IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 19, 2021, accepted October 14, 2021, date of publication October 26, 2021, date of current version November 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3122894

Height Prediction and Refinement From Aerial Images With Semantic and Geometric Guidance

MAHDI ELHOUSNI, ZIMING ZHANG, AND XINMING HUANG, (Senior Member, IEEE)
Department of Electrical and Computer Engineering, Worcester Polytechnic Institute (WPI), Worcester, MA 01609, USA

Corresponding author: Xinming Huang (xhuang@wpi.edu)

This work was supported in part by U.S. NSF under Grant CCF-2006738 and in part by The MathWorks Fellowship.

ABSTRACT Deep learning provides a powerful new approach to many computer vision tasks. Height prediction from aerial images is one of those tasks which benefited greatly from the deployment of deep learning, thus replacing traditional multi-view geometry techniques. This manuscript proposes a two-stage approach to solve this task, where the first stage is a multi-task neural network whose main branch is used to predict the height map resulting from a single RGB aerial input image, while being augmented with semantic and geometric information from two additional branches. The second stage is a refinement step, where a denoising autoencoder is used to correct some errors in the first stage prediction results, producing a more accurate height map. Experiments on two publicly available datasets show that the proposed method is able to outperform state-of-the-art computer vision based and deep learning-based height prediction methods. Code is publicly available at: https://github.com/melhousni/DSMNet.

INDEX TERMS UAV, height, DSM, CNN, autoencoders, multi-task.

I. INTRODUCTION

Aerial imagery analysis was known as a very tedious task owing to the low quality of the acquired images and the lack of some appropriate automated process that could extract the relevant information from the data. Fortunately, recent advances in computer vision have made it possible to directly extract predefined patterns from the images, by applying some carefully designed algorithms. Moreover, deep learning brings in a new revolution to the field of aerial imagery analysis with more intelligence and better accuracy. As a result, multiple deep learning challenges related to aerial imagery processing, such as semantic segmentation [1], [2] and object detection [3], [4], have been routinely featured each year by the geoscience and remote sensing (GRSS) community [5]–[7].

This work focuses on the height prediction task that is to predict and reconstruct the corresponding height map, or in other words, predict the height value for every pixel in the input aerial image. Predicting such height maps can be very useful in the subsequent task of 3D reconstruction. By obtaining the accurate height of each building or structure appearing in the input images, 3D models can be generated as an accurate representation of the surrounding world. These 3D models are crucial for GPS-denied navigation, or other fields such as urban planning or telecommunications. Theses reconstructions are traditionally done using Structure from

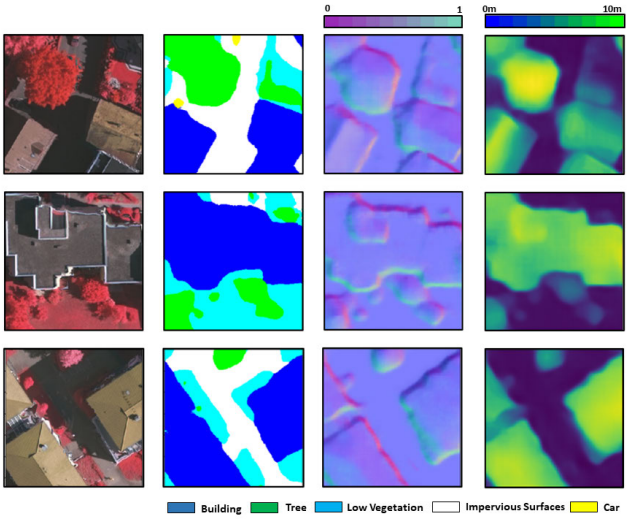

FIGURE 1. The outputs of our multi-task network. From left to right: The input RGB image, the output semantic labels, surface normals and height predictions.

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Shariq Imran.

Motion (SfM) [8], [9] technique with stereo camera rigs, which can be very sensible to noise and changes in lighting condition.

For the task of height prediction from aerial images, we propose a multi-task learning framework where additional branches are introduced to improve height prediction accuracy. Previous works have showed that multi-task learning helps improving the accuracy of height prediction networks by including semantic labels [10]. We propose to add a third branch to the multi-task network which will be devoted to predicting the surface normals, as shown on Fig. 1. In this configuration, the main height prediction branch will have access to both semantic and geometric guidance, improving the results of the height prediction network.

However, since the input is only an aerial image, our predictions sometimes can be noisy due to artefacts such as shadows or unexpected changes in color. Therefore, we introduce a refinement network which is a denoising autoencoder taking the outputs from the prediction network, removing the noise present in the prediction and producing a higher quality and more accurate height map. By combining these two steps, we are able to produce results that surpass the current state-of-the-art on multiple datasets. We are also able to produce reasonable semantic labels and surface normal predictions without additional optimizations.

In summary, our contributions in this work are the following:

- We propose a triple-branch multi-task learning network, including semantic label, surface normal and height prediction.
- We introduce a denoising autoencoder as a refinement step for the final height prediction results.
- We achieve state-of-the-art performance on two publicly available datasets, and an extensive ablation study shows the importance of each step in the 3D reconstruction pipeline.
- We show through two applications how our height prediction pipeline can be used to reconstruct dense 3D point clouds with semantic labels.

II. RELATED WORK

A. MULTI-TASK LEARNING

This learning framework aims at optimizing a single neural network that can predict multiple related outputs, each represented by a task-specific loss function [11]. Lately, this approach has become increasingly popular, especially in the area of autonomous driving cars, where multiple outputs (such as object detection, semantic segmentation, motion classification) are derived simultaneously from the input of camera images [12], [13].

B. HEIGHT PREDICTION FROM AERIAL IMAGES

This task has received a considerable amount of attention by the deep learning and remote sensing communities, especially after the use of UAVs to collect aerial images has become widely accessible. The goal here is to generate a height value for each pixel in an input aerial image. In works such as [14]–[16], deep learning methods such as residual networks, skip connections and generative adversarial networks are leveraged in order to predict the expected height maps.

Other works such as [10], [17] proposed to reformulate the task as a multi-learning problem, by introducing neural networks capable of predicting both the height maps and the semantic labels simultaneously. These works showed that both outputs can benefit from each other, during the simultaneous optimization process of the multi-task network. We choose to extend that formulation by including a third branch in our network tasked for predicting surface normals, which was inspired by previous works [18], [19] in the depth prediction task for autonomous driving cars. Surface normals are also known to be extremely useful during 3D reconstruction tasks and are required for surface and mesh reconstruction algorithms such as the Poisson surface reconstruction algorithm [20] or the Ball pivoting algorithm [21].

