U-IMG2DSM: Unpaired Simulation of Digital Surface Models With Generative Adversarial Networks

M. E. Paoletti, *Graduate Student Member, IEEE*, J. M. Haut, *Senior Member, IEEE*, P. Ghamisi, *Senior Member, IEEE*, N. Yokoya, *Senior Member, IEEE*, J. Plaza, *Senior Member, IEEE*, and A. Plaza, *Fellow, IEEE*

***Abstract*— High-resolution digital surface models (DSMs) provide valuable height information about the Earth's surface, which can be successfully combined with other types of remotely sensed data in a wide range of applications. However, the acquisition of DSMs with high spatial resolution is extremely time-consuming and expensive with their estimation from a single optical image being an ill-possed problem. To overcome these limitations, this letter presents a new unpaired approach to obtain DSMs from optical images using deep learning techniques. Specifically, our new deep neural model is based on variational autoencoders (VAEs) and generative adversarial networks (GANs) to perform image-to-image translation, obtaining DSMs from optical images. Our newly proposed method has been tested in terms of photographic interpretation, reconstruction error, and classification accuracy using three well-known remotely sensed data sets with very high spatial resolution (obtained over Potsdam, Vaihingen, and Stockholm). Our experimental results demonstrate that the proposed approach obtains satisfactory reconstruction rates that allow enhancing the classification results for these images. The source code of our method is available from: https://github.com/mhaut/UIMG2DSM.**

***Index Terms*— Digital surface models (DSMs), generative adversarial networks (GANs), image-to-image problems, optical imaging, variational autoencoder (VAEs).**

I. Introduction

DIGITAL surface models (DSMs) [1] are 2-D-data products that capture the height information from the Earth's surface, taking into account all natural and man-made objects to provide detailed elevation (height) data. In addition to laser scanning or digitized topographic maps, very high resolution (VHR) DSMs can be obtained by processing optical and stereoscopic images captured by both aerial and satellite instruments with submetric ground resolution (e.g., Quickbird, IKONOS, WoldView-2, GeoEye-2, or Cartosat-1 missions) [2]. These instruments exhibit great spatial details that can be exploited to provide accurate DSMs for a wide range of applications, such as management of agricultural and natural resources [3], urban planning [4], and catastrophe/damage assessment [5]. Also, the combination of DSMs with VHR optical image data has been proven to be very useful in land-cover classification tasks [6], where the spectral characteristics of the materials are combined with their height in order to obtain highly discriminative features, thus facilitating the categorization of the elements present in a remotely sensed scene. However, the acquisition of VHR DSMs is quite expensive and time-consuming. Although there have been some efforts to create public DSM databases [7], the availability of pairs made up of optical image data and their corresponding VHR DSMs is still limited.

In order to overcome these limitations, several algorithms have been developed for the retrieval of DSMs from VHR optical images. Of particular importance are those based on deep learning techniques [8], which have aroused great interest in the remote sensing community [9]–[11] due to their flexibility to implement different neural architectures and learning modes. In particular, the convolutional neural network (CNN) has become the current state of the art in this domain, due to its high generalization power and performance when extracting mid- and high-level abstract features. The CNN has been successfully employed to estimate DSMs from single monocular-optical images [12]–[14]. However, the CNN works as a discriminative model, that is, as a mapping function $f(\cdot, \theta_f)$ with trainable parameters θ_f and maps the original input data $\mathbf{x} \in \mathcal{X}$ to some desired output $\mathbf{y} \in \mathcal{Y}$, learning conditional distributions $p(\mathbf{y}|\mathbf{x})$ by minimizing a loss function (i.e., modeling the decision boundary). This generally demands a huge effort to design an effective loss function, resulting in overfitting and blurring problems. In contrast, generative approaches such as variational autoencoders (VAEs) [15] and generative adversarial networks (GANs) [16] model the data distributions by learning the joint probability $p(\mathbf{x}, \mathbf{y})$, generating new samples instead of evaluating the available ones. In particular, GANs are able to map a complex data distribution $p_{\mathcal{I}}$ from a low-dimensional latent space $p_{\mathcal{Z}}$ by optimizing a loss function that recognizes whether the data are real or generated. In this sense, the GAN is a very interesting deep learning model for the generation of DSMs. For instance, in the pioneering work by Ghamisi and Yokoya [17], a conditional GAN (cGAN) [18], [19] was implemented with

Manuscript received December 1, 2019; revised March 17, 2020 and May 5, 2020; accepted May 21, 2020. This work was supported in part by Spanish Ministry under Grant FPU15/02090, in part by the Junta de Extremadura under Grant GR18060, and in part by European Union under Grant 734541-EXPOSURE. *(Corresponding author: M. E. Paoletti.)*

M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza are with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, 10003 Cáceres, Spain (e-mail: mpaoletti@unex.es).

