Height Estimation From Single Aerial Images Using a Deep Ordinal Regression Network

Xiang Li, Mingyang Wang, and Yi Fang, *Member, IEEE*

***Abstract*— Understanding the 3-D geometric structure of the Earth's surface has been an active research topic in photogrammetry and remote sensing community for decades, serving as an essential building block for various applications such as 3-D digital city modeling, change detection, and city management. Previous research studies have extensively studied the problem of height estimation from aerial images based on stereo or multiview image matching. These methods require two or more images from different perspectives to reconstruct 3-D coordinates with camera information provided. In this letter, we deal with the ambiguous and unsolved problem of height estimation from a single aerial image. Driven by the great success of deep learning, especially deep convolutional neural networks (CNNs), some research studies have proposed to estimate height information from a single aerial image by training a deep CNN model with large-scale annotated data sets. These methods treat height estimation as a regression problem and directly use an encoder–decoder network to regress the height values. In this letter, we propose to divide height values into spacing-increasing intervals and transform the regression problem into an ordinal regression problem, using an ordinal loss for network training. To enable multiscale feature extraction, we further incorporate an Atrous Spatial Pyramid Pooling (ASPP) module to extract features from multiple dilated convolution layers. After that, a postprocessing technique is designed to transform the predicted height map of each patch into a seamless height map. Finally, we conduct extensive experiments on International Society for Photogrammetry and Remote Sensing (ISPRS) Vaihingen and Potsdam data sets. Experimental results demonstrate significantly better performance of our method compared to state-of-the-art methods.**

***Index Terms*— Aerial image, convolutional neural networks (CNNs), digital surface model (DSM), height estimation, ordinal regression.**

I. Introduction

DUE to the rapid advancement of sensor technologies and high-resolution earth observation platforms, it becomes possible to explore the fine-grained 3-D structure of ground objects from satellite and aerial images. Height estimation, as one of the key building blocks for digital surface modeling, has been a hot topic in photogrammetry and remote sensing communities for decades and plays a crucial role in various applications such as city management [1], [2] and disaster monitoring [3], [4], just to mention a few. It has also been proved to benefit many challenging remote sensing problems, such as semantic labeling [5], [6] and change detection [7], [8].

Regarding height estimation from aerial images, previous research studies mostly focus on methods based on stereo or multiview image matching. Notable methods include shape from motion [9], shape from stereo [10], and shape from focus [11]. These methods require two or more images from different perspectives to reconstruct 3-D coordinates with camera information provided. A straightforward question is: can we generate height from a single image? To achieve this goal, the desired model should learn to identify certain height cues such as object size, perspectives, atmospheric effects, occlusion, texture, and shading [12].

In recent years, with the advancement of depth cameras, such as Microsoft Kinect and ZED, many large-scale depth data sets have been collected and widely used in the computer vision field. Moreover, the advancement of deep learning, especially convolutional neural networks (CNNs) has made it possible for estimating depth from monocular images through training on large-scale data sets [12]. Monocular depth estimation has, therefore, become a hot topic in the computer vision field with wide applications in various 3-D related tasks [13].

In the remote sensing field, some recent works tried to adopt CNNs for height estimation from a single aerial or satellite image. Although monocular images do not include explicit 3-D information and monocular height estimation from single aerial images is an ill-posed problem in nature, it is possible to learn 3-D cues from single aerial images and perform height estimation. For example, Mou and Zhu [14] proposed a CNN-based method for height estimation from optical images. They employed a fully convolutional encoder–decoder architecture to regress the height maps from a single aerial image. A regression loss between the predicted height values and the real ones is used to supervise network training in an end-to-end manner. Some other research studies [15], [16] also adopted similar encoder–decoder architectures for height regression. Nevertheless, because the real-value regression problem is unbounded and hard to optimize, these methods suffer from slow convergence and sometimes lead to suboptimal solutions [17].