C. DENOISING AUTOENCODERS

Removing noise from images is a traditional task in computer vision. Over the years, many techniques were presented in the literature which can be broadly divided into two categories [22]: spatial filtering methods and variational denoising methods. The spatial filtering methods can either be linear, such as mean filtering [23] or Wiener filtering [24], [25], or nonlinear such as median filtering [26] or bilateral filtering [27]. These filtering methods work reasonably well but are limited. If the noise level becomes too high, these methods tend to lead to over-smoothing of the edges that are present in the image. On the other hand, in variational denoising methods, an energy function is defined and minimized to remove the noise, based on image priors or the noise-free images. Some popular variational denoising methods include total variation regularization [28], non-local regularization [29] and low-rank minimization [30].

Lately, a new trend based on deep learning autoencoders has shown great potential on image denoising. Autoencoder is a class of popular neural networks that has shown to be very powerful across multiple tasks such as segmentation of medical imagery [31], decoding the semantic meaning of words [32] or solving facial recognition challenges [33]. For our task, the most useful type of autoencoders available in the literature is the denoising autoencoder. As shown in [34], autoencoders can be trained to remove noise from an arbitrary input signal such as an image. We propose to use denoising autoencoder to refine the height predictions from the multi-task learning network.

III. METHOD

A. PROBLEM SETUP

Our main objective is to predict an accurate height map using only a monocular aerial image as input. We attempt to do so by constructing a two-stage pipeline, where two different

networks are cascaded in serial. The first stage of our pipeline is a multi-task learning network, where the main branch is tasked with predicting preliminary height images, aided by semantic and surface normal information that was extracted by two additional branches of the neural network. The second stage can be seen as a denoising autoencoder: All the predictions from the multi-task network are concatenated and fed into the autoencoder, in order to deal with noisy areas remaining in the height results from the first stage. This effectively produces sharper images that are closer to the ground truth. An overview of the full pipeline can be seen in Fig. 3.

Fundamentally, the height prediction task is a non-linear regression problem that can be formulated as:

$$\min_{\psi \in \Psi} \sum_{i} \ell(\mathbf{y}_i, \psi(\mathbf{x}_i)) \tag{1}$$

where $\psi : \mathcal{X} \rightarrow \mathcal{Y}$ denotes the height prediction mapping function from the feasible space Ψ, $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ denotes a loss function such as the least-square, $\mathbf{x}_i$ is the input aerial image and y_i is the output height map.

Predicting height only using a single branch neural network is possible. However, previous works such as [10], [17] showed that including additional branches to predict other related information such as segmentation labels can be beneficial for both tasks. In our case, in addition to predicting the height maps, we also predict semantic labels and surface normals, which provide semantic and geometric guidance by augmenting the main height prediction branch with information from the semantics and surface normal branches. More details can be found in the height prediction section below. Hence, our ψ function can now be defined as:

$$\psi(\mathbf{x}_i) = \{\mathbf{P}_h, \mathbf{P}_s, \mathbf{P}_n\} \tag{2}$$

where $\mathbf{P}_h$, $\mathbf{P}_s$ and $\mathbf{P}_n$ are the height, semantic and surface normal predictions respectively, that are trying to approximate $\mathbf{y}_i = \{\mathbf{P}^*_h, \mathbf{P}^*_s, \mathbf{P}^*_n\}$ where $\mathbf{P}^*_h$, $\mathbf{P}^*_s$ and $\mathbf{P}^*_n$ are the height, semantic and surface normal ground truth respectively. Finding a good approximation of the ψ function can be seen as the first stage in our proposed method.

Regression problems such as the one we are facing are difficult to solve due to the high number of values expected to be predicted. This makes our height prediction $\mathbf{P}_h$ noisy by definition, so the use of denoising autoencoders is appropriate in this situation.

First, we can write: $\mathbf{P}_h = \mathbf{P'}_h + e$ where $\mathbf{P'}_h$ is the clean height value, and e the noise inherent to our approximation of the function ψ. By introducing a denoising autoencoder, we can approximate the noise function γ such as $\mathbf{P}_h = \mathbf{P'}_h + \gamma(\mathbf{z_i})$, where $\mathbf{z_i}$ is the concatenation of the outputs of ψ with the input aerial image x_i. This makes it possible to re-write equations (2) as $\psi(\mathbf{x}_i) = \{\mathbf{P'}_h + \gamma(\mathbf{z_i}), \mathbf{P}_s, \mathbf{P}_n\}$. We can also now define the objective of the second stage of our method such as:

$$\min_{\gamma \in \Gamma} \sum_{i} \ell(\mathbf{P}^*_h, \mathbf{P}_h - \gamma(\mathbf{z_i})) \tag{3}$$

In this paper, our goal is to approximate both function ψ and γ by using two cascaded deep neural networks.

B. HEIGHT PREDICTION NETWORK

We solve the height prediction problem via multi-task learning where, in addition to the main height prediction, semantic and surface normals predictions are conducted too. We found that by re-routing the information in the semantic and surface normal branches to the main height branch, our neural network can learn to predict more accurate height values, especially around the edges.

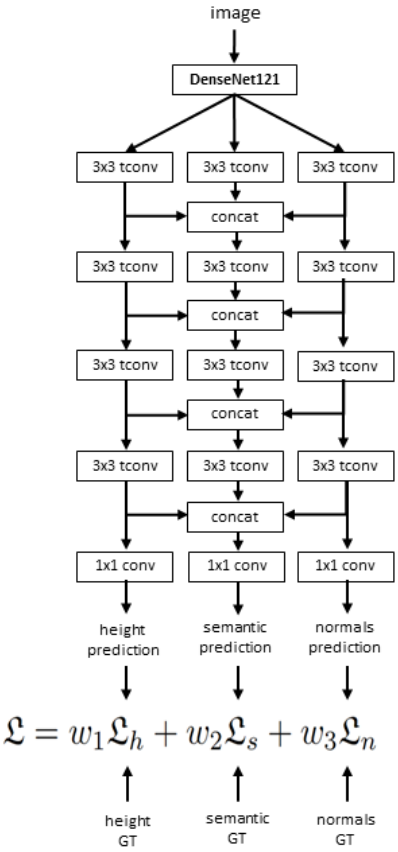

FIGURE 2. Architecture of our multi-task learning network for height, semantic and surface normals predictions. Note that each tconv block is followed by the ReLu function and drop out layers are inserted after each tconv layers in the main height prediction branch.

