Geometry-Aware Segmentation of Remote Sensing Images via Joint Height Estimation

Xiang Li, Congcong Wen, Lingjing Wang, and Yi Fang, *Member, IEEE*

***Abstract*— Recent studies have shown the benefits of using additional elevation data [e.g., digital surface model (DSM) or normalized DSM (nDSM)] for enhancing the performance of the semantic labeling of aerial images. However, previous methods mostly adopt 3-D elevation information as additional inputs, while, in many real-world applications, one does not have the corresponding DSM images at hand, and the spatial resolution of acquired DSM images usually does not match the aerial images. To alleviate this data constraint and also take advantage of 3-D elevation information, in this letter, a geometry-aware segmentation model is introduced to achieve accurate semantic labeling of aerial images via joint height estimation. Instead of using a single-stream encoder–decoder network for semantic labeling, we design a separate decoder branch to predict the height map and use the DSM images as side supervision to train this newly designed decoder branch. With the newly designed decoder branch, our model can distill the 3-D geometric features from 2-D appearance features under the supervision of ground-truth DSM images. Moreover, we develop a new geometry-aware convolution module that fuses the 3-D geometric features from the height decoder branch and the 2-D contextual features from the semantic segmentation branch. The fused feature embeddings can produce geometry-aware segmentation maps with enhanced performance. Our model is trained with DSM images as side supervision, while, in the inference stage, it does not require DSM data and directly predicts the semantic labels. Experiments on International Society for Photogrammetry and Remote Sensing (ISPRS) Vaihingen and Potsdam data sets demonstrate the effectiveness of the proposed method for the semantic segmentation of aerial images.**

***Index Terms*— Feature fusion, geometry-aware convolution (GAC), height estimation, semantic segmentation.**

I. Introduction

In RECENT years, convolutional neural networks (CNNs) have drawn huge attention in remote sensing and photogrammetry due to the remarkable performance in many applications. The encouraging performance drives researchers to develop CNN-based methods for the semantic labeling of remote sensing images (RSIs). In this direction, early efforts adopt patch-based CNNs to perform classification for the center pixel of each input patch; recent methods mostly perform pixelwise segmentation using fully convolutional networks. For example, Maggiori *et al.* [1] developed a fully convolutional model for the classification of RSIs. Substantial researches have tried to enhance the performance by using more powerful encoder networks [2], [3], incorporating dilated convolution modules [4], [5], or using more powerful output representations [6], [7].

It is commonly known that objects in RSIs are characterized by complex spectral–spatial properties and need a comprehensive feature extraction process to ensure the classification performance. Nevertheless, existing CNN-based methods mostly focus on spectral and contextual feature extraction using a single encoder–decoder network, while geometric features (such as height above ground and implicit 3-D structure) are often not fully explored. A direct remedy to this issue is to explicitly incorporate geometric-related data [such as digital surface model (DSM)] as additional inputs. Audebert *et al.* [8] proposed to enhance the segmentation performance of RSIs by fusing the feature representations from both RGB images and elevation composite images [normalized difference vegetation index (NVDI), DSM, and normalized DSM (nDSM)]. Concretely, they propose a two-stream network that simultaneously learns spectral and auxiliary geometric features, and a residual correction module is leveraged to fuse the features from two encoder networks.

In this letter, instead of directly taking elevation data (e.g., DSM or nDSM) as additional inputs, we propose to jointly learn geometric features using a height estimation network. Our key insight is that geometric information (height above ground) is naturally preserved by the aerial images and can be estimated from monocular inputs [9], [10]. The learned 3-D geometric features are further fused with the 2-D contextual features using the newly designed geometry-aware convolution (GAC) module. Our model is, thus, able to distinguish those objects that have similar 2-D appearances but with distinct geometric characteristics, e.g., rooftop and impervious surface. Through joint training of these two tasks, the implicit 3-D geometric information can be well extracted and fused with contextual features, which further contributes to better semantic labeling performance. More importantly, after training, our model does not need elevation data and can directly produce the segmentation labels for the test images.

We also note that some recent works [11], [12] explore a multitask learning strategy for simultaneously height estimation and semantic labeling, which are quite similar to the proposed method. Unlike these methods that decouple two tasks in the middle or top layers of decoder networks, our method uses two task-specific decoder branches: one for semantic labeling and the other for height estimation. More importantly, a GAC module is proposed to effectively fuse semantic feature embeddings and geometric feature embeddings to enhance performance.

The main contributions of this letter are summarized as follows.