In this letter, instead of treating height estimation as a real-value regression problem, we divide height values into spacing-increasing intervals and transform the regression problem as an ordinal-based regression problem. By means of space-increasing height discretization strategy, smaller height values would get more precise representations, while larger height values get coarse representations. More specifically,

Manuscript received June 4, 2020; revised July 21, 2020; accepted August 18, 2020. Date of publication September 9, 2020; date of current version December 29, 2021. This work was supported by the Ecological Quality Meteorological Monitoring and Evaluation, Mountain Flood Geological Disaster Prevention Meteorological Guarantee Project 2020, Zhejiang Province Climate Center. *(Corresponding author: Yi Fang.)*

Xiang Li and Yi Fang are with the NYU Multimedia and Visual Computing Laboratory, NYU Abu Dhabi, United Arab Emirates, and also with the NYU Multimedia and Visual Computing Laboratory, NYU Tandon, New York, NY 10012 USA (e-mail: yfang@nyu.edu).

Mingyang Wang is with the Department of Electrical and Computer Engineering, New York University, Abu Dhabi, United Arab Emirates.

Digital Object Identifier 10.1109/LGRS.2020.3019252

we develop an encoder–decoder network where the encoder part extracts multiscale image features, and the decoder part transforms the feature embeddings into discrete height maps. To enable multiscale feature extraction, we further incorporate an atrous spatial pyramid pooling (ASPP) [18] module to extract features from multiple dilated convolution layers. Moreover, since our model is trained with small image patches due to memory constraints, a postprocessing technique is then used to transform the patched height predictions into a seamless height map. The main contributions of the proposed method can be summarized as follows:

1) This letter introduces a fully convolutional network to perform height estimation from single aerial images. Instead of directly regressing the height values, we proposed to divide height values into spacing-increasing intervals representations and our model predicts the probability distribution over all intervals.
2) An ordinal loss function is proposed by comparing the predicted and ground truth discrete height maps.
3) To enable multiscale feature extraction, we apply an ASPP module to aggregate features from multiple dilated convolution layers to enhance the height estimation performance.

II. Methods

A. Height Discretization

To divide a height interval $[a, b]$ into discrete classes, the most commonly used method is uniform discretization (UD), which means each subinterval represents an equal height range. However, as the height values going larger, the available information from a single image becomes less rich; meanwhile, the height estimation error of larger height values become larger. In this regard, using the UD will lead to an over-strengthened loss for larger height values. To address this issue, we introduce the so-called space-increasing height discretization strategy which uniformed divides a given height interval in the logarithm space so that the smaller height values would get more precise representations while larger height values get coarse representations. In this way, the training losses in areas with larger height values are down-reweighted and our height estimation network can thus perform more accurate height prediction in areas with relatively small height values.

Given a height interval $[a, b]$ to be discretized with K subintervals, our space-increasing discretization (SID) can be formulated as

$$\text{SID: } t_i = \log(a) + \log(b/a) * i/K \tag{1}$$

where K denotes the total number of subintervals and the output $t_i \in \{t_0, t_1, \ldots, t_K\}$ denotes the thresholds for height discretization. To avoid negative value of log function, we add a shift of 1 to both a and b, and then apply this discretization strategy to transform the ground truth height maps into discrete class representations $\mathcal{C}$.

B. Ordinal Loss

After transforming the real-value height maps $\mathcal{D}$ into discrete height maps $\mathcal{C}$, one can directly transform the height estimation problem into a classification problem. However, we note that our discrete height values have a well-defined order, i.e., larger height values have larger class representations. While traditional classification loss functions treat each class independently and ignore this order information. To address this issue, we regard the height estimation as an ordinal regression problem and formulate an ordinal regression loss to train network parameters. Experimental comparisons of these two loss functions can be found in Section III-E.

Given input aerial image $\mathcal{I}_i$, discrete ground truth height map $\mathcal{C}_i$, our ordinal loss $\mathcal{L}(\mathcal{I}_i, \mathcal{C}_i)$ is defined as

$$\begin{aligned}\mathcal{L}(\mathcal{I}_i, \mathcal{C}_i) &= -\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(\sum_{k=0}^{\mathcal{C}_i(w,h)-1}\log(P^k(w,h)) \right.\\ &\left. + \sum_{k=\mathcal{C}_i(w,h)}^{K-1}\log(1-P^k(w,h))\right)\end{aligned} \tag{2}$$

where h and w denote the pixel location, H and W denote the width and height of the input image, $\mathcal{C}_i(w, h)$ denotes the ground truth discrete class label of pixel location (w, h), and $P^k(w, h)$ denotes the probability of predicted height class is larger than k. Our ordinal loss is averaged across all pixel locations.