Fig. 2 shows our multi-task learning network architecture. We propose a convolutional neural network where we combine a pretrained encoder (tasked with extracting relevant features from the input aerial images), with three inter-connected decoder branches, one for each type of predictions respectively. We chose to use a DenseNet121 network, pretrained on ImageNet, as our main encoder. We show later in the experimentation section that DenseNet121 yields the best accuracy when compared to other popular architectures. Our decoders on the other hand is inspired by [35] and are characterized by being able to reconstruct the expected predictions efficiently. We list in Table 1 the different layers that we used.

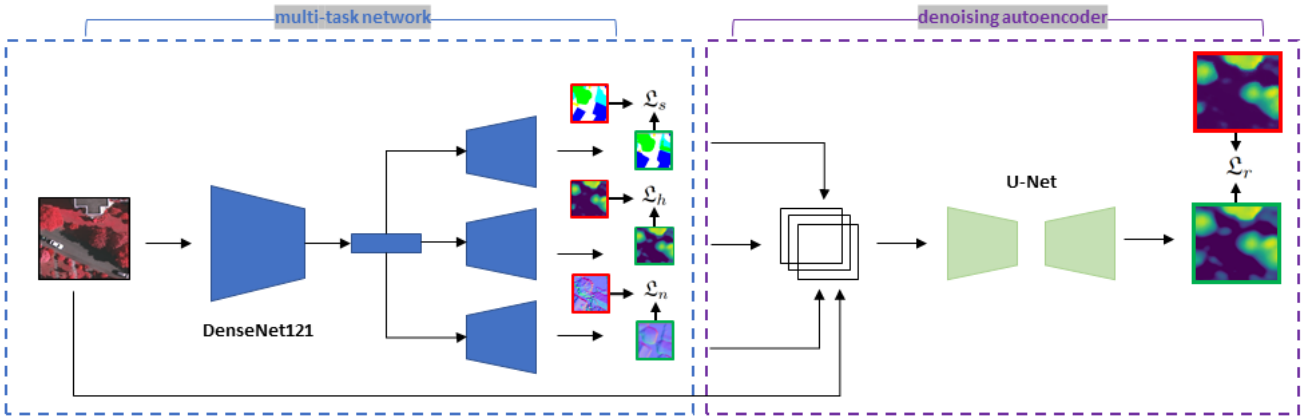

FIGURE 3. Our two stage height prediction and refinement pipeline. We use DenseNet121 to extract a global feature vector from the input aerial images, which is used to predict the normals map, semantic labels and a first guess at the height map (first stage, in blue). These results are concatenated with the input aerial image and fed into a denoising autoencoder to generate the refined final height map (second stage, in purple). Red boxes represent the ground truth, while green ones represent the networks predictions.

This network is optimized by using a multi-objective loss function defined as:

$$\mathfrak{L} = w_1\mathfrak{L}_h + w_2\mathfrak{L}_s + w_3\mathfrak{L}_n \tag{4}$$

where $\mathfrak{L}_h = \frac{1}{n}\sum_{i=1}^{n}(P_h - P_h^*)^2$, $\mathfrak{L}_s = -\frac{1}{n}\sum_{i=1}^{n} P_s^* log(P_s)$, $\mathfrak{L}_n = \frac{1}{n}\sum_{i=1}^{n}(P_n - P_n^*)^2$ and w_1, w_2 and w_3 are weights set up according to the training dataset and the scale of each loss function: We found that by using weights that keep all the loss functions at the same scale, the CNN would converge faster and achieve higher final accuracy levels.

C. HEIGHT REFINEMENT NETWORK

As mentioned previously, the height prediction map $\mathbf{P}_h$ produced by the multi-task learning network still contains some noisy areas that must be refined in order to generate the final height prediction $\mathbf{P'}_h$. We introduce an autoencoder to estimate the noise and produce more accurate height map predictions.

We choose the popular U-Net architecture [31] as network structure. The input of the network is the concatenation of the multi-task network outputs $\mathbf{P}_h$, $\mathbf{P}_s$ and $\mathbf{P}_n$ with the aerial image $\mathbf{x}_i$, as shown in Fig. 3. Details of the different layers forming the denoising network are listed in Table 2. The loss function used to optimize this network is the mean square error between the refined height map and the ground truth: $\mathfrak{L}_r = \frac{1}{n}\sum_{i=1}^{n}(P'_h - P_h^*)^2 = \frac{1}{n}\sum_{i=1}^{n}(P_h - \gamma - P_h^*)^2$, with γ being the noise function defined in Eq. 3.

IV. EXPERIMENTS

A. DATASETS

2018 DFC [36] dataset was released during the 2018 Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society. It was collected over the city of Houston, which contains multiple optical resources geared toward urban machine learning tasks such multispectral LiDAR, hyperspectral imaging, Very High-Resolution (VHR) imagery and semantic labels. Using the results of the multispectral LiDAR, it is possible to obtain Digital Structural Models (DSM) and Digital Elevation Models (DEM), which, if subtracted from one another, produces height maps that we can use as ground truth. Four tiles of data are used for training while ten tiles are used for testing.

ISPRS Vaihingen [37] dataset was released during the semantic labeling contest of ISPRS WG III/4. It was collected over the city of Vaihingen, Germany and consists of very high resolution true ortho photo (TOP) tiles, corresponding Digital Surface Models (DSM) and semantic labels. As it is usually done when dealing with this dataset, we use the normalized DSM (nDSM) produced by [38] as ground truth for our height prediction. Sixteen tiles were used for training while seventeen tiles are used for testing.

TABLE 1. Height prediction network details.

