Geometry-Aware Segmentation of Remote Sensing Images via implicit height estimation

Xiang Li and Yi Fang

Abstract—Deep learning-based methods, especially deep convolutional neural networks (CNNs), have made significant breakthroughs in the field of remote sensing and greatly advanced the performance of the semantic segmentation of remote sensing images. Recent studies have shown the benefits of using additional elevation data (e.g., DSM or nDSM) for enhancing the performance of the semantic segmentation of aerial images. However, previous methods mostly adopt 3D elevation information as additional inputs. While in many real-world applications, one does not have the corresponding DSM information at hand and the spatial resolution of acquired DSM images usually do not match the aerial images. To alleviate this data constraint and also take the advantage of 3D elevation information in this paper, we propose a geometry-aware segmentation model that achieves accurate semantic segmentation of aerial images via implicit 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. In this way, our model does not require DSM as model input and still benefits from the helpful 3D geometric information during training. With the newly designed decoder branch, our model can distill the 3D geometric features from 2D appearance features under the supervision of ground truth DSM images. Moreover, we develop a new geometry-aware convolution module that fuses the 3D geometric features from the height decoder branch and the 2D 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 can directly predict the semantic labels in an end-to-end fashion. Experiments on ISPRS Vaihingen and Potsdam datasets demonstrate the effectiveness of our proposed method for the semantic segmentation of aerial image. Our proposed model achieves remarkable performance on both datasets without using any hand-crafted features or post-processing.

Index Terms—Geometry-Aware Convolution, Feature Fusion, Semantic Segmentation, Height Estimation

I. INTRODUCTION

The semantic segmentation problem, which is often called image classification in the field of remote sensing, is generally defined as determining the semantic classes of all pixels in the input images. Automatic semantic segmentation has been a long-standing problem in the field of remote sensing and plays a crucial role in many applications, such as land use/land cover mapping, agricultural production estimate, building extraction, city planning, and etc.

Recent years, convolutional neural networks (CNNs) have drawn huge attentions in remote sensing and photogrammetry due to the remarkable performance in many applications, such as scene classification [1], [2], image classification [3], [4], [5], object detection [6], [7], building extraction [8], [9], [10], [11], land use classification [12], [13], point cloud classification [14], [15], [16]. The encouraging performance drives researchers to develop CNN-based methods for the semantic segmentation of remote sensing images (RSIs). In this direction, early efforts adopt patch-based CNNs to predict the class label for the center pixel of each input patch, recent methods mostly perform pixel-wise segmentation using fully convolutional networks. For example, [3] develops a fully convolutional model for pixel-wise classification of remote sensing images in an end-to-end fashion. Substantial researches have tried to enhance the performance by using more powerfully encoder network [17], [18], incorporating dilated convolution module [19], [20] or using more powerful output representations [21], [22].

It is commonly known that objects in remote sensing images 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, implicit 3D structure) are often not fully explored. A direct remedy to this issue is to explicitly incorporate geometric-related data (such as DSM) as additional inputs. [5] proposes to enhance the segmentation performance of remote sensing images by fusing the feature representations from both RGB images and elevation composite images (NVDI, DSM, nDSM). Concretely, they develop a two-stream network that jointly learns RGB and auxiliary geometric features, and a residual correction module is leveraged to fuse the features from two encoder networks.

In this paper, instead of directly taking elevation data (e.g., DSM or nDSM) as additional inputs, we propose to implicitly 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 [23], [24], [25], [26]. The learned 3D geometric features are further fused with the 2D contextual features using the newly designed geometry-aware convolution module. Our model is thus able to distinguish those objects that have similar 2D appearance but with different geometric characteristics, e.g., rooftop and impervious surface. Figure 1 gives an illustration of the proposed framework for simultaneous semantic segmentation and...
height estimation. Through joint training of these two tasks, the implicit 3D geometric information can be well extracted and fused with contextual features, which further contributes to better semantic labeling performance.

The main contributions of this paper are summarized as follows:

1) This paper introduces a geometry-aware neural network for the semantic labeling of aerial images. Instead of taking the DSM images as additional inputs, our model simultaneously predicts the segmentation maps as well as the height maps from input aerial images.

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

3) We validate the effectiveness of the proposed method on the ISPRS 2D Semantic Labeling datasets and report remarkable performance compared to the state-of-the-art methods.