1) This letter introduces a geometry-aware neural network model for the semantic labeling of aerial images. Instead

Manuscript received October 23, 2020; revised December 7, 2020 and January 19, 2021; accepted February 3, 2021. Date of publication February 24, 2021; date of current version January 5, 2022. This work was supported in part by the ADEK under Grant AARE-18150 and in part by the NYU Abu Dhabi Institute under Grant AD131. *(Corresponding author: Yi Fang.)*

Xiang Li, Lingjing Wang, and Yi Fang are with the Multimedia and Visual Computing Laboratory, NYU Tandon, Abu Dhabi, United Arab Emirates, and also with the Multimedia and Visual Computing Laboratory, NYU Tandon, New York City, NY 11201 USA (e-mail: yfang@nyu.edu).

Congcong Wen is with the Department of Electrical and Computer Engineering, New York University, New York City, NY 11201 USA.

Digital Object Identifier 10.1109/LGRS.2021.3058168

of taking the DSM images as additional inputs, our model simultaneously predicts the segmentation maps and the height maps from input aerial images. After training, it does not need DSM data and can directly produce the segmentation labels for the test images.

2) A GAC module is proposed to effectively fuse semantic feature embeddings and geometric feature embeddings to enhance the performance of semantic labeling.

II. Methods

A. Method Overview

Traditional single-stream encoder–decoder-based networks use a successive of convolutional and pooling layers to obtain high-level contextual features from input images, and then, a successive of convolutional and unpooling layers is adapted to decode the learned features into classification score maps. In this way, the network can learn only 2-D contextual/appearance features while neglecting the 3-D geometric information, which is also important for distinguishing those objects that have similar 2-D appearances but with different geometric characteristics, e.g., rooftop and impervious surface.

In this letter, our proposed method explicitly enables geometric feature learning by incorporating a new decoder branch. During training, the 3-D information from ground-truth height maps is used to guide the training procedure of the newly designed decoder branch. Fig. 1 illustrates the proposed Geometry-Aware segmentation network (GANet) for aerial image classification. Our GANet model contains three main components: the encoder network, the segmentation decoder, and the height decoder. The encoder network aims to learn both contextual and geometric features from input images, which will be introduced in Section II-B. The segmentation decoder predicts classification maps, while the height decoder learns the geometric embeddings by predicting height maps. After getting the contextual and geometric feature embeddings, a GAC module is used to fuse these two forms of features to enables geometry-aware semantic labeling. The GAC module is illustrated in Section II-C.

B. Encoder–Decoder Network

Our GANet follows the prevalent Deeplab V3+ [13] architecture to design its encoder and decoder parts. In the encoder part, a backbone network is used to extract multiscale feature representations. An atrous spatial pyramid pooling (ASPP) module is applied after the backbone network to learn multiscale features.

In the decoder part, the combined features are fed into two separate 3×3 convolution layers to learn independent feature representations for the task of semantic segmentation and height estimation, respectively. The learned height-related geometric feature embeddings are directly upscaled by $4\times$ and fed into another convolutional layer to predict the height maps. The learned semantic-related contextual features are fused with the geometric features by leveraging the newly proposed GAC module for enhancement. The GAC module is introduced in Section II-C. The fused feature maps are then passed to a convolutional layer to predict the semantic labels.

C. GAC Module

We introduce a GAC module that leverages the learned geometric embeddings as guidance for the convolution operation. Instead of using original height values as convolution inputs, the proposed convolution operation takes as input both contextual and geometric features in the embedding space. Given an input contextual feature map x and the learned geometric embeddings $G \in \mathcal{R}^{H\times W\times E}$, the convolution output y_i at location i can be formulated as

$$y_i = \sigma\left(\sum_{j\in\mathcal{N}_i} W_{ij}(G)x_j + b\right) \tag{1}$$

where W_{ij} is the kernel weights derived from the geometry guidance G and b is the bias term. Here, W_{ij} can be regarded as a geometric similarity between pixel i and j defined in the embedding space. To better calculate W_{ij}, we follow [14] to decouple it as a dot-product of two subspace embeddings:

$$W_{ij}(G) = \phi(G_i)\cdot\psi(G_j) \tag{2}$$

where $\phi(\cdot)$ and $\psi(\cdot)$ denote the features in subembedding space. In this letter, the subembedding functions $\phi(\cdot)$ and $\psi(\cdot)$ are implemented by two independent convolutional layers. Then, the proposed GAC operation is defined as

$$y_i = \sigma\left(\sum_{j\in\mathcal{N}_i} \phi(G_i)\cdot\psi(G_j)\cdot x_j + b\right). \tag{3}$$

D. Multitask Objective Function

Our GANet model gets supervision from both semantic segmentation branch and height estimation branch. The overall loss function is formulated as

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda\mathcal{L}_h \tag{4}$$

where $\mathcal{L}_{\text{seg}}$ denotes the segmentation loss, $\mathcal{L}_h$ denotes height estimation loss, and λ is a hyperparameter to balance these two loss terms. By default, λ is set to 1 in our experiments.