	Layer	Output Size
Encoder	DenseNet121	(10,10,1024)
Decoder	$DeConv_1$	(20,20,1024)
	$Concat$	(20,20,3072)
	$Conv_{11}$	(20,20,1024)
	$Conv_{12}$	(20,20,1024)
	$DeConv_2$	(40,40,512)
	$Concat$	(40,40,1536)
	$Conv_{21}$	(40,40,512)
	$Conv_{22}$	(40,40,512)
	$DeConv_3$	(80,80,256)
	$Concat$	(80,80,768)
	$Conv_{31}$	(80,80,256)
	$Conv_{32}$	(80,80,256)
	$DeConv_4$	(160,160,64)
	$Concat$	(160,160,192)
	$Conv_{41}$	(160,160,64)
	$Conv_{42}$	(160,160,64)
	$DeConv_5$	(320,320,32)
	$Concat$	(320,320,96)
	$Conv_{51}$	(320,320,32)
	$Conv_{52}$	(320,320,32)
	$Conv_{out}$	(320,320,1)

TABLE 2. Height refinement network details.

	Layer	Output Size
Encoder	$Conv_1$	(320,320,64)
	$MaxPooling$	(160,160,64)
	$Conv_2$	(160,160,128)
	$MaxPooling$	(80,80,128)
	$Conv_3$	(80,80,256)
	$MaxPooling$	(40,40,256)
	$Conv_4$	(40,40,512)
	$MaxPooling$	(20,20,512)
	$Conv_5$	(20,20,1024)
Decoder	Upsampling	(40,40,512)
	$Concat$	(40,40,1024)
	$Conv_6$	(40,40,512)
	$Upsampling$	(80,80,256)
	$Concat$	(80,80,512)
	$Conv_7$	(80,80,256)
	$Upsampling$	(160,160,128)
	$Concat$	(160,160,256)
	$Conv_8$	(160,160,128)
	$Upsampling$	(320,320,64)
	$Concat$	(320,320,128)
	$Conv_8$	(320,320,64)
	$Conv_{out}$	(320,320,1)

1) SURFACE NORMAL MAPS

The surface normal maps for both dataset are generated using the given height maps, following practices usually used for surface normal estimation from dense depth maps based on the Sobel operator [39]. The details are listed in Alg 1.

Algorithm 1: Surface Normals Generation

Input : Height map P_h
Output: Surface normals map P_n
$zx \leftarrow Sobel(P_h, 0)$
$zy \leftarrow Sobel(P_h, 1)$
$N \leftarrow stack(-zx, -zy, 1)$
$P_n \leftarrow \frac{N/\|N\|}{2} + 1$
return P_n

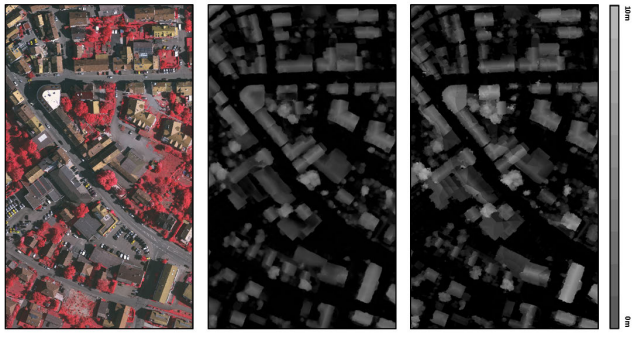

FIGURE 4. Qualitative comparison of a reconstructed tile from the testing dataset. From left to right: The input RGB tile, the height prediction and the height ground truth.

B. NETWORK TRAINING AND RESULTS

1) TRAINING

Our training process is not end-to-end. Instead, we follow a two stages approach: we first remove the denoising autoencoder and only focus on training the multi-task network. To do so, random 320 × 320 crops are sampled from the aerial tiles and corresponding semantic, surface normals and height ground truth are used for training. Once the multi-task network converges, we freeze its weights and then plug into the denoising autoencoder to obtain the final height predictions. We train this second network following the same random sampling process used to train the first one. We use Tensorflow [40], a learning rate of 0.0002, a batch size of 64, the Adam optimizer [41] and a single RTX2080Ti to train both stages. During training, we saw that altering the network's hyper parameters can sometimes have a slight effect of the convergence speed, but no significant effect on the final accuracy level.

Note that in the case of the DFC2018 dataset, the input VHR aerial tiles are ten times bigger than their corresponding DSM, DEM and semantic labels. To deal with that, we first down sample the aerial tiles ten times before starting to collect training crops.

2) RESULTS

The aerial tiles were reconstructed using a sliding window of the same size as of the training samples and with a constant step size. We use Gaussian smoothing to deal with overlapping areas. This makes it possible to deal with cases where different crops of the same area produce different height values, while also protecting the final result from the "checkerboard effect". We report the results of our height prediction and refinement pipeline on both datasets in Table 3, where we use the mean square error (MSE), the mean absolute error (MAE) and root-mean-square error (RMSE) as metrics, all in meters. We also show a qualitative comparison in Fig. 4. When comparing with previous proposed methods in the literature, we can see that by using our multi-task network combined with the refinement step, we are able to surpass the state-of-the-art performance across all metrics on both datasets, with improvement up to 25%.

We credit this increase in accuracy to multiple factors. Firstly, the choice of our encoder (in this case DenseNet121), which is capable of extracting features that are relevant to this task. The second is the context information brought by our 2 additional branches in the multi-task prediction network. Knowing if a pixel falls on a building rather than the road, in addition to the orientation of its associated surface normal vector, helps the network predict height values better. Finally, the denoising autoencoder helps us deal with certain artefacts that tend to confuse the prediction network. We provide numerical analysis of these observations in the ablation study.

It is also interesting to note that we are able to achieve similar scores to methods which were trained on the high-definition aerial tiles directly without any down sampling as shown in Table 4. For reconstruction of the same sized area, such networks would take much longer processing time and significantly more computing resources than our proposed method.

TABLE 3. Comparison with other height prediction methods on the ISPRS Vaihingen and the 2018 DFC datasets in meters.

	ISPRS Vaihingen			2018 DFC		
Method	MSE	MAE	RMSE	MSE	MAE	RMSE
Ours	**0.0042**	**0.036**	**0.062**	**6.92**	**1.37**	**2.57**
Carvalho [10]	0.0060	0.045	0.074	9.34	1.53	2.97
Srivastava [17]	-	0.063	0.098	-	-	-
IMG2DSM [15]	-	-	0.090	-	-	-

TABLE 4. Comparison with method trained on VHR aerial images.

Method	MSE	MAE	RMSE	Time (s)	Input Resolution
Ours	**6.92**	1.37	**2.57**	72	1192x1202
Carvalho VHR [10]	7.27	**1.26**	2.59	774	11920x12020

Missing values in Table 3 were not reported by the cited publications. We also exclude the results reported by [16] because it did not follow the same training/testing split of the data.