The predicted class probability $P^k(w, h)$ is defined as

$$P^k(w,h) = P(\hat{\mathcal{C}}_i(w,h) > k|\mathcal{I}_i, \theta) \tag{3}$$

where θ denotes network parameters. Given (2) and (3), our ordinal loss enforces the predicted probabilities to have larger values in the first k feature maps where k is smaller than the ground discrete class $\mathcal{C}_i(w, h)$, meanwhile it enforces predicted probabilities to have small values in the last $(K - \mathcal{C}_i(w, h))$ feature maps where k is larger than the ground truth discrete class. In this way, predictions far from the ground truth label $\mathcal{C}_i(w, h)$ will receive larger penalties than those close to $\mathcal{C}_i(w, h)$.

In the inference stage, our model predicts the height class map with $2K$ channels, we then transform the discrete height values into real-values $D' \in R^{W\times H}$. To achieve this, we average the height thresholds of predicted class and its successive class, calculated as

$$D'(w,h) = \frac{t_{d(w,h)} + t_{d(w,h)+1}}{2}. \tag{4}$$

C. Network Architecture

Fig. 1 gives an overview of the proposed method. In this letter, we use a ResNet network as our backbone to extract height-related cues from aerial images. The original ResNet network consists of four residual blocks. To maintain the spatial resolution and reconstruct high-resolution height maps, all convolution layers after block2 are replaced by dilated convolutions [19] with dilation set to 2. By doing so, all feature maps after block2 have a fixed spatial resolution of 1/8 of the input image. Moreover, in a dilated convolution, the convolutional layer has a larger receptive field which enables informative inputs from a larger area. When the

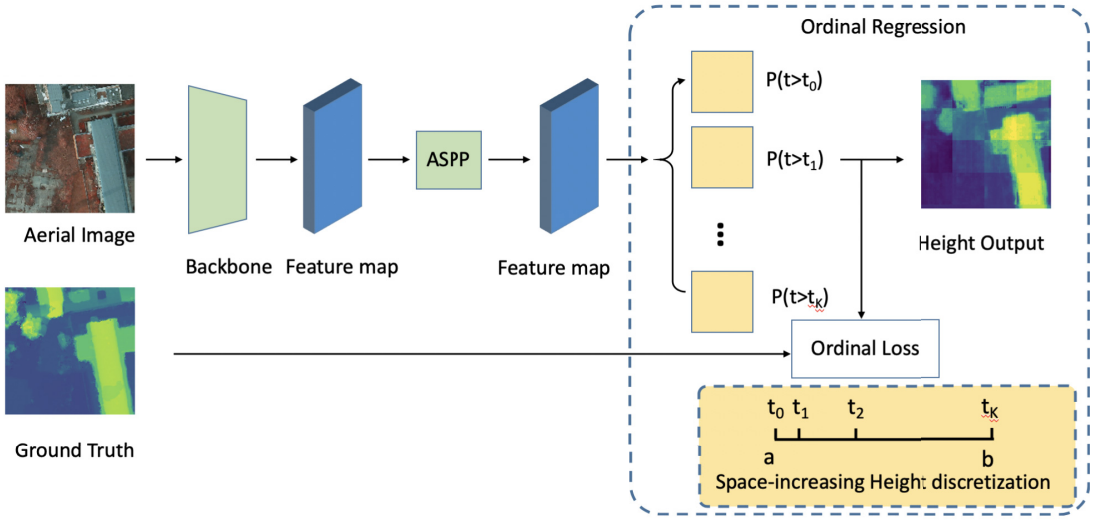

Fig. 1. Overview of the proposed method of height estimation from single aerial images. Our model starts with a deep CNN to extract high-level height-related feature maps. Then, an ASPP module is utilized to aggregate features from multiple enlarged receptive fields via dilated convolution layers. Finally, our model predicts the discrete height maps. Our model is optimized with an ordinal regression loss defined by comparing the predicted and ground truth height maps in an end-to-end fashion.

dilated rate equals 1, the dilated convolution becomes a vanilla convolutional layer.

To further enable multiscale feature learning, we apply an ASPP module to combine features at different scales and keep the same resolutions for them. The ASPP module is first proposed in Deeplab-v3 [18] for the problem of semantic segmentation. In this letter, our ASPP module contains one convolution layer with a kernel size of 3×3 and three dilated convolutional layers with a kernel size of 3×3 and a dilation rate of 6, 12, and 18, respectively.