P. Ghamisi is with the Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, 09599 Freiberg, Germany.

N. Yokoya is with the RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan.

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2020.2997295

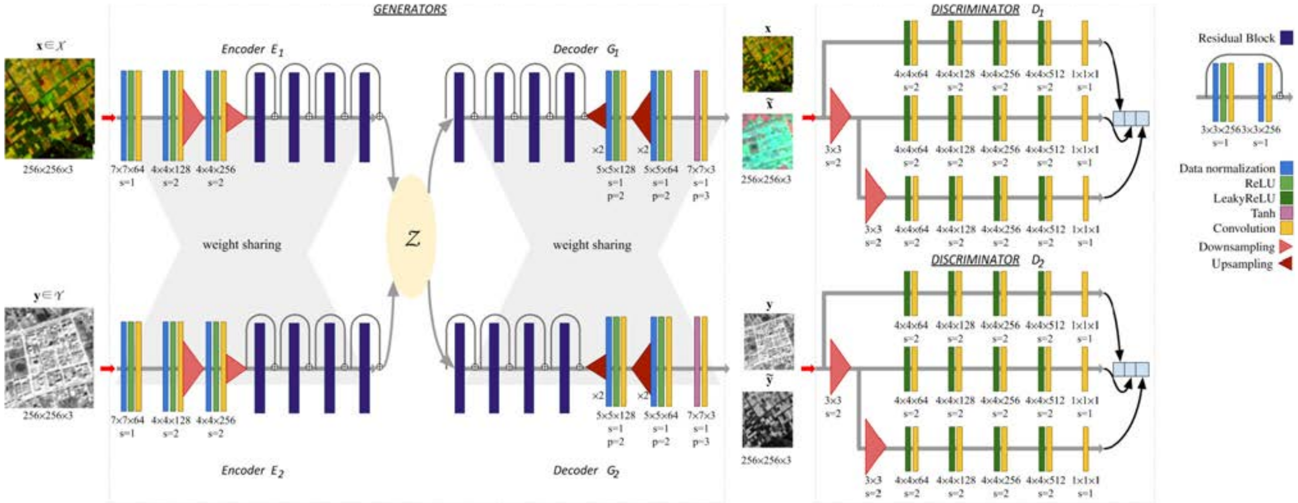

Fig. 1. Proposed network topology. Both branches receive data patches of $256 \times 256 \times 3$, being the DSM data replicated until obtaining three spectral bands. The spatial-downsampling has been made (using a factor of two in E_1 and E_2) by setting stride $= 2$ in the convolutional layers, while in D_1 and D_2 average pooling is implemented, also with a factor of 2.

DSM data as additional information during the training stage to simulate the desired elevation information from a single optical image. However, despite its good performance during reconstruction and classification, this model needs aligned pairs of corresponding optical-DSM data, which may not always be readily available in real scenarios.

In order to overcome the aforementioned limitation, this letter presents a new unpaired image-to-image translation (UNIT) method, based on GANs and VAEs, for automatic DSM generation from optical images. The UNIT problem is a hard and ill-posed one, where an infinite number of joint distributions can reach the marginal distributions of $\mathbf{x}$ and $\mathbf{y}$. In this context, additional assumptions must be adopted in order to correctly infer the information about the joint distribution [20]. In particular, we adopt the shared-latent space assumption of Coupled GANs (CoGANs) [21], where the corresponding data $\mathbf{x}$ and $\mathbf{y}$ are mapped into the same latent code $\mathbf{z} \in \mathcal{Z}$, from which the DSM images are generated by implementing a two-branch VAE-GAN architecture.