C. SEMANTIC LABEL AND SURFACE NORMAL PREDICTIONS

Although this work does not focus on the semantic label and surface normal predictions and only uses them to improve the height predictions, we share the results of those two branches and compare them with available methods in the literature in Table 5. Our results in Table 5 show that our multi-task network is able to produce semantic label results that are comparable with the state of the art on the Vaihingen dataset and acceptable ones on the DFC2018 (which has 20 classes compared to the 6 of the Vaihingen dataset). We use the following metrics for the semantic segmentation: The overall accuracy (OA), defined as the sum of accuracies for each class predicted, divided by the number of class, the average accuracy (AA), defined as the number of correctly predicted pixels, divided by the total of pixels to predict and Cohen's coefficient (Kappa), which is defined as $\mathfrak{Kappa} = \frac{p_0 - p_e}{1 - p_e}$, such as p_e is the probability of the network classifying a pixel correctly and p_0 is the probability of the pixel being correctly classified by chance. The network is also able to produce meaningful surface normal maps as seen on Fig. 1. Missing values in Table 5 were not reported by the cited publications.

TABLE 5. Semantic labels and surface normals results on the ISPRS Vaihingen and the 2018 DFC datasets.

	ISPRS Vaihingen			2018 DFC		
	Semantic Labels					
Method	OA	AA	Kappa	OA	AA	Kappa
Ours	85.6	74.8	**80.1**	51.89	47.01	49
Carvalho [10]	**87.7**	**85.4**	75.9	**64.70**	**58.85**	**63**
Srivastava [17]	78.8	73.4	71.9	-	-	-
Cerra [42]	-	-	-	58.60	55.60	56
Fusion-FCN [43]	-	-	-	63.28	-	61
	Surface Normals					
Method	MSE	MAE	RMSE	MSE	MAE	RMSE
Ours	0.0115	0.0642	0.1066	0.0620	0.2119	0.2572

D. ABLATION STUDY

1) HEIGHT REFINEMENT

To demonstrate the usefulness of the aforementioned refinement network, we test our method with and without the denoising autoencoder, on both datasets. In Table 6, we compare the results obtained after both experiments and show that the refinement step always produces more accurate height maps, resulting in an increase of up to 16% in accuracy. By combining the information present in the semantic and surface normal inputs with the initial guess of the height produced by the previous network, the refinement network is able to concentrate on noisy areas where the height values are abnormal and fix them automatically. In addition, we compare our deep learning based denoiser with other popular non-learning denoising algorithms such as Bilateral Filtering (BF) [27] and Non-local Means (NIM) regularization [29].

TABLE 6. Comparison of our height prediction methods with and without refinement, on the ISPRS Vaihingen and the 2018 DFC datasets in meters.

	ISPRS Vaihingen			2018 DFC		
Method	MSE	MAE	RMSE	MSE	MAE	RMSE
multi-task only	0.0045	0.043	0.065	7.36	1.50	2.64
multi-task + BF	0.0046	0.043	0.065	7.27	1.51	2.62
multi-task + NIM	0.0045	0.043	0.065	7.34	1.48	2.63
multi-task + Unet	**0.0042**	**0.036**	**0.062**	**6.92**	**1.37**	**2.57**

We also show qualitatively on Fig. 5 that the refinement height maps are much closer to the ground truth and contains less noise than the direct output of the multi-task network.

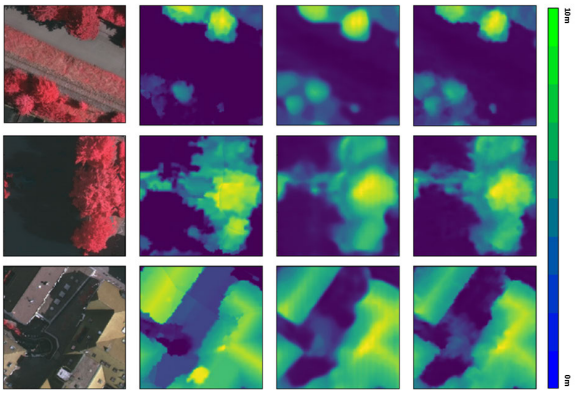

FIGURE 5. Qualitative comparison. From left to right: The input RGB image, the height prediction of our multi-task network, the refined height map of our denoising autoencoder and the ground truth.

2) CHOOSING THE RIGHT ENCODER

Our network structure for height prediction is generic, since any off-the-shelf encoder can be used in the first stage to extract features from the input aerial image.

However, we show in Table 7 that DenseNet121 outperforms other popular encoder structures and produces the most accurate height maps. This is owing to the fact that DenseNet121 is much deeper than the other two networks and contains a higher number of skip connections between layers, making it possible to extract much finer features from the input image. All the networks are trained for the same number of epochs and using the same hyper parameters, such that it

TABLE 7. Encoder comparison on the DFC2018 dataset in meters.

Encoder	MSE	MAE	RMSE
ResNet101 [44]	18.95	3.33	4.19
VGG19 [45]	8.57	1.87	2.85
DenseNet121 [46]	**7.36**	**1.50**	**2.64**

ensures the fairness when comparing both the convergence speed and accuracy scores.

3) GEOMETRIC AND SEMANTIC GUIDANCE

In this section, we show the effect of the geometric and semantic guidance in our method in both height prediction and height refinement stages. First, we show in Table 8 that using a multi-task network instead of a single task one improves the overall height prediction results. We also show in Table 9 that by concatenating all the results of the first stage as the input to the denoising autoencoder, we are able to generate more accurate and refined results compared to only using the height image as input. This shows that the semantic and geometric context information brought by two additional branches assist in producing more accurate height values.

TABLE 8. Comparison of height prediction results of single and multi-task networks in meters.

	ISPRS Vaihingen			2018 DFC		
Method	MSE	MAE	RMSE	MSE	MAE	RMSE
single-task	0.0048	0.046	0.067	8.17	1.64	2.78
multi-task	**0.0045**	**0.043**	**0.065**	**7.36**	**1.50**	**2.64**

TABLE 9. Comparison of height refinement results of single and multi-input denoiser in meters.