The remainder of this paper is organized as follows. In Section II, we briefly review the deep learning-based methods for semantic labeling and height estimation on remote sensing images in section II. The proposed geometry-aware segmentation model is introduced in Section III. In Section IV, we conduct experiments to validate the effectiveness of the proposed method for semantic labeling of remote sensing images. We investigate the effect of our proposed height estimation module and geometry-aware convolution module in Section V. Finally, our paper is concluded in Section VI.

II. RELATED WORK

A. Semantic Segmentation of Remote Sensing Images

Benefiting from the powerful feature learning abilities of deep neural networks, semantic segmentation of remote sensing images have achieved significant improvements. In this direction, [27] is the first successful work that utilizes patch-based CNN for road and building extraction. [9] develop a method for simultaneous road and building extraction on an aerial imagery dataset using a single CNN network. A moving average technique is designed to ensemble model predictions with different spatial displacements and further enhance the performance. [28] further extend the approach to multispectral images and validate the performance of their model for building extraction. Nevertheless, these patch-based methods need to divide the original input images into small patches and can only produce one classification label for each patch. They need to slide over the whole image plane to get the final prediction, which makes these methods inefficient for large-scale datasets.

A breakthrough comes from Fully Convolutional Network (FCN) [29] which converts fully connected layers of CNN into convolutional layers, and therefore enables dense pixel-wise segmentation of input images. After FCN, numerous variants have been proposed to enhance the performance by using more powerful encoder networks, skip-connections, or dilated convolution modules. For the specific task of remote sensing image classification, recently proposed methods mostly base their method on encoder-decoder architecture. For example, [3] develop an end-to-end fully convolutional architecture for fine-grained pixel-wise classification on large-scale remote sensing images. [30] design an FCN to perform pixel-wise classification on the ISPRS semantic labeling benchmark. They also discuss different design choices of the proposed method and demonstrate an ensemble of CNNs can achieve better results. [4] explicitly adds a boundary detection branch to the SegNet [18] architecture to preserve high-frequency details of object boundaries. They show that adding a boundary detection network can significantly improve the semantic labeling performance of remote sensing images. Instead of focusing on the design of network architecture, other researchers try to boost performance by using more powerful output representations. [21] propose a novel method for building extraction which uses the signed distance from each pixel to building boundaries to represent output and report significant performance boost over the baseline model using traditional label maps. A more recent work [22] extend this method to multi-class segmentation of remote sensing images and train a CNN model to predict multiple signed distance maps, one for each class.

Moreover, some recent work also tries to improve performance by taking advantage of multi-modal fusion. [31] develop two-stream network for VHR image segmentation. An FCN branch is used to generate classification probabilities from optical images, meanwhile, a multinomial logistic regression branch is used to generate probability maps from LiDAR inputs. A high-order CRF model is leveraged to fuse these two probabilistic results. [5] propose a two-stream network that jointly learns RGB and depth features in two branches and they also investigate the early and late fusion of multi-modal remote sensing data. Instead of using DSM as additional inputs, [32] proposes to treat DSM as a mid-level feature and feed it to the hypercolumn layer of a VGG network.

B. Height Estimation from Single Aerial Images

In the field of remote sensing, existing methods for height estimation mostly focus on 3D reconstruction based on stereo or multi-view image matching. There are only a few numbers of researches focus on estimating height from single aerial images. Early efforts mostly start by identifying object shades based on pixel-wise or object-based features and then estimate height values using camera information [33], [34], [35]. For example, [35] propose a method for building height estimation which firstly projects the building shadow onto the ground plane and then adjusts the building height until the projected shadow can be well-aligned with the real one. Much attention has been paid to enhance the shadow detection performance by using more powerful contextual and geometric features [33], [34]. Other researches also try to estimate height from single remote sensing images using a small number of control points [36] or sparse DTM [37].

Height estimation from aerial images is highly related to the task of depth estimation in the field of computer vision. Recent progress in the computer vision field has frequently shown the capabilities of deep neural networks for learning representative
In this section, we introduce the proposed geometry-aware semantic segmentation model. First, we give an overview of the proposed method in Section III-A. The encoder-decoder network is introduced in Section III-B. The proposed geometry-aware convolution module and objective loss function are illustrated in Section III-C and Section III-D respectively.