For the semantic segmentation task, we leverage the weighted cross-entropy loss for model training, where the inverse class frequencies are used as the balance weights for all pixels of that class. In this letter, we adopt L1 loss to train the height estimation network.

E. Results on Vaihingen

III. Experiments and Results

A. Data Sets

To verify the effectiveness of our proposed model for the semantic labeling of RSIs, we conduct experiments on the International Society for Photogrammetry and Remote Sensing (ISPRS) 2-D Semantic Labeling Challenge data set. This data set contains very high-resolution aerial images from two cities of Germany: Vaihingen and Potsdam. We follow the same settings as [8] to prepare the data set for model training and evaluation.

B. Implementation Details

Our GANet model is developed based on PyTorch Library. The network is trained using a momentum stochastic gradient descent (SGD) algorithm with a weight decay of 0.0005. The initial learning rate is set to 0.01, respectively. We train our model for 100 epochs until convergence on 4 T P100 GPUs with a batch size of 4. We use synchronized batch normalization [15] after each convolutional layer. We use overall accuracy (OA) and per-class F1 score to evaluate the classification performance of our GANet model.

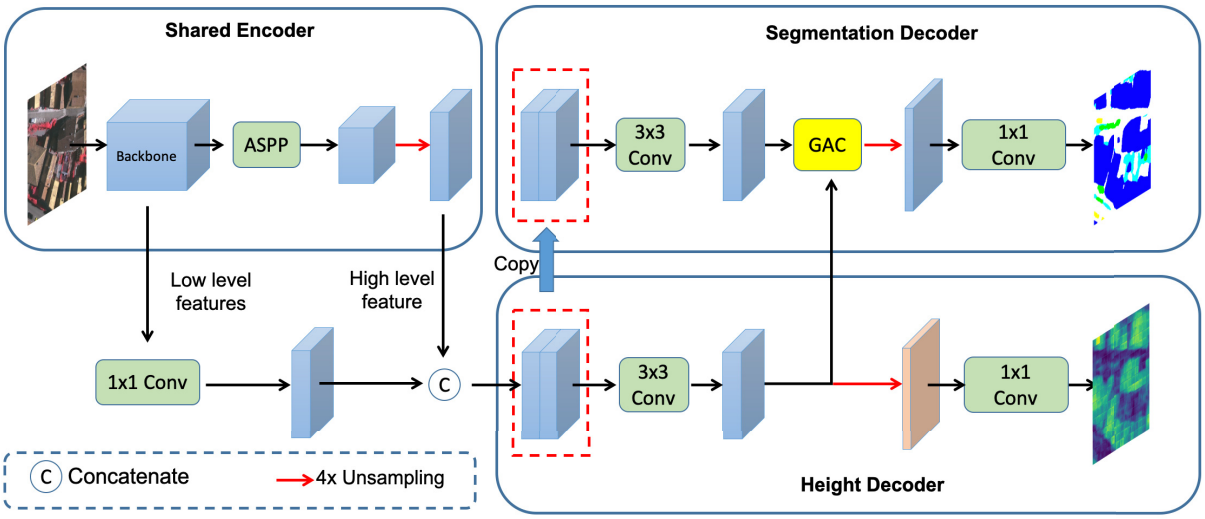

Fig. 1. Overview of our GANet model for RSI semantic labeling. Our model receives a single aerial image as input and predicts the classification map and height map simultaneously. "ASPP" denotes the atrous spatial pyramid pooling module, and "GAC" denotes our newly proposed geometry-aware convolution module.