II. Proposed Methodology

The traditional GAN [16] is a generative mapping function that mimics a data distribution $p_{\mathcal{I}}$ from a random noise vector $\mathbf{z} \in \mathcal{Z}$ with prior $p_{\mathcal{Z}}(\mathbf{z})$, following an adversarial process where two neural models are simultaneously trained: 1) the *generative model* $\mathcal{G}(\cdot, \theta_{\mathcal{G}}) : \mathcal{Z} \rightarrow \mathcal{I}$, which tries to learn the data distribution by approximating $p_{\mathcal{G}} \approx p_{\mathcal{I}}$ by adjusting its trainable parameters $\theta_{\mathcal{G}}$ and 2) the *discriminative model* $\mathcal{D}(\cdot, \theta_{\mathcal{D}}) : \mathcal{I}$ or $\mathcal{G} \rightarrow \{0, 1\}$, which obtains the probability that a sample belongs to $p_{\mathcal{I}}$ or $p_{\mathcal{G}}$ by adjusting its parameters $\theta_{\mathcal{D}}$. In this context, the GAN follows a two-player `minimax` game with a value function $V(\mathcal{D}, \mathcal{G})$ in order to learn the original data distribution $p_{\mathcal{G}} = p_{\mathcal{I}}$

$$\min_{\mathcal{G}} \max_{\mathcal{D}} V(\mathcal{D}, \mathcal{G}) = \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{I}}(\mathbf{x})}[\log \mathcal{D}(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathcal{Z}}(\mathbf{z})}[\log(1 - \mathcal{D}(\mathcal{G}(\mathbf{z})))]. \tag{1}$$

The ultimate goal of GANs is to train $\mathcal{G}$ in order to maximize $\mathcal{D}(\mathcal{G}(\mathbf{z}))$, so that $\mathcal{D}$ is unable to differentiate between the original and artificially generated samples.

Usually, the GAN-based model for image-to-image translation takes an image from one domain $\mathbf{x} \in \mathcal{X}$ as input of $\mathcal{G}$ (instead of the fixed-size noise vector $\mathbf{z}$) and outputs the corresponding image in the target domain $\mathbf{y} \in \mathcal{Y}$, employing a $\mathcal{G}$ with an encoder–decoder architecture and skip connections. Also, $\mathcal{D}$ usually returns a matrix of values instead of the traditional $[0, 1]$ in order to preserve the image details [17]. In this sense, the goal of $\mathcal{G}$ is to learn the joint distribution $p_{\mathcal{G}} \approx p_{\mathcal{X},\mathcal{Y}}$ that relates each image $\mathbf{x}$ in domain $\mathcal{X}$ with its counterpart $\mathbf{y}$ in domain $\mathcal{Y}$. From an unpaired point of view, we assume that there are no pairs of $\{\mathbf{x}, \mathbf{y}\}$ available to train the model, where $\mathbf{x}$ and $\mathbf{y}$ exhibit their corresponding and independent marginal distributions $p_{\mathcal{X}}(\mathbf{x})$ and $p_{\mathcal{Y}}(\mathbf{y})$. In this context, there are infinite solutions (i.e., joint distributions) that can yield the marginal distributions, this being an inherently ambiguous and ill-posed problem.

To address this issue and further develop our U-IMG2DSM method, we combine the weight-sharing constraint of CoGANs [21] with the latent-enconding of VAE-GAN [22] to adopt a shared-latent space assumption [20]. The proposed network architecture is graphically illustrated in Fig. 1. It is composed of six subnetworks: two encoder–decoder pairs (E_1, G_1 and E_2, G_2, respectively) of two VAE-generators and two adversarial discriminators (D_1 and D_2). Focusing on the VAE-generators, for any pair of (not necessarily correlated) images $\mathbf{x}$ and $\mathbf{y}$ introduced into the encoder-branches, there is a shared-latent representation $\mathbf{z} \in \mathcal{Z}$, being $\mathbf{z} = E_1(\mathbf{x}) = E_2(\mathbf{y})$, from which the two images are recovered by the decoder-branches, that is, $\mathbf{x} = G_1(E_1(\mathbf{x})) = G_1(\mathbf{z} \sim Q_1(\mathbf{z}|\mathbf{x}))$ and $\mathbf{y} = G_2(E_2(\mathbf{y})) = G_2(\mathbf{z} \sim Q_2(\mathbf{z}|\mathbf{y}))$, where both Q_* follows a normal distribution $\mathcal{N}(\mathbf{z}|E_{\mu,*}(\cdot), I)$ being $E_{\mu,*}(\cdot)$ the encoder's output mean vector. In this sense, our goal is to train both generators to learn the mapping functions $F_{\mathbf{x}\rightarrow\mathbf{y}}$ and $F_{\mathbf{y}\rightarrow\mathbf{x}}$ in such a way that each one translates the images given from one domain to another, that is, providing their corresponding pairs in the destination domains

$$\begin{aligned} \hat{\mathbf{x}} &= F_{\mathbf{y}\rightarrow\mathbf{x}}(\mathbf{y}) = G_1(E_2(\mathbf{y})) = G_1(\mathbf{z} \sim Q_2(\mathbf{z}|\mathbf{y})) \\ \hat{\mathbf{y}} &= F_{\mathbf{x}\rightarrow\mathbf{y}}(\mathbf{x}) = G_2(E_1(\mathbf{x})) = G_2(\mathbf{z} \sim Q_1(\mathbf{z}|\mathbf{x})) \end{aligned} \tag{2}$$