A. Method Overview

Given a group of aerial images \( \mathcal{I} = \{I_1, I_2, ..., I_N\} \) and the corresponding label maps \( \mathcal{Y} = \{Y_1, Y_2, ..., Y_N\} \), where \( I_i \in \mathcal{R}^{H \times W \times 3} \) and \( Y_i \in \mathcal{R}^{H \times W} \), \( H \) and \( W \) represent the image height and width respectively. Our semantic segmentation model aims to predict a classification map \( \hat{Y}_i \) for each input image \( I_i \). 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 group of convolutional and unpooling layers are adopted to decode the learned features into classification score maps. A pre-defined classification loss, e.g., cross-entropy loss, formulated on the predicted classification maps and the ground truth ones are used to optimize the network parameters. In this way, the network is forced to learn only 2D contextual/appearance features, while neglecting the 3D geometric information which is also important for distinguishing those objects that have similar 2D appearance but with different geometric characteristics, e.g., rooftop and impervious surface.

In this paper, our proposed method explicitly enables geometric feature learning by incorporating a new decoder branch. During training, the 3D information from ground truth height maps is used to guide the training procedure of the newly designed decoder branch. Figure 2 gives an overview of the proposed Geometry-Aware segmentation network (GANet) for aerial image classification. Our GANet model contains three main components: the encoder network, segmentation decoder, and height decoder. The encoder network aims to learn both
Fig. 2. Overview of our GANet model for remote sensing image semantic labeling. Our model takes a single aerial image as input and outputs the semantic segmentation map and height map simultaneously. The shared encoder adopts ResNet-101 as a backbone network to extract multi-scale high-level features from input images. An atrous spatial pyramid pooling (ASPP) module is appended at the end of the backbone network to fuse features at different scales. The combined feature maps are upsampled by 4x and further concatenated with the low-level features of the same resolution from the backbone network. In the decoder part, the combined feature maps are fed into two separate $3 \times 3$ convolution layers to learn independent feature representations for semantic segmentation and height estimation respectively. A geometry-aware convolution module is then used to further fuse these two forms of features thus enables geometry-aware semantic labeling.

contextual and geometric features from input images, which will be introduced in Section III-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 geometry-aware convolution module (GAC) is used to fuse these two forms of features thus enables geometry-aware semantic labeling. The GAC module is illustrated in Section III-C.

B. Encoder-Decoder Network

Our GANet follows the prevalent Deeplab V3+ [47] architecture to design its encoder and decoder parts. In the encoder part, a backbone network (e.g., VGG-16, ResNet-101) is used to extract multi-scale feature representations. An atrous spatial pyramid pooling (ASPP) module is applied after the backbone network to combine features at different scales. In our method, the ASPP module consists of one regular convolutional layer with a kernel size of $1 \times 1$ and 3 dilated convolutional layers with a kernel size of $3 \times 3$ and a dilated rate of 6, 12 and 18 respectively. A global average pooling layer along is also leveraged in the ASPP module to encode full-image information and is further up-sampled to the original resolution. Figure 3 gives an illustration of the ASPP module used in this paper. The multi-scale feature representations after the ASPP module are then upsampled by 4x and further concatenated with low-level features of the same resolution from the backbone network. Note that the low-level features are fed into another convolution layer before concatenation.

In the decoder part, the combined features are fed into two separate $3 \times 3$ 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 upsampled by 4x and fed into another convolutional layer to predict the height maps. While the learned semantic-related contextual features are fused with the geometric features by leveraging the newly proposed geometry-aware convolution (GAC) module for enhancement. The GAC module is introduced in the next section. The fused features are then fed into a convolutional layer to predict the semantic labels. In this paper, our feature extraction network is build upon ResNet-101 [48] architecture and is pre-trained on Pascal VOC 2012 dataset [49].

C. Geometry-Aware Convolution Module

Before going to our newly proposed convolution module, let's revisit conventional convolutional layer first. A conventional 2D convolution operation aggregate all activations
within a local neighborhood. Given an input feature map $X \in \mathbb{R}^{H \times W \times C}$, where $H$ and $W$ denote the width and height of the feature map, $C$ denotes the number of input channel, the output feature vector at pixel $i$ can be calculated as:

$$y_i = \sigma(\sum_{j \in N_i} W_{ij} x_j + b)$$  \hspace{1cm} (1)$$

where $x_j \in \mathbb{R}^C$ denotes the input feature vector at neighbor pixel $j$, $N_i$ denotes the local neighborhood of pixel $i$, and $y_i \in \mathbb{R}^E$ ($E$ is the output dimension) denote the output feature vector at pixel $i$, $\sigma$ denotes the activation function (e.g., sigmoid), $W \in \mathbb{R}^{C \times E}$ is the convolution kernel which is shared across all pixel locations, $b \in \mathbb{R}^E$ denotes the bias.