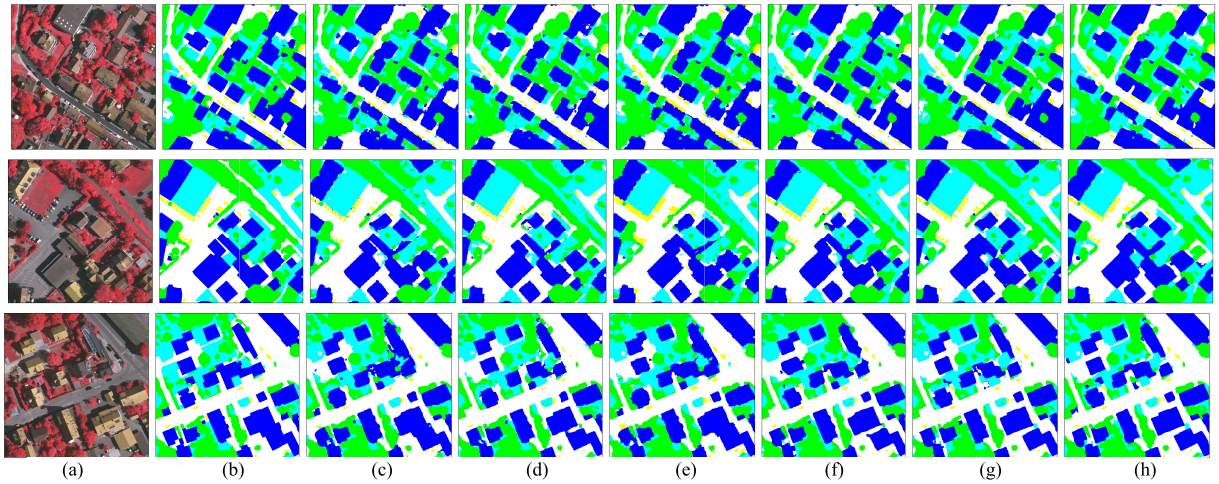

Fig. 2. Selected examples of the classification results on the ISPRS Vaihingen data set. (a) Image. (b) Ground truth. (c) DST_2. (d) ONE_7. (e) ADL_3. (f) DLR1_0. (g) CASIA_2. (h) GANet.

In our experiments, we compare our GANet model with state-of-the-art methods submitted to the ISPRS 2-D Semantic Labeling Contest. Comparing methods include ADL [16], DST [2], DLR [17], ONE [18], RIT [19], SWJ [20], CASIA [21], and TreeUNet [22].

The quantitative performance of our GANet model on the test split of the ISPRS Vaihingen data set is shown in Table I. As indicated in Table I, our GANet model gets better performance than all comparing methods with an OA of 91.3% and an average F1 score of 90.4%. The best-published CASIA2 model achieves quite close performance to our model. Note that CASIA2 pretrains their model on the PASCAL VOC 2012 data set and then fine-tunes their model on the ISPRS Vaihingen data set. Ablation analysis shows that the CASIA2 model improves the performance a lot by using the fine-tune technique. Our GANet model is trained from scratch and does not need pretraining from other data sets.

Fig. 2 shows the classification results of our GANet model and the compared methods on several sampled patches. As can be seen in this figure, our GANet model produces satisfying classification results on all test samples. Moreover, our model can better distinguish between the building and impervious surface categories, as well as vegetation-tree categories. We owe this to the 3-D geometric difference between these objects.

C. Results on Potsdam

We report the performance of our GANet model and the comparing models on the test split of ISPRS Potsdam data set in Table II. As shown in Table II, our GANet model obtains the best performance on average F1 score and the second-best performance on OA. We note that the SWJ_2 model obtains better performance than our GANet model on OA, and CASIA_2 obtains quite close performance compared to our

TABLE I

CLASSIFICATION PERFORMANCE ON THE VAIHINGEN DATA SET. "DSM(S)" INDICATES THE MODEL USING DSM AS ADDITIONAL SUPERVISION

Method	Input	Imp. surf.	Buildings	Low veg.	Trees	Cars	OA	Average F1
DST_2	IRRG+DSM	90.5	93.7	83.4	89.2	72.6	89.1	85.9
ONE_7	IRRG+DSM+NDSM	91.0	94.5	84.4	89.9	77.8	89.8	87.5
ADL_3	DSM+nDSM	89.5	93.2	82.3	88.2	63.3	88.0	83.3
DLR_10	IRRG+DSM+Edge	92.3	95.2	84.1	90.0	79.3	90.3	88.2
CASIA2	IRRG	**93.2**	**96.0**	**84.7**	89.9	86.7	91.1	90.1
TreeUNet	IRRG+DSM	92.5	94.9	83.6	89.6	85.9	90.4	89.3
Ours	IRRG+DSM(s)	93.1	95.9	84.6	**90.1**	**88.4**	**91.3**	**90.4**