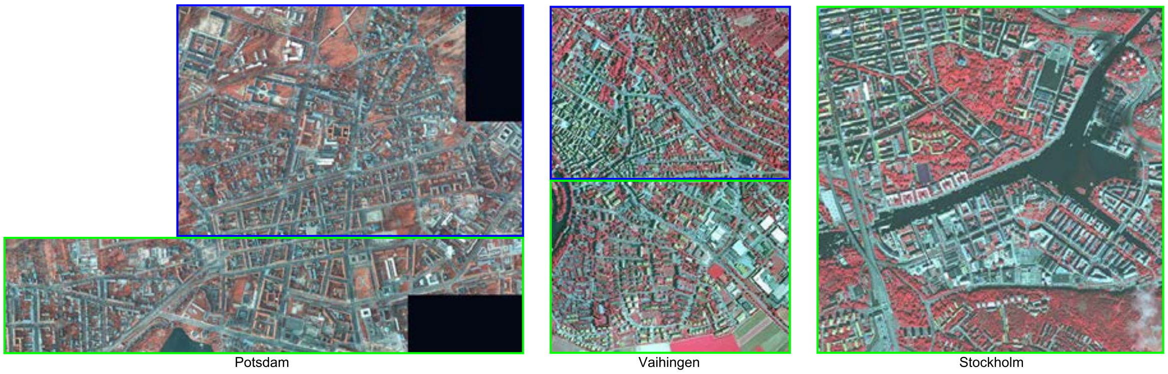

Fig. 2. IRRG images for Potsdam, Vaihingen, and Stockholm. Blue and green rectangles indicate training and test areas, respectively.

where $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are the corresponding pairs of $\mathbf{y}$ and $\mathbf{x}$ in the target domains, respectively, creating two image translation streams. In this sense, $G_1 : \mathcal{Z} \rightarrow \mathcal{X}$ mimics the distribution $p_{\mathcal{X}}$, while $G_2 : \mathcal{Z} \rightarrow \mathcal{Y}$ mimics $p_{\mathcal{Y}}$. Regarding the discriminators, D_1 distinguishes between real $\mathcal{X}$-domain images and images generated by G_1, while D_2 discriminates between real $\mathcal{Y}$-domain images and images generated by G_2, following a three-branch structure to exploit multiscale features.

To correctly adjust the model's parameters, a three-component loss given by (3) is designed during training: 1) $\mathcal{L}_{re_1}(E_1, G_1)$ and $\mathcal{L}_{re_2}(E_2, G_2)$ are the image-reconstruction losses, where the Kullback–Leibler divergence indicates the deviation between the distribution of the obtained latent code $\mathbf{z}$ and its prior distribution $p_\eta(\mathbf{z}) = \mathcal{N}(\mathbf{z}|0, I)$; 2) $\mathcal{L}_{tr_1}(E_2, G_1, D_1)$ and $\mathcal{L}_{tr_2}(E_1, G_2, D_2)$ are the image-translation losses (based on the cGAN's function) to reinforce the learning of the target domains; and iii) $\mathcal{L}_{cc_1}(E_1, G_1, E_2, G_2)$ and $\mathcal{L}_{cc_2}(E_2, G_2, E_1, G_1)$ are the cycle reconstruction losses, in the sense that an image translated twice needs to resemble the input one, that is, $\mathbf{x} = F_{\mathbf{y}\rightarrow\mathbf{x}}(F_{\mathbf{x}\rightarrow\mathbf{y}}(\mathbf{x}))$ and $\mathbf{y} = F_{\mathbf{x}\rightarrow\mathbf{y}}(F_{\mathbf{y}\rightarrow\mathbf{x}}(\mathbf{y}))$, while the obtained latent code $\mathbf{z}$ is constrained by its prior distribution. $\lambda_0 = 1$, $\lambda_1 = \lambda_3 = 0.01$, and $\lambda_2 = \lambda_4 = 10$ are some hyperparameters needed to control the weight of each element. This `minimax` problem is solved as two-player zero-sum game, following the gradient update scheme of [16]. The parameters of our model have been initialized with the Kaiming method [23] and optimized by the ADAM algorithm with parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$, a learning rate of $1 \cdot 10^{-4}$, and weight decay of $1 \cdot 10^{-4}$. Although VAE-generators' optimizers work on (3), the discriminators' optimizer minimizes the least squares loss between the translated and the original data. The maximum number of epochs has been set to 1750, using a batch size of two images (one per image domain), and 400 training images. The data have been scaled into the [0, 1] range, with mirrored borders.