Previous researches [50], [51] have explored the geometric correlations between pixels by adding a geometric similarity term to the convolution operation. In [50], the author propose a depth-aware convolution operation calculated as:

$$y_i = \sigma(\sum_{j \in N_i} S(d_i, d_j) W_{ij} x_j + b)$$  \hspace{1cm} (2)$$

where $d_i$ and $d_j$ denote the depth values at pixel location $i$ and $j$ respectively, $S(\cdot, \cdot)$ measures the depth similarity between two depth values. By this formulation, the neighbor pixel which has a similar depth with the center pixel $i$ will have a larger impact on the convolution output.

Motivated by this formulation, we introduce a geometry-aware convolution 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 $G \in \mathbb{R}^{H \times W \times E}$, the convolution output $y_i$ at location $i$ can be formulated as:

$$y_i = \sigma(\sum_{j \in N_i} W_{ij}(G) x_j + b)$$  \hspace{1cm} (3)$$

where $W_{ij}$ is the kernel weights derived from the geometry guidance $G$. 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 [52] to decouple it as a dot-product of two sub-space embeddings:

$$W_{ij}(G) = \phi(G_i) \cdot \psi(G_j)$$  \hspace{1cm} (4)$$

where $\phi(\cdot)$ and $\psi(\cdot)$ denote the features in sub-embedding space. Then, the proposed geometry-aware convolution operation is defined as:

$$y_i = \sigma(\sum_{j \in N_i} \phi(G_i) \cdot \psi(G_j) \cdot x_j + b)$$  \hspace{1cm} (5)$$

In this paper, we achieve this convolution operation by several conventional convolutional layers and several basic element operations. Figure 4 gives an illustration of the proposed GAC convolution module. The geometric embeddings are first fed into two convolutional layers to get the sub-embeddings feature maps $\phi(G)$ and $\psi(G)$. Then the geometric similarity matrix $W_{ij}(G)$ is produced by dot product of two sub-embeddings feature maps. After that, the geometric affinity is fused with semantic features by another dot product operation. Finally, the fused information is combined with original contextual features through an element-wise summation to produce the final output. Note that the whole convolution process maintains the dimension and size of the contextual features.

D. Multi-task Objective Function

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

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda \mathcal{L}_{\text{h}}$$  \hspace{1cm} (6)$$

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

For the semantic segmentation task, existing methods mostly use cross-entropy loss to penalize the difference between the ground truth labels and predicted label maps. In this paper, we note that in remote sensing datasets, different semantic classes can have a very different number of pixels (e.g., the car category has a much smaller number of pixels than the vegetation category). To address this issue, we leverage the weighted cross-entropy loss as our loss function, where the inverse class frequencies are used as the balance weights for all pixels of that class. Our semantic segmentation loss function can be calculated as follows:

$$\mathcal{L}_{\text{seg}} = \sum_i w_i \sum_c \ell_i \log(p_{ic})$$  \hspace{1cm} (7)$$

where $i$ indicates the pixel location, $c$ is the category index, $\ell_i$ is the ground truth label of pixel $i$, $p_{ic}$ denotes the predicted probability of pixel $i$ belonging to class $c$, $w_i$ denotes the balance weight for pixel $i$.

In this paper, we use standard L1 loss to train height estimation network, calculated as:

$$\mathcal{L}_{\text{h}} = \sum_i |\hat{H}_i - H_i|$$  \hspace{1cm} (8)$$

where $\hat{H}_i$ and $H_i$ denote the predicted and ground truth height at pixel $i$. 

Fig. 4. Illustration of geometry-aware distillation module. $\otimes$ represents dot-product and $\oplus$ represents element-wise summation.
IV. EXPERIMENTS AND RESULTS

A. Datasets

To verify the effectiveness of our proposed model for the semantic labeling of remote sensing images, we conduct experiments on the ISPRS 2D Semantic Labeling Challenge dataset [53]. This dataset contains very high-resolution aerial images from two cities of Germany: Vaihingen and Potsdam. And for each aerial image, the ground truth labels are provided on six classes: buildings, impervious surfaces (e.g. roads), trees, low vegetation, cars, and clutter. The corresponding DSM information generated by dense image matching is also provided.