TABLE II

CLASSIFICATION PERFORMANCE ON THE POTSDAM DATA SET. "DSM(S)" INDICATES THE MODEL USING DSM AS ADDITIONAL SUPERVISION

Method	Input	Imp. surf.	Buildings	Low veg.	Trees	Cars	OA	Average F1
DST_5	IRRGB+DSM	92.5	96.4	86.7	88.0	94.7	90.3	91.7
RIT_L7	IRRGB+nDSM+NDVI	91.2	94.6	85.1	85.1	92.8	88.4	89.8
SWJ_2	IRRG	**94.4**	97.4	87.8	87.6	94.7	**91.7**	92.4
CASIA_2	IRRGB	93.3	97.0	87.7	88.4	96.2	91.1	92.5
TreeUNet	IRRGB+DSM+nDSM	93.1	**97.3**	86.8	87.1	95.8	90.7	92.0
Ours	IRRG+DSM(s)	93.0	**97.3**	**88.2**	**89.5**	**96.8**	91.3	**93.0**

TABLE III

EFFECT OF HEIGHT SUPERVISION. GANET* DENOTES OUR BASELINE MODEL WITHOUT HEIGHT ESTIMATION

Method	OA	Average F1
GANet* w/o GAC	90.7	88.2
GANet ($\lambda = 0.5$) w/o GAC	91.2	89.3
GANet ($\lambda = 1$) w/o GAC	**91.3**	**89.6**
GANet ($\lambda = 2$) w/o GAC	91.2	89.1

TABLE IV

EFFECT OF GAC MODULE. "SUM FUSION" DENOTES OUR MODEL USING ELEMENTWISE SUMMATION FOR FEATURE FUSION. "MS TEST" DENOTES MULTISCALE TEST

Method	OA	Average F1
GANet w/o feature fusion	91.3	89.6
GANet w/ Sum Fusion	91.6	90.1
GANet w/ GAC	**92.0**	**90.7**
GANet w/ GAC + ms test	**92.3**	**91.1**

model. However, these two comparing models were pretrained on the PASCAL VOC 2012 data set and then fine-tuned on the ISPRS Potsdam data set. In contrast, our GANet model does not require additional data sets for pretraining. One should also note that OA is sensitive to the class distribution, while F1-score is a better metric when there are imbalanced classes as in the above case. Moreover, our model gets new state-of-the-art performance on four out of five categories, including building, low vegetation, tree, and car.

IV. DISCUSSION

A. Effect of Height Supervision

First, we investigate the effectiveness of height supervision in comparison to methods that directly using DSM images as inputs. To achieve this, we remove the height decoder branch and the GAC module from our GANet model and use it as a baseline model. The baseline model now becomes a traditional single-stream encoder–decoder network. We report the performance of the baseline and our GANet model with the height decoder branch in Table III. Moreover, we also list the performance of our GANet model with different configurations of λ. Note that all models do not include the GAC module in this section.

From Table III, one can find out that, by using the proposed height decoder branch, our GANet model gets a significant performance improvement, which demonstrates the benefits of using height information as side supervision. Specifically, the baseline model obtains an OA of 90.7% and an average F1 score of 88.2% on the Vaihingen validation set, while our GANet model with the height decoder branch achieves an OA of 91.3% and an average F1 score of 89.6% when λ equals 1. Moreover, Table III also show that different values of λ give similar performance. Our GANet model gets the best performance when λ equals 1.

B. Effect of GAC Module

Then, we explore the benefits of using our newly proposed GAC module for contextual and geometrical feature fusion. We investigate the performance of our model with and without the GAC module with λ set to 1. We also explore another variant of our GANet model using elementwise summation for feature fusion instead of the proposed GAC module. Table IV lists the quantitative performance of our GANet model and the comparing methods. The results show that both the elementwise summation fusion strategy and GAC module can improve the segmentation performance. More importantly, the proposed GAC module performs better than the elementwise summation fusion strategy. This is because our GAC module can effectively learn geometric affinity from the geometrical embeddings and use it to reweight the convolutional kernels.