III. Experimental Results

A. Data Sets

Three remotely sensed images (over the cities of Potsdam, Vaihingen, and Stockholm) have been considered. The Potsdam and Vaihingen images were provided by the ISPRS Working Group II/4,[1] while the Stockholm image was provided by DigitalGlobe[2] and captured by WorldView-2. **Potsdam** is composed of 38 tiles of orthophotographs with four bands [near-infrared response (NIR), red, green, and blue]. Its DSM has 6000 × 6000 pixels at ground sampling distance (GSD) of 5 cm. The second data set, **Vaihingen**, comprises an orthophotograph with three bands (NIR, red, and green), while its DSM is composed of 2000 × 2889 pixels at GSD of 50 cm. Finally, **Stockholm** comprises multispectral and panchromatic images. In this case, the target area is of 4000 × 4000 pixels at GSD of 50 cm. In order to avoid the spatial overlap between the training and test samples, Potsdam and Vaihingen data sets have been spatially divided into two subimages, while Stockholm is only used in the test stage. In particular, the training set consists of 400 samples with 256 × 256 pixels (i.e., 200 from Potsdam and 200 from Vaihingen), while the test set comprises 800 samples (200 from Potsdam, 200 from Vaihingen, and 400 from Stockholm). Fig. 2 shows the considered training and test data, using blue and green rectangles, respectively.

[1] http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html

[2] https://www.digitalglobe.com/resources/product-samples

$$
\begin{aligned}
&\min_{E_1,E_2,G_1,G_2}\ \max_{D_1,D_2}\ \mathcal{L}_{re_1}(E_1, G_1)+\mathcal{L}_{tr_1}(E_2, G_1, D_1)+\mathcal{L}_{cc_1}(E_1, G_1, E_2, G_2)+\mathcal{L}_{re_2}(E_2, G_2)+\mathcal{L}_{tr_2}(E_1, G_2, D_2)+\mathcal{L}_{cc_2}(E_2, G_2, E_1, G_1)\\
&\quad \text{where}\mathcal{L}_{re_1}(E_1, G_1) = \lambda_1 \mathrm{KL}(Q_1(\mathbf{z}|\mathbf{x})||p_\eta(\mathbf{z})) - \lambda_2 \mathbb{E}_{\mathbf{z}\sim Q_1(\mathbf{z}|\mathbf{x})}[\log p_{G_1}(\mathbf{x}|\mathbf{z})]\\
&\quad\ \text{and } \mathcal{L}_{re_2}(E_2, G_2) = \lambda_1 \mathrm{KL}(Q_2(\mathbf{z}|\mathbf{y})||p_\eta(\mathbf{z})) - \lambda_2 \mathbb{E}_{\mathbf{z}\sim Q_2(\mathbf{z}|\mathbf{y})}[\log p_{G_2}(\mathbf{y}|\mathbf{z})]\\
&\quad \text{where}\mathcal{L}_{tr_1}(E_2, G_1, D_1) = \lambda_0 \mathbb{E}_{\mathbf{x}\sim p_{\mathcal{X}}}[\log D_1(\mathbf{x})] + \lambda_0 \mathbb{E}_{\mathbf{z}\sim Q_2(\mathbf{z}|\mathbf{y})}[\log(1 - D_1(G_1(\mathbf{z})))]\\
&\quad\ \text{and } \mathcal{L}_{tr_2}(E_1, G_2, D_2)=\lambda_0 \mathbb{E}_{\mathbf{y}\sim p_{\mathcal{Y}}}[\log D_2(\mathbf{y})]+\lambda_0 \mathbb{E}_{\mathbf{z}\sim Q_1(\mathbf{z}|\mathbf{x})}[\log(1-D_2(G_2(\mathbf{z})))]\\
&\quad \text{where } \mathcal{L}_{cc_1}(E_1, G_1, E_2, G_2) = \lambda_3 \mathrm{KL}(Q_1(\mathbf{z}|\mathbf{x})||p_\eta(\mathbf{z})) + \lambda_3 \mathrm{KL}(Q_2(\mathbf{z}|\hat{\mathbf{y}})||p_\eta(\mathbf{z})) - \lambda_4 \mathbb{E}_{\mathbf{z}\sim Q_2(\mathbf{z}|\hat{\mathbf{y}})}[\log p_{G_1}(\mathbf{x}|\mathbf{z})]\\
&\quad\ \text{and } \mathcal{L}_{cc_2}(E_2, G_2, E_1, G_1) = \lambda_3 \mathrm{KL}(Q_2(\mathbf{z}|\mathbf{y})||p_\eta(\mathbf{z})) + \lambda_3 \mathrm{KL}(Q_1(\mathbf{z}|\hat{\mathbf{x}})||p_\eta(\mathbf{z})) - \lambda_4 \mathbb{E}_{\mathbf{z}\sim Q_1(\mathbf{z}|\hat{\mathbf{x}})}[\log p_{G_2}(\mathbf{y}|\mathbf{z})] \qquad (3)
\end{aligned}
$$