1) ISPRS Vaihingen: The Vaihingen dataset contains 33 image tiles with a spatial resolution of 9cm/pixel, and each image has a size of around $2500 \times 2500$ pixels. Each aerial image is composed of three channels of infrared, red, and green. Following the official split, 16 tiles with provided ground truth are used for model training, and the remaining 17 tiles are used for held-out evaluation by the challenger organizers. To enable a fair comparison with existing methods, we select 4 tiles (image numbers 5, 7, 23, 30) from the training split and use them for model evaluation.

2) ISPRS Potsdam: The Potsdam dataset contains 38 image tiles with a spatial resolution of 5cm/pixel, and each image has a size of $6000 \times 6000$ pixels. Each aerial image is composed of three channels of infrared, red, green, and blue. Following the official split, 24 tiles with provided ground truth are used for model training, and the remaining 14 tiles are used for held-out evaluation by the challenger organizers. To enable a fair comparison with existing methods, we select 4 tiles (image numbers 7_8, 4_10, 2_11, 5_11) from the training split and use them for model evaluation.

B. Implementation details

Our GANet model is implemented based on PyTorch Library. The network is optimized using a momentum SGD algorithm with the momentum set to 0.9. We train our model for 80 epochs with a cosine learning rate decay schedule. The initial and minimum learning rate are set to 1e-2 and 2e-5 respectively. We used a weight decay of 5e-4 for regularization. We train our model on 4 Tesla P100 GPUs with the batch size set to 4. We use Synchronized BN [54] after each convolutional layer.

Considering each image tile in both Vaihingen and Potsdam datasets has quite a large size, it can not be directly used for model training due to the GPU memory limit. An intuitive remedy to this problem is to divide the original image tiles into small patches. Previous methods mostly use a sliding window strategy to extract image patches. In this paper, we randomly crop a small patch of size $320 \times 320$ pixels ($512 \times 512$ pixels for Potsdam dataset) from each tile to formulate the training batch. By doing so, the input patches can be chosen from all possible positions in the image tile instead of the pre-defined locations when using a sliding window strategy. For data augmentation, we randomly flip the training patches horizontally and vertically with a probability of 0.5.

We choose the best hyper-parameter configurations based on the performance on the validation dataset and used it for online evaluation. In the test stage, we use a sliding window approach to generate small patches of size $320 \times 320$ pixels ($512 \times 512$ for Potsdam dataset). We set the sliding stride to 32 pixels to ensure overlaps between consecutive patches. We feed all image patches into our trained model and generate the probability map for each patch and average the probability values in overlapping regions. Considering objects in aerial images can have very different sizes, we use multi-scale inputs (scales of 0.8, 1, 1.2) to enhance the testing performance.

C. Evaluation metric

We use overall accuracy and per-class F1 score to validate the performance of our GANet model. The overall accuracy evaluates the classification performance for all semantic classes as a whole and is defined as the percentage of correctly classified pixels over all pixels. The F1 score is defined as the harmonic average of precision and recall of a given class and is calculated as follows,

\[
\text{precision} = \frac{TP}{TP + FP} \tag{9}
\]

\[
\text{recall} = \frac{TP}{TP + FN} \tag{10}
\]
\[ F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \] (11)

, where \( TP \) denotes the number of true positive pixels, \( FP \) denotes the number of false positive pixels and \( FN \) denotes the number of false negative pixels.

Moreover, to reduce the impact of uncertain border definitions during evaluation, we follow contest protocols to erode object boundaries by a 3px radius circle and ignoring those pixels during evaluation.

D. Comparing methods

In our experiments, we compare our GANet model with other methods submitted to the ISPRS 2D semantic labeling contest. The implementation details of each comparing method are listed below.

1) SVL_*: This is the baseline method provided by the challenge organizer [55]. This method takes as input both aerial images and several additional pre-calculated features, including SVL-features [56], normalized digital vegetation index (NDVI), saturation and normalized DSM (nDSM). The Adaboost algorithm is used as the base classifier and the conditional random field (CRF) model is used as postprocessing.

2) UZ_1: The method is developed by [57]. It takes as input both aerial images and nDSM images. This method uses an encoder-decoder network in which a successive of convolutional layers are used to learn high-level features and then deconvolution layers are used to recover the spatial resolution and produce the final segmentation outputs.