Fig. 3 shows an example of the semantic segmentation results with different fusion strategies. As can be seen in Fig. 3, the model without feature fusion misclassifies some building pixels as impervious surface and leads to incorrect classification between low vegetation and tree categories. The two models with feature fusion modules can successfully correct the errors between the building and impervious surface categories using the geometrical information from the height decoder

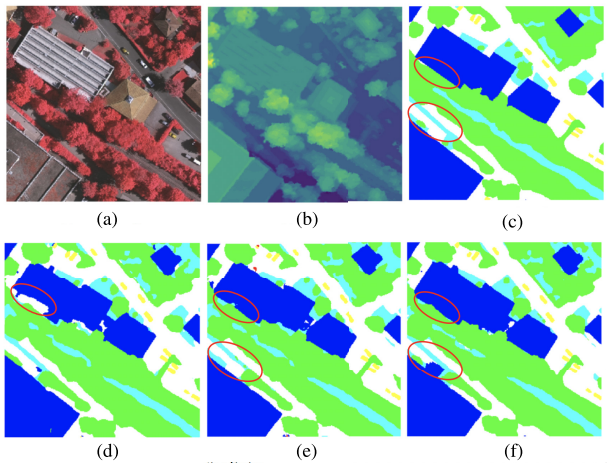

Fig. 3. Effect of feature fusion on an example of the ISPRS Vaihingen data set. (a) Aerial image. (b) DSM. (c) Ground truth. (d) GANet w/o feature fusion. (e) GANet w/Elementwise Summation fusion. (f) GANet w/GAC fusion.

branch. Moreover, our model using the GAC module obtains better performance than its counterpart using an elementwise summation fusion strategy. The improvement mainly comes from a better classification between the tree and low vegetation categories.

Moreover, we also investigate the effectiveness of a multi-scale test strategy. As can be seen in Table IV, our GANet model enjoys a further performance boost on both evaluation metrics by using a multiscale test strategy.

V. Conclusion

In this letter, we introduce a geometry-aware CNN to approach the problem of semantic segmentation of RSIs. Our model benefits from the 3-D geometric information via joint height estimation. Unlike previous methods that mostly use a single decoder network to predict pixelwise semantic labels, in our model, a newly designed height decoder branch is developed to predict the height map under the supervision of DSM images. Furthermore, we introduce a novel GAC module to combine the learned 3-D geometric features and 2-D contextual features from two decoder branches. Experiments on ISPRS Vaihingen and Potsdam data sets demonstrate the effectiveness of our proposed method for aerial image classification.

Acknowledgment

The authors would like to thank the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) (http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html) for providing the Vaihingen data set. They would also like to thank the ISPRS WG II/4 for releasing the Vaihingen and Potsdam data sets and organizing the 2-D semantic labeling contest.

References

[1] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.

[2] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," 2016, *arXiv:1606.02585*. [Online]. Available: http://arxiv.org/abs/1606.02585

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[4] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 182–186.

[5] X. Wei *et al.*, "Semantic pixel labelling in remote sensing images using a deep convolutional encoder-decoder model," *Remote Sens. Lett.*, vol. 9, no. 3, pp. 199–208, Mar. 2018.

[6] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.

[7] D. Chai, S. Newsam, and J. Huang, "Aerial image semantic segmentation using DCNN predicted distance maps," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 309–322, Mar. 2020.

[8] N. Audebert, B. L. 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.

[9] 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.

[10] L. Mou and X. X. 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

[11] 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.

[12] M. Carvalho, B. L. 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.

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[14] J. Jiao, Y. Wei, Z. Jie, H. Shi, R. Lau, and T. S. Huang, "Geometry-aware distillation for indoor semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2869–2878.

[15] H. Zhang *et al.*, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.

[16] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Van Den Hengel, "Semantic labeling of aerial and satellite imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 7, pp. 2868–2881, Jul. 2016.

[17] 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. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 2016, no. 3, pp. 473–480, 2016.

[18] N. Audebert, B. L. Saux, and S. Lefèvre, "Semantic segmentation of Earth observation data using multimodal and multi-scale deep networks," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 180–196.

[19] S. Piramanayagam, W. Schwartzkopf, F. Koehler, and E. Saber, "Classification of remote sensed images using random forests and deep learning framework," *Proc. SPIE*, vol. 10004, Oct. 2016, Art. no. 100040L.

[20] J. Wang, L. Shen, W. Qiao, Y. Dai, and Z. Li, "Deep feature fusion with integration of residual connection and attention model for classification of VHR remote sensing images," *Remote Sens.*, vol. 11, no. 13, p. 1617, Jul. 2019.

[21] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 78–95, Nov. 2018.

[22] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, "TreeUNet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation," *ISPRS J. Photogramm. Remote Sens.*, vol. 156, pp. 1–13, Oct. 2019.