After getting the multiscale features maps from the above ASPP module, we concatenate all feature maps and use two additional convolutional layers to produce dense ordinal map predictions. The first convolutional layer is designed to compress input feature maps to a lower dimension, while the second convolutional layer produces the desired $2K$-channel dense ordinal outputs.

D. Training and Test Process

During training, we divide each aerial image and corresponding digital surface model (DSM) image into small image patches with a size of 256×256 pixels. Previous methods mostly divide the whole image into gridded patches with overlaps. In this letter, to enable more training patches, we random crop image patches and height maps at the same location from original images. This can be regarded as one of the data augmentation strategies to make our model more robust to input variations.

In the test phase, we also divide each aerial image into small patches of size 256×256 and obtain the predicted depth map for each patch through one network forward pass. As the height maps used for model training are localized, the predicted height map for each patch also contains localized values, i.e., the relative height with respect to their minimum value. To connect the patch-wise height maps into seamless predictions, we need to determine the height shift (minimum height) of each patch prediction. To achieve this goal, we add a small overlap of 2 pixels when dividing the test images. By doing so, we can decide the height shift of one image if we know the height shift of its adjacent patches. More specifically, we merge the height predictions of two adjacent patches: we compute the height shift of adjacent patches using the difference of their mean values in the overlapping 2×256 region. In this letter, the top left patch is chosen as the base height image, the height values in all other image patches are shifted to the absolute ones based on it.

III. Experiments and Results

A. Experimental Data Sets

To show the effectiveness of our model for height estimation from a single aerial image, we conduct experiments on the public International Society for Photogrammetry and Remote Sensing (ISPRS) 2-D Semantic Labeling Challenge data set. This data set contains high-resolution aerial images in two Germany cities: Vaihingen and Potsdam. The Vaihingen data set contains 33 image tiles with a spatial resolution of 9 cm/pixel, and each image has a size of around 2500×2500 pixels. To enable a fair comparison with existing methods, we randomly choose 22 images for model training, and the remaining 11 images are used for model evaluation. The Potsdam data set contains 38 image tiles with a spatial resolution of 5 cm/pixel, and each image has a size of 6000×6000 pixels. To enable a fair comparison with existing methods, we randomly choose 25 tiles for model training and the remaining 13 tiles are used for model evaluation. All results are reported on the test set.

B. Implementation Details

Our model involves two parts: a ResNet-101 trained on ImageNet data set [20] as a feature extractor (backbone network) and the ordinal regression network with ASPP module. We use Adam as the optimizer for both parts: the initial learning rate for the ordinal regression network is set to 1e-3,

TABLE I

TOP: RESULTS ON THE VAIHINGEN DATA SET, BOTTOM: RESULTS ON THE POTSDAM DATA SET. NOTE THAT FOR [15], APPROXIMATELY HALF OF THE STUDIED AREA IS USED FOR TRAINING AND THE REMAINING FOR EVALUATION

Method	Rel↓	Rel (log10)↓	RMSE (m)↓	RMSE (log10)(m)↓	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
[15]	-	-	2.58 ± 0.09	-	-		-
[16]	1.163	0.234	2.871	0.334	0.330	0.572	0.741
Ours	0.314	0.126	1.698	0.155	0.451	0.817	0.939
[15]	-	-	3.89 ± 0.11	-	-	-	-
[16]	0.571	0.195	3.468	0.259	0.342	0.601	0.782
Ours	0.253	0.114	1.404	0.132	0.557	0.782	0.974

which is $10\times$ that of the feature extractor network, and the weight decay for both parts is set to 5e-4. We introduce this difference in learning rates to accelerate the training process of the ASPP module and, to a certain degree, prevent the feature extractor network from overfitting the training set. We divide the learning rate by 10 for both parts when the average loss stops decreasing for two epochs consecutively. We use a batch size of 15 and parallelize training on three GPUs. We train our model for 20 epochs for both ISPRS Vaihingen and Potsdam data set with each epoch that includes 10 000 randomly sampled patches from the predetermined training set. We implement our method with the public deep learning platform PyTorch [21] on TESLA K80 GPUs.