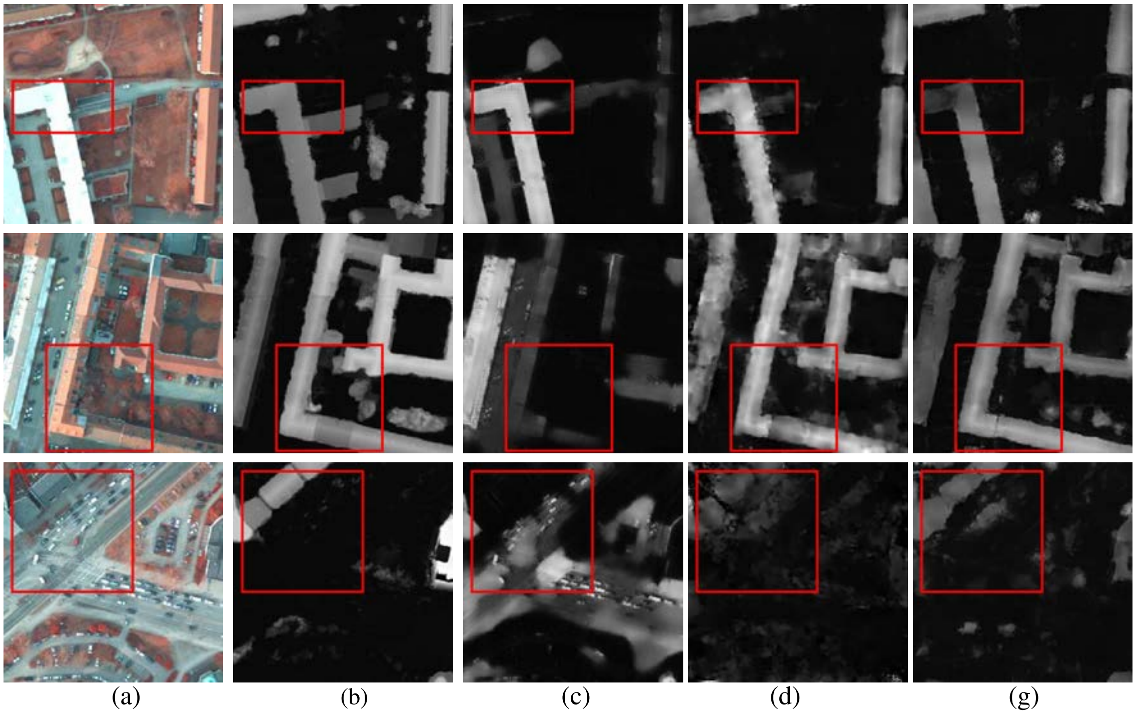

Fig. 3. Results of elevation data simulation (obtained by different unpaired and paired methods) as compared with the ground truth for Potsdam. (a) RGB. (b) Ground-Truth DSM. (c) CycleGAN. (d) U-IMG2DSM. (g) IMG2DSM.

