 7 selected buildings existed on both input photogrammetric DSM and ground truth LoD2-DSM of the first area depicted in Fig. 5.

Method	Error		
	RMSE (m)	NMAD (m)	MAE (m)
photogrammetric DSM	2.72	**1.09**	**1.57**
single-stream cGAN [10]	4.13	1.88	2.68
WNet-cGAN [11]	3.89	2.03	2.64
Hybrid-cGAN (ours)	**2.64**	1.34	1.69

Our results demonstrate that deep learning models visually produce fairly reasonable reconstructed DSMs. However, the used RMSE, NMAD, MAE metrics do not give good enough insight into the depth estimation quality, as they mainly consider the overall accuracy by reporting

Table 3: Quantitative results for RMSE, NMAD, MAE metrics evaluated on 4 selected buildings existed on both input photogrammetric DSM and ground truth LoD2-DSM of the first area depicted in Fig. 6.

Method	Error		
	RMSE (m)	NMAD (m)	MAE (m)
photogrammetric DSM	2.19	0.62	**1.12**
single-stream cGAN [10]	3.38	**0.45**	2.99
WNet-cGAN [11]	3.16	1.15	2.02
Hybrid-cGAN (ours)	**1.91**	0.56	1.22

global statistics of depth residuals. Moreover, the available ground truth data with only sufficient quality also influences evaluation procedure.

5. Conclusion

We presented a methodology for automatic building shape refinement from low-quality *digital surface models (DSMs)* to the *level of detail (LoD)* 2 from multiple space-borne remote sensing data on the basis of *conditional generative adversarial networks (cGANs)*. The network automatically combines the advantages of *pan-chromatic (PAN)* imagery and photogrammetric DSM, while complementing their individual drawbacks, and, from obtained results, shows the potential of generating DSMs with completed not only residential but also industrial building structures. Moreover, the generated roof surfaces are smoother and more planar, giving evidence of the positive influence of the auxiliary *normal vector loss* function. Besides, a 3D visualization of the generated elevation models illustrates the

realistic appearance of the buildings and their strong resemblance to the ground truth.

References

[1] J. P. Walker and G. R. Willgoose, "A comparative study of australian cartometric and photogrammetric digital elevation model accuracy," *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 7, pp. 771–779, 2006.

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

[3] A. M. Felicísimo, "Parametric statistical method for error detection in digital elevation models," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 49, no. 4, pp. 29–33, 1994.

[4] E. Anderson, J. Thompson, and R. Austin, "Lidar density and linear interpolator effects on elevation estimates," *International Journal of Remote Sensing*, vol. 26, no. 18, pp. 3889–3900, 2005.

[5] T. Krauß and P. Reinartz, "Enhancement of dense urban digital surface models from vhr optical satellite stereo data by pre-segmentation and object detection," *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 38, no. 1, p. 6, 2010.

[6] B. Sirmacek, P. dAngelo, T. Krauss, and P. Reinartz, "Enhancing urban digital elevation models using automated computer vision techniques," *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2010.

[7] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374.

[8] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2024–2039, 2016.

[9] K. Bittner and M. Körner, "Automatic large-scale 3d building shape refinement using conditional generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018, pp. 2000–2002.

[10] K. Bittner, P. dAngelo, M. Körner, and P. Reinartz, "Dsm-to-lod2: Spaceborne stereo digital surface model refinement," *Remote Sensing*, vol. 10, no. 12, p. 1926, 2018.

[11] K. Bittner, M. Körner, and P. Reinartz, "DSM Building Shape Refinement from Combined Remote Sensing Images based on WNet-cGANs," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, accepted, 2019.

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

[13] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," 2018. arXiv: 1803.08673 [cs.ML].

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2016. arXiv: 1611.07004 [cs.CV].

[15] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier Nonlinearities improve Neural Network Acoustic Models," in *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[16] P. d'Angelo and P. Reinartz, "Semiglobal matching results on the isprs stereo matching benchmark," *ISPRS Hannover Workshop*, vol. 38, no. 4/W19, pp. 79–84, 2011.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. arXiv: 1412.6980 [cs.ML].

[18] J. Höhle and M. Höhle, "Accuracy assessment of digital elevation models by means of robust statistical methods," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 64, no. 4, pp. 398–406, 2009.

[19] J. Zhang, T. Zhu, Y. Tang, and W. Zhang, "Geostatistical approaches to refinement of digital elevation data," *Geo-spatial Information Science*, vol. 17, no. 4, pp. 181–189, 2014.

[20] A. F. Elaksher and J. Bethel, "Refinement of digital elevation models in urban areas using breaklines via a multi-photo least squares matching algorithm," *Journal of Terrestrial Observation*, vol. 2, no. 2, p. 7, 2010.

[21] M. L. Hobi and C. Ginzler, "Accuracy assessment of digital surface models based on worldview-2 and ads80 stereo remote sensing data," *Sensors*, vol. 12, no. 5, pp. 6347–6368, 2012.