C. Evaluation Metrics

Following [22], we use four metrics to evaluate the height estimation performance, including the average relative error (Rel), average log10 error (Rel(log10)), root mean squre error (RMSE), and log10 RMSE. We also evaluate the performance by the ratio of pixels that have a predicted height value close to the ground truth. Following [22], we define the following evaluation metrics:

$$\delta^i = \max\left(\frac{\hat{h}}{h}, \frac{h}{\hat{h}}\right) < 1.25^i, \quad i \in \{1, 2, 3\} \tag{5}$$

where $\hat{h}$ and h denote the predicted and ground truth height value.

D. Experimental Results

1) Results on Vaihingen: For the Vaihingen data set, we randomly sample 10k patches with a size of 256×256 pixels for each training epoch. We train our model for 20 epochs, where each epoch takes approximately 48 min. The entire training process finishes in 20 h for the ISPRS validation data set. As shown in Table I, our proposed method obtains a significantly better performance on all evaluation metrics. More specifically, state-of-the-art method [16] achieves an RMSE of 2.871 m while our proposed method gets an RMSE of 1.698 m. Moreover, our method achieves significant better height estimation accuracy, indicated by δ_1, δ_3, and δ_3. Fig. 2 shows some selected patches of height estimation on the Vaihingen data set.

2) Results on Potsdam: For the Potsdam data set, we use the same experimental setting as we use for the Vaihingen data set. We note that the spatial resolution of this data set is higher than that of the Vaihingen data set. This makes our height estimation problem more challenging where each patch covers a smaller area of the region and thus contains less salient height-related structure. Our proposed method overcomes this challenge by applying an ASPP module where it combines features learned from different scales, alleviating the problem of having a relatively small receptive field. We list the height estimation results in Table I. From this table, we can see that our proposed method achieves significantly better performance than existing methods on all evaluation methods.

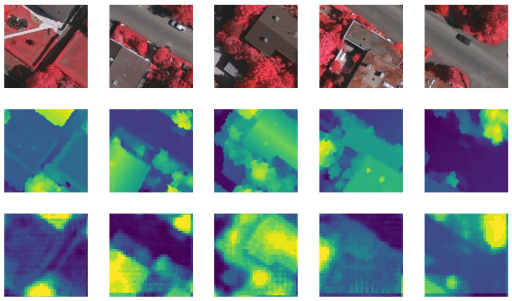

Fig. 2. Selected results from ISPRS Vaihingen data set. (Top row) infrered-red-green (IRRG) images. (Middle row) DSMs. (Bottom row) our predictions.

E. Ablation Analysis

We conduct ablation analysis to demonstrate the effectiveness of different modules in our proposed method, including height discretization and the ASPP module.

1) Height Discretization: We apply three variants of our model to achieve height estimation: 1) our model with MSE loss; 2) our model with UD strategy; and 3) our model with the proposed SID discretization strategy. The height estimation performances are listed in the top part of Table II. As shown in this table, training our model with vanilla regression loss (e.g., MSE) on continuous height values leads to significantly worse performance than our method with discretization, and our model with SID obtains the best performance. Specifically, our model with MSE loss gets a relative loss of 0.814, while our model with SID discretization achieves a relative loss of 0.314. Moreover, by using SID discretization rather than UD, the relative loss and RMSE decrease from 0.323 to 1.756 m and 0.314 to 1.698 m, respectively. This demonstrates the superiority of the SID discretization strategy. We also report the performance of our model with multiclass classification (MCC) loss instead of using ordinal regression loss, marked as "Ours MCC w SID." By comparing the last two rows of the top part of Table II, it can be seen that our model with the proposed ordinal regression loss achieves better performance than its counterpart with a MCC loss.

TABLE II

TOP: HEIGHT ESTIMATION PERFORMANCE OF OUR PROPOSED DISCRETIZATION STRATEGY AND UD ON VAIHINGEN DATA SET. BOTTOM: EFFECT OF ASPP MODULE ON THE VAIHINGEN DATA SET

Method	Rel↓	Rel (log10)↓	RMSE (m)↓	RMSE (log10)(m)↓	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
Ours w MSE	0.814	0.298	3.217	0.371	0.375	0.582	0.707
Ours w UD	0.323	0.127	1.756	0.163	0.473	0.814	0.915
Ours w SID	0.314	0.126	1.698	0.155	0.451	0.817	0.939
Ours MCC w SID	0.549	0.173	2.192	0.475	0.414	0.603	0.784
Ours w/o ASPP	0.588	0.256	3.522	0.296	0.204	0.373	0.583
Ours w ASPP	0.314	0.126	1.698	0.155	0.451	0.817	0.939