TABLE I

COMPARISON BETWEEN THE PROPOSED (UNPAIRED) U-IMG2DSM AND THE (PAIRED) IMG2DSM IN [17]

		CycleGAN	U-IMG2DSM	IMG2DSM
Potsdam	RMSE	6.79±1.05	4.62±1.24	3.89±0.11
	ZNCC	0.477±0.180	0.740±0.124	0.718±0.008
Vaihingen		CycleGAN	U-IMG2DSM	IMG2DSM
	RMSE	5.86±0.57	3.46±0.78	2.58±0.09
	ZNCC	0.485±0.110	0.808±0.073	0.813±0.072
Stockholm		CycleGAN	U-IMG2DSM	IMG2DSM
	RMSE	5.75±0.96	3.93±0.79	3.66±0.23
	ZNCC	0.375±0.191	0.555±0.129	0.524±0.136
Parameters (Millions)		28.30	38.82	57.18
Training time (hours)		24.85	16.62	2.65

Inspired by Ghamisi and Yokoya [17], infrared, red, and green (IRRG) images have been employed, unifying the GSD at 50 cm. Also, consistent with [17], we considered two scenarios in our experiments to evaluate the performance of the proposed approach. In the first scenario (which applies only to the Potsdam and Vaihingen data sets), the training and test data were selected from the same data sets with spatially separated areas, as shown in Fig. 2. In the second scenario, we took one step further and selected the training and test data from different cities (imaged with entirely different data acquisition platforms) to carefully investigate the generalization ability and transferability of the method. In particular, we trained with Potsdam and Vaihingen, and used Stockholm for testing.

B. Experimental Settings

Our experiments have been conducted on a 6th Generation Intel Core i7-6700K processor with 8M of cache and up to 4.20 GHz (4 cores/8 way multitask processing), an NVIDIA GeForce GTX 1080 graphics processing unit (GPU) with 8-GB GDDR5X of video memory and 10 Gbps of memory frequency, 40 GB of DDR4 RAM with a serial speed of 2400 MHz, a Toshiba DT01ACA HDD with 7200 RPM and 2 TB, and an ASUS Z170 pro-gaming motherboard. Regarding our software environment, it is composed of Ubuntu 18.04.2×64 as the operating system, CUDA 9 and cuDNN 7.0.5, PyTorch framework [24] and Python 2.7.15 as the programming language.

In order to evaluate the quality of generated DSMs, two widely used metrics have been used: the root-mean-square-error (RMSE), which provides information on the degree of absolute error at each pixel in the unit of meters, and the zero-mean normalized cross correlation (ZNCC), which quantifies the spatial correlation between the output and ground truth. In addition, we consider the overall accuracy (OA), average accuracy (AA), and kappa coefficient (K) of the classification of the resulting products.

C. Experimental Discussion

Two experiments have been conducted to evaluate the proposed U-IMG2DSM model. The first one tests the performance of the developed model (as compared to the paired IMG2DSM in [17]) when generating the corresponding DSMs, while the second experiment test the reliability of the obtained DSMs by performing a classification task.

Table I shows a comparison between the proposed (unpaired) U-IMG2DSM and the paired IMG2DSM in [17]. In this experiment, the training has been made using approximately half of the Potsdam and Vaihingen data sets, testing with the other half of these images (first scenario) and the full Stockholm image (second scenario) to generate the corresponding DSMs. Thus, the first scenario tests the performance of the proposed method when training and test data follow the same distribution (as they belong to the same data set), while

TABLE II

CLASSIFICATION RESULTS WITH DIFFERENT METHODS FOR POTSDAM

			SIMULATED DSM			
			Unsupervised		Supervised	
Algorithm	Metric	IRRG	CycleGAN	U-IMG2DSM	IMG2DSM	IRRG+DSM
RF	OA	56.09	60.16	67.65	68.35	78.12
	AA	46.46	50.26	56.17	56.54	66.75
	K (×100)	40.8	46.13	56.25	57.2	70.46
MLP	OA	61.88	62.99	69.99	70.95	80.71
	AA	50.63	51.88	58.79	59.54	70.18
	K (×100)	48.38	49.83	59.37	60.66	73.88

in the second scenario the training and test samples follow different distributions, as they belong to different scenes. Although the IMG2DSM outperforms the proposed method in terms of RMSE, ZNCC and runtime, our U-IMG2DSM is fully unpaired and does not need any image-DSM pairs in advance, which represents an advantage over the paired model. Also, it requires fewer parameters than IMG2DSM to achieve very close results, also being faster than CycleGAN. In addition, Fig. 3 shows that the DSMs obtained by the proposed U-IMG2DSM for the Potsdam data set (used here as an example) are visually competitive with those obtained for the same image by the IMG2DSM, reducing the confusion between impervious surfaces and buildings, as indicated by the red framed areas. This suggests the potential of the proposed (fully unpaired) method to distinguish land covers that are similar in spectral characteristics but different in elevation, without using any prior information.