	ISPRS Vaihingen			2018 DFC		
Method	MSE	MAE	RMSE	MSE	MAE	RMSE
single-input	0.0043	0.037	0.063	7.13	1.47	2.62
multi-input	**0.0042**	**0.036**	**0.062**	**6.92**	**1.37**	**2.57**

TABLE 10. Comparison of our reconstruction results (meters) based on the step size (pixels).

	ISPRS Vaihingen			2018 DFC		
Step	MSE	MAE	RMSE	MSE	MAE	RMSE
80	0.00421	0.0363	0.0625	6.98	1.38	2.58
60	**0.00420**	**0.0362**	**0.0623**	**6.92**	**1.37**	**2.57**
40	0.00421	0.0362	0.0623	6.93	1.37	2.58

4) FINDING THE RIGHT RECONSTRUCTION STEP

The accuracy of our final tile reconstruction depends also on the step size of the sliding window that we choose when collecting the aerial crops. We show in Table 10 the different results corresponding to different step sizes. We found that a step size of 60 pixels results the best across both datasets.

5) VISUALIZING THE UNCERTAINTY

In order to investigate the performance of our pipeline more thoroughly, we generate uncertainty maps according to the method proposed in [47]. The results are displayed in Fig. 6

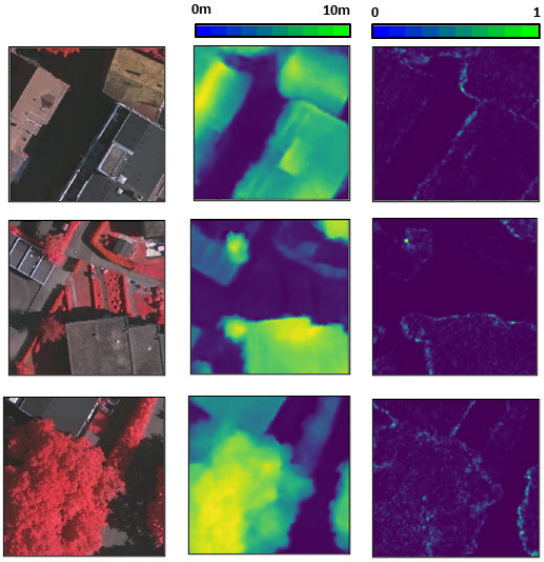

FIGURE 6. Uncertainty results. From left to right RGB Image, Height Prediction, Uncertainty Map. Prediction errors are mostly concentrated around the edges.

and show that most of the prediction errors can be attributed to the areas such as the edges of buildings due to the sudden changes in brightness and color, and trees where shadows introduce a significant amount of color noise.

V. APPLICATIONS FOR 3D RECONSTRUCTION

In this section, we propose two applications to show how to take advantage of the results generated by our proposed pipeline. The first is 3D reconstruction of select buildings from a single aerial image. In the second application, we simulate a UAV flight over a certain area and show that we can reconstruct the entire 3D area by combining odometry and aerial images. In comparison to the classic SfM algorithm, our method provides a significant gain in speed, accuracy and density. More importantly, our proposed method requires significantly less number of images since only minimal overlaps are necessary when taking the aerial shots.

A. SINGLE AERIAL IMAGE 3D RECONSTRUCTION

Usually, in order to reconstruct the 3D shape of a building, multiple shots from multiple angles with significant overlap are necessary in order to apply the sequential surface from motion algorithm. We show in Fig. 7(b) that owing to our multi-task network, we are able to produce accurate 3D point clouds of the buildings using a single image only.

The proposed method is also capable of generating semantic point clouds in Fig. 7(c) and 3D meshes of buildings and their surrounding areas in Fig. 7(d) by leveraging the semantic labels and surface normals generated by the networks. Specifically, semantic point clouds are generated by projecting the semantic labels onto the point clouds, while the meshes are generated by combining the surface normals with the reconstructed point clouds using the ball pivoting algorithm [21].

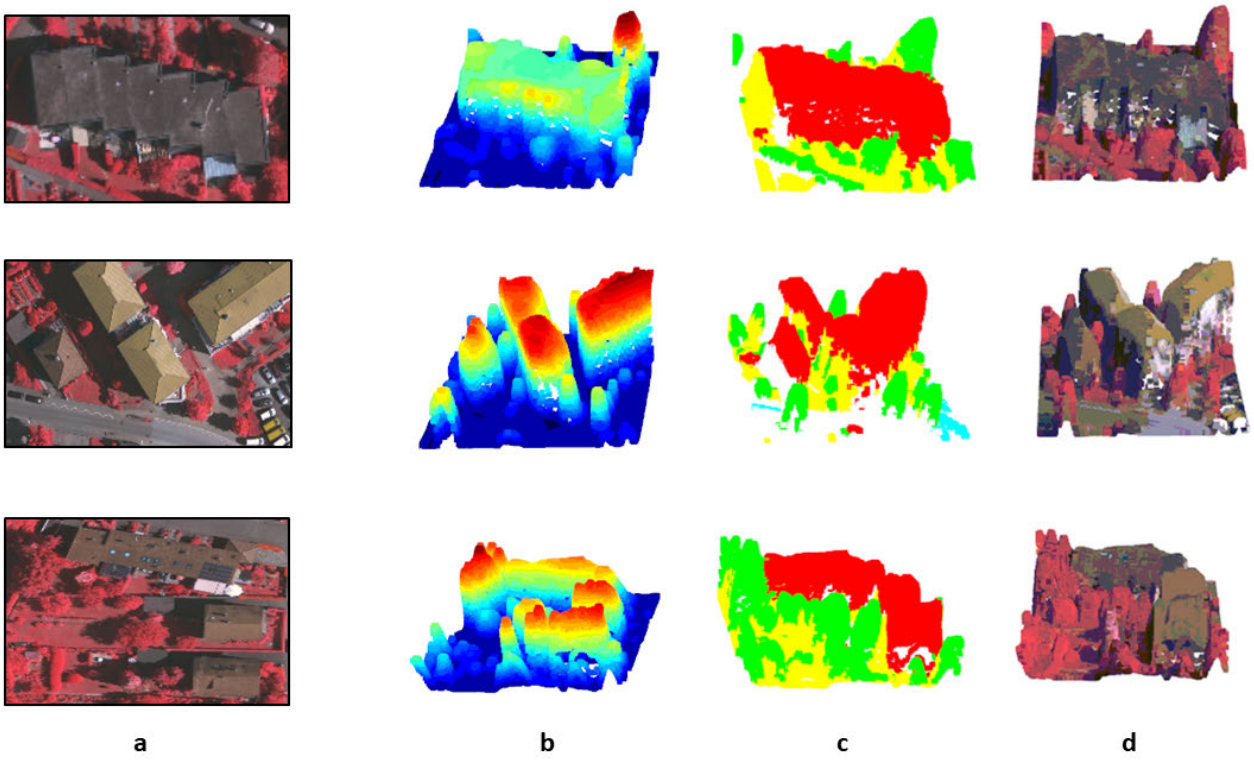

FIGURE 7. 3D reconstructions using a single image. (a) RGB Image, (b) Height Colorized Pointcloud, (c) Semantic Pointcloud, (d) RGB Colorized Mesh.