2) Effect of ASPP: Finally, we investigate the effect of the ASPP module for multiscale feature learning and aggregation. For comparison, we conduct further experiments using our model without the ASPP module, i.e., we directly feed the feature representations generated by the backbone network to two convolutional layers to produce the dense ordinal maps for height prediction. We list the performance of our models with and without ASPP modules in the bottom part of Table II. As can be seen, our model achieves a significantly better performance with the ASPP module. Specifically, our model with and without the ASPP module gets an RMSE of 3.522 and 1.698 m, respectively. This finding demonstrates multiscale feature learning is crucial for height estimation from aerial images.

IV. CONCLUSION

In this letter, we proposed a deep CNN-based method for height estimation from a single aerial image. Unlike previous methods that treat height estimation as a regression problem and directly use an encoder–decoder network to regress the height values, we proposed to divide the height values into spacing-increasing intervals and transform the regression problem into an ordinal regression problem and design an ordinal loss for network training. To enable multiscale feature learning, we further incorporate an ASPP module to aggregate features from multiple dilated convolution layers. Moreover, a postprocessing technique was used to transform the patched height predictions into a seamless depth map. Finally, we demonstrate the effectiveness of our proposed method on ISPRS Vaihingen and Potsdam data sets with various experimental settings. We also demonstrate the semantic labeling abilities of the generated depth maps. Our proposed method achieves an RMSE of 1.698 and 1.404 m on Vaihingen and Potsdam data sets, and a relative error of 0.314 and 0.253, respectively. Extensive experiments are conducted to show the effectiveness of each core module of our method.

REFERENCES

[1] X.-Z. Pan, Q.-G. Zhao, J. Chen, Y. Liang, and B. Sun, "Analyzing the variation of building density using high spatial resolution satellite images: The example of Shanghai city," *Sensors*, vol. 8, no. 4, pp. 2541–2550, Apr. 2008.

[2] K. Lwin and Y. Murayama, "A GIS approach to estimation of building population for micro-spatial analysis," *Trans. GIS*, vol. 13, no. 4, pp. 401–414, Aug. 2009.

[3] J. Tu, H. Sui, W. Feng, and Z. Song, "Automatic building damage detection method using high-resolution remote sensing images and 3D GIS model," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vols. 3–8, pp. 43–50, Jun. 2016.

[4] C. N. Koyama, H. Gokon, M. Jimbo, S. Koshimura, and M. Sato, "Disaster debris estimation using high-resolution polarimetric stereo-SAR," *ISPRS J. Photogramm. Remote Sens.*, vol. 120, pp. 84–98, Oct. 2016.

[5] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[6] N. Audebert, B. Le Saux, and S. Lefevrey, "Fusion of heterogeneous data in convolutional networks for urban semantic labeling," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, Mar. 2017, pp. 1–4.

[7] R. Qin, X. Huang, A. Gruen, and G. Schmitt, "Object-based 3-D building change detection on multitemporal stereo images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2125–2137, May 2015.

[8] R. Qin, J. Tian, and P. Reinartz, "3D change detection-approaches and applications," *ISPRS J. Photogramm. Remote Sens.*, vol. 122, pp. 41–56, Dec. 2016.

[9] M. J. Westoby, J. Brasington, N. F. Glasser, M. J. Hambrey, and J. M. Reynolds, "'Structure-from-motion' photogrammetry: A low-cost, effective tool for geoscience applications," *Geomorphology*, vol. 179, pp. 300–314, Dec. 2012.

[10] S. C. de Vries, A. M. L. Kappers, and J. J. Koenderink, "Shape from stereo: A systematic approach using quadratic surfaces," *Perception Psychophysics*, vol. 53, no. 1, pp. 71–80, Jan. 1993.

[11] S. K. Nayar and Y. Nakagawa, "Shape from focus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 8, pp. 824–831, Aug. 1994.

[12] 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.

[13] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang, and X. Fan, "Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6851–6860.

[14] 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

[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] 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.

[17] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2002–2011.

[18] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[19] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[21] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[22] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2366–2374.