Table II shows the classification results obtained by the random forest (RF) and multilayer perceptron (MLP) for the Potsdam image. Although the best results are obtained using the IRRG image combined with the ground-truth DSM, the table shows that, when simulated DSMs are used in addition to the IRRG images, the classification accuracy can also significantly increase (regardless of the considered classifier). In fact, the proposed U-IMG2DSM performs better than the (unpaired) CycleGAN and close to the (paired) IMG2DSM. These results suggest that our new unpaired approach produces appropriate elevation information for classification purposes without any prior knowledge. These results suggest that our new unpaired approach produces appropriate elevation information for classification purposes, without any prior knowledge.

IV. CONCLUSION

This letter introduces U-IMG2DSM, a new fully unpaired model (based on VAEs and GANs) for the automatic generation of DSMs from optical images. Its architecture is composed of six subnetworks—two encoder–decoder pairs of two VAE-generators and two adversarial discriminators. Our experimental results, conducted using data sets collected over different cities by different data acquisition platforms, indicate that our U-IMG2DSM outperforms unpaired approaches (e.g., CycleGAN) and performs very close to paired approaches (e.g., IMG2DSM) in terms of RMSE, ZNCC, classification accuracy, and visual interpretation (without using any prior knowledge).

REFERENCES

[1] A. A. Aldosari and K. Jacobsen, "Quality of height models covering large areas," *PFG-J. Photogramm., Remote Sens. Geoinf. Sci.*, vol. 87, pp. 177–190, Oct. 2019.

[2] E. P. Baltsavias, "A comparison between photogrammetry and laser scanning," *ISPRS J. Photogramm. Remote Sens.*, vol. 54, nos. 2–3, pp. 83–94, Jul. 1999.

[3] K. Nurminen, M. Karjalainen, X. Yu, J. Hyyppä, and E. Honkavaara, "Performance of dense digital surface models based on image matching in the estimation of plot-level forest variables," *ISPRS J. Photogramm. Remote Sens.*, vol. 83, pp. 104–115, Sep. 2013.

[4] C. Beumier and M. Idrissa, "Digital terrain models derived from digital surface model uniform regions in urban areas," *Int. J. Remote Sens.*, vol. 37, no. 15, pp. 3477–3493, Aug. 2016.

[5] M. Bisson, B. Behncke, A. Fornaciai, and M. Neri, "LiDAR-based digital terrain analysis of an area exposed to the risk of lava flow invasion: The zafferana Etnea territory, Mt. Etna (Italy)," *Natural Hazards*, vol. 50, no. 2, pp. 321–334, Aug. 2009.

[6] S. van Beijma, A. Comber, and A. Lamb, "Random forest classification of salt marsh vegetation habitats using quad-polarimetric airborne SAR, elevation and optical RS data," *Remote Sens. Environ.*, vol. 149, pp. 118–129, Jun. 2014.

[7] G. Jorayev, K. Wehr, A. Benito-Calvo, J. Njau, and I. de la Torre, "Imaging and photogrammetry models of Olduvai gorge (Tanzania) by unmanned aerial vehicles: A high-resolution digital database for research and conservation of early stone age sites," *J. Archaeological Sci.*, vol. 75, pp. 40–56, Nov. 2016.

[8] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, Apr. 2016.

[9] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.)*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[10] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[11] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 279–317, Dec. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271619302187

[12] S. Srivastava, M. Volpi, and D. Tuia, "Joint height estimation and semantic labeling of monocular aerial images with CNNS," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 5173–5176.

[13] L. Mou and X. Xiang Zhu, "IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," 2018, *arXiv:1802.10249*. [Online]. Available: http://arxiv.org/abs/1802.10249

[14] H. A. Amirkolaee and H. Arefi, "Convolutional neural network architecture for digital surface model estimation from single remote sensing image," *J. Appl. Remote Sens.*, vol. 13, no. 1, 2019, Art. no. 016522.

[15] Y. Pu *et al.*, "Variational autoencoder for deep learning of images, labels and captions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2352–2360.

[16] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2014, pp. 2672–2680.

[17] P. Ghamisi and N. Yokoya, "IMG2DSM: Height simulation from single imagery using conditional generative adversarial net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 794–798, May 2018.

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

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

[20] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 700–708.

[21] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 469–477.

[22] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," 2015, *arXiv:1512.09300*. [Online]. Available: http://arxiv.org/abs/1512.09300

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[24] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *Proc. NIPS Autodiff Workshop*, 2017, pp. 1–4.