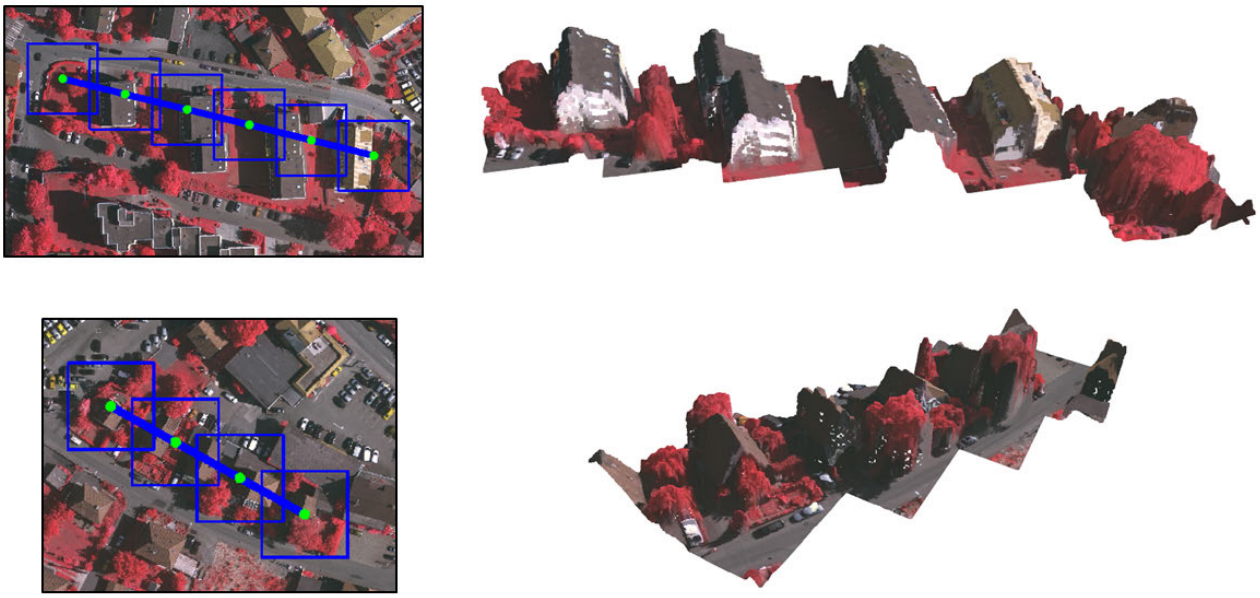

FIGURE 8. 3D reconstructions from simulated UAV flight. From left to right: Positions of the UAV images, Reconstructed 3D scene.

B. AREA RECONSTRUCTION WITH SIMULATED UAV FLIGHT

3D reconstruction of urban areas is a very useful application. Similarly to what we mentioned in the first application, reconstructing an entire area would generally require a series of captured images with significant overlaps, by flying the drones in multiple passes over the same area, in order to generate a semi-dense point cloud.

In our case, we show in Fig. 8 that by using a single pass with a small number of captured images and minimal overlap

(only to avoid gaps in the final reconstruction) we are able to produce accurate and dense 3D reconstructions. We also note that when we feed the same data to an SfM algorithm, it typically leads to failures since only a small number of features can be matched among the single-pass aerial shots. The data is collected by simulating a constant altitude UAV flight over a certain neighborhood in one of the tiles available in the testing datasets. The odometry is assumed to be known from on-board IMU or GPS sensors.

VI. CONCLUSION

In this work, we propose a deep learning based two-stage pipeline that can predict and refine height maps from a single aerial image. We leverage the power of multi-task learning by designing a three-branch neural network for height, semantic label and surface normal predictions. We also introduce a denoising autoencoder to refine the predicted height maps and largely eliminate the noise remaining in the results of the first stage height prediction network. Experiments on two publicly available datasets show that our method is capable of outperforming state-of-the-art results in height prediction accuracy. In future work, we plan on exploring the computational efficiency of the proposed neural networks for their applications towards real-time processing of aerial images.

REFERENCES

[1] K. Chen, K. Fu, M. Yan, X. Gao, X. Sun, and X. Wei, "Semantic segmentation of aerial images with shuffling convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 173–177, Feb. 2018.

[2] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of CNSS," *ISPRS Ann. Photogram., Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 473–480, Jul. 2016.

[3] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2849–2858.

[4] I. Ševo and A. Avramović, "Convolutional neural network based automatic object detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 5, pp. 740–744, May 2016.

[5] B. Le Saux, N. Yokoya, R. Hansch, and S. Prasad, "2018 IEEE GRSS data fusion contest: Multimodal land use classification [technical committees]," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 1, pp. 52–54, Mar. 2018.

[6] B. Le Saux, N. Yokoya, R. Hansch, M. Brown, and G. Hager, "2019 data fusion contest [technical committees]," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 1, pp. 103–105, Mar. 2019.

[7] N. Yokoya, P. Ghamisi, R. Haensch, and M. Schmitt, "2020 IEEE GRSS data fusion contest: Global land cover mapping with weak supervision [Technical Committees]," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 1, pp. 154–157, Mar. 2020.

[8] P. Moulon, P. Monasse, and R. Marlet, "Adaptive structure from motion with a *Contrario* model estimation," in *Proc. Asian Conf. Comput. Vis.* Daejeon, South Korea: Springer, 2012, pp. 257–270.

[9] P. Moulon, P. Monasse, and R. Marlet, "Global fusion of relative motions for robust, accurate and scalable structure from motion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3248–3255.

[10] M. Carvalho, B. Le Saux, P. Trouve-Peloux, F. Champagnat, and A. Almansa, "Multitask learning of height and semantics from aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1391–1395, Aug. 2020.

[11] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[12] M. Teichmann, M. Weber, M. Zollner, R. Cipolla, and R. Urtasun, "MultiNet: Real-time joint semantic reasoning for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1013–1020.

[13] D. Zhou, J. Fang, X. Song, L. Liu, J. Yin, Y. Dai, H. Li, and R. Yang, "Joint 3D instance segmentation and object detection for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1839–1849.

[14] H. A. Amirkolaee and H. Arefi, "Height estimation from single aerial images using a deep convolutional encoder-decoder network," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 50–66, Mar. 2019.

[15] 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.

[16] C.-J. Liu, V. A. Krylov, P. Kane, G. Kavanagh, and R. Dahyot, "IM2ELEVATION: Building height estimation from single-view aerial imagery," *Remote Sens.*, vol. 12, no. 17, p. 2719, Aug. 2020.

[17] 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.

[18] T. Dharmasiri, A. Spek, and T. Drummond, "Joint prediction of depths, normals and surface curvature from RGB images using CNNs," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 1505–1512.

[19] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2650–2658.

[20] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proc. 4th Eurograph. Symp. Geometry Process.*, Cagliari, Italy, Jun. 2006.

[21] F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, and G. Taubin, "The ball-pivoting algorithm for surface reconstruction," *IEEE Trans. Vis. Comput. Graphics*, vol. 5, no. 4, pp. 349–359, Oct./Dec. 1999.

[22] L. Fan, F. Zhang, H. Fan, and C. Zhang, "Brief review of image denoising techniques," *Vis. Comput. for Ind., Biomed., Art*, vol. 2, no. 1, p. 7, Dec. 2019.

[23] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing Using MATLAB*. London, U.K.: Pearson, 2004.

[24] A. K. Jain, *Fundamentals of Digital Image Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 1989.

[25] J. Benesty, J. Chen, and Y. Huang, "Study of the widely linear Wiener filter for noise reduction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 205–208.

[26] I. Pitas and A. N. Venetsanopoulos, *Nonlinear Digital Filters: Principles and Applications*, vol. 84. NY, USA: Springer, 2013.

[27] S. Paris, P. Kornprobst, J. Tumblin, and F. Durand, *Bilateral Filtering: Theory and Applications*. New York, NY, USA: Now, 2009.

[28] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D, Nonlinear Phenomena*, vol. 60, nos. 1–4, pp. 259–268, 1992.

[29] G. Gilboa and S. Osher, "Nonlocal operators with applications to image processing," *Multiscale Model. Simul.*, vol. 7, no. 3, pp. 1005–1028, 2008.

[30] I. Markovsky and K. Usevich, *Low Rank Approximation*, vol. 139. London, U.K.: Springer, 2012.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2015, pp. 234–241.

[32] C.-Y. Liou, W.-C. Cheng, J.-W. Liou, and D.-R. Liou, "Autoencoder for words," *Neurocomputing*, vol. 139, pp. 84–96, Sep. 2014.

[33] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *Proc. Int. Conf. Artif. Neural Netw.* Espoo, Finland: Springer, 2011, pp. 44–51.

[34] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, 2010.

[35] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 239–248.

[36] Y. Xu, B. Du, L. Zhang, D. Cerra, M. Pato, E. Carmona, S. Prasad, N. Yokoya, R. Hänsch, and B. Le Saux, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.

[37] M. Cramer, "The DGPF-test on digital airborne camera evaluation–overview and test design," *Photogrammetrie-Fernerkundung-Geoinf.*, vol. 2010, no. 2, pp. 73–82, May 2010.

[38] M. Gerke, "Use of the stair vision library within the ISPRS 2D semantic labeling benchmark," Univ. Twent, Enschede, The Netherlands, Tech. Rep., Jan. 2015.

[39] I. Sobel, "An isotropic 3×3 image gradient operator," Stanford, CA, USA, Artif. Project 271-272, 1968.
[40] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*. [Online]. Available: http://arxiv.org/abs/1603.04467
[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[42] D. Cerra, M. Pato, E. Carmona, S. M. Azimi, J. Tian, R. Bahmanyar, F. Kurz, E. Vig, K. Bittner, C. Henry, P. d'Angelo, R. Müller, K. Alonso, P. Fischer, and P. Reinartz, "Combining deep and shallow neural networks with ad hoc detectors for the classification of complex multi-modal urban scenes," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 3856–3859.
[43] Y. Xu, B. Du, and L. Zhang, "Multi-source remote sensing data classification via fully convolutional networks and post-classification processing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 3852–3855.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[46] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[47] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," 2015, *arXiv:1511.02680*. [Online]. Available: http://arxiv.org/abs/1511.02680

MAHDI ELHOUSNI received the B.S. degree in computer science and the M.S. degree in embedded systems from the National School for Computer Science, Rabat, Morocco. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Worcester Polytechnic Institute, Worcester, MA, USA. His main research interests include computer vision, deep learning, and SLAM.

ZIMING ZHANG received the Ph.D. degree from Oxford Brookes University, U.K., in 2013, under the supervision of Prof. Philip H. S. Torr (now in the University of Oxford). He is currently an Assistant Professor with Worcester Polytechnic Institute. Before joining WPI, he was a Research Scientist at the Mitsubishi Electric Research Laboratories (MERL), from 2016 to 2019. Prior to that, he was a Research Assistant Professor with Boston University. His research interests include computer vision and machine learning, especially in object recognition/detection, data-efficient learning, namely zero-shot learning, and applications, namely person re-identification, deep learning, and optimization. He won the Research and Development 100 Award 2018. His works have appeared in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, CVPR, ICCV, ECCV, and NIPS. He serves as a Review Member/a PC Member for top conferences, namely CVPR, ICCV, NIPS, ICML, ICLR, AAAI, AISTATS, and IJCAI, and journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *IJCV*, and *JMLR*.

XINMING HUANG (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Virginia Tech, in 2001. He was a Technical Staff Member with the Wireless Advanced Technology Laboratory, Bell Labs of Lucent Technologies. Since 2006, he has been a Faculty Member with the Department of Electrical and Computer Engineering, Worcester Polytechnic Institute (WPI), where he is currently a Full Professor. His main research interests include the areas of circuits and systems, with an emphasis on reconfigurable computing, wireless communications, information security, computer vision, and machine learning.