3) ADL_3: The method is developed by [58]. The model input combines aerial images, nDSM, and nDSM data. In this method, both CNN and Random forest (RF) classifiers are used to produce per-pixel classification probability maps and the results are fused. CRF is applied to further improve performance.

4) DST_2: The method is developed by [17]. It takes as input both aerial images and DSM data. A deep FCN model is proposed to obtain per-pixel labels with no downsampling and upsampling layers. CRF is used as a post-processing to enhance the performance.

5) DLR_10: This method is developed by [30]. It takes as input both aerial images and DSM data. This method combines semantic segmentation with edge detection in an end-to-end network using a SetNet-based architecture.

6) ONE_7: This method is developed by [59]. This method contains two multi-scale branches, one branch is trained with aerial images and the other is trained with the composite image of NDVI, DSM, and nDSM.

7) RIT_2: This method is developed by [60]. In this method, two independent convolutional layers are adopted to extract representative features from RGB images and composite images (IR, NDVI, and nDSM) respectively. The output feature maps are fused and fed into an FCN network for pixel-wise labeling.

8) SWJ_2: This method is developed by [61]. It uses ResNet-101 as the backbone network for multi-scale feature extraction and build a fully convolutional network that adaptively fuses multi-scale features. Only IRRG images are used as the inputs for model training and evaluation.

9) CASIA_2: This method is developed by [62]. It only takes aerial image data as input. This method is based on ResNet-101 architecture and it does not require the elevation data (DSM and nDSM) or any post-processing techniques.

10) TreeUNet: This method is developed by [63]. This method directly takes the aerial image data as well as DSM data as inputs for model training. No hand-crafted features or post-processing are used.

E. Results on Vaihingen

The quantitative performance of our GANet model on the ISPRS Vaihingen test set is shown in Table I. In this table, we compare our GANet model with other best-published methods on the ISPRS Vaihingen challenge. As shown in Table I, our GANet model achieves superior performance than all comparing methods with an overall accuracy of 91.1% and an average F1 score of 90.2%. The best-published CASIA2 model achieves quite close performance as our model. Note that CASIA2 pre-trains their model on PASCAL VOC 2012 dataset and then fine-tunes their model on the ISPRS Vaihingen dataset. Ablation analysis shows that CASIA2 model improves the performance a lot by using the finetune technique. While our GANet model is trained from scratch and does not need pretraining from other datasets.

Figure 6 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 3D geometric difference between these objects.

F. Results on Potsdam

We show the quantitative performance of our GANet model as well as the comparing methods on the ISPRS Potsdam test set in Table II. As can be seen in Table II, our GANet method achieves superior performance than all comparing methods on average F1 score and the second-best performance on overall accuracy. We note that SWJ_2 model obtains better performance than our GANet model on overall accuracy and CASIA_2 obtains quite close performance compared to our model. We owe this to the fact that these two models are pre-trained on PASCAL VOC 2012 dataset and then fine-tuned on ISPRS Potsdam dataset. While our model does not require additional dataset for pretraining. Besides, our GANet model only uses simple data augmentation strategy (i.e., randomly horizontal and vertical flip), while SWJ_2 and CASIA_2 models use more complicated data augmentation techniques.

Moreover, unlike previous methods (e.g., such as One_7 [5] and TreeUNet [63]) that mostly use elevation data (DSM and nDSM) as additional inputs, our GANet instead uses DSMs as side supervision during training. In the inference stage, our model only needs optical images as inputs and can surprisingly get better performance than previous methods.
Fig. 6. Selected examples of the segmentation results on ISPRS Vaihingen dataset.

| Method | Input | Imp. surf | Buildings | Low veg. | Trees | Cars | Overall | Average F1 |
|--------|-------|-----------|-----------|----------|-------|------|---------|------------|
| SVL_3  | IRRG+nDSM+NDVI+SVL | 86.6 | 91.0 | 77.0 | 85.0 | 55.6 | 84.8 | 79.0 |
| DST_2  | IRRG+DSM | 90.5 | 93.7 | 83.4 | 89.2 | 72.6 | 89.1 | 85.9 |
| UZ_1   | IRRG+nDSM | 89.2 | 92.5 | 81.6 | 86.9 | 57.3 | 87.3 | 81.5 |
| ONE_7  | IRRG+DSM+NDSM | 91.0 | 94.5 | 84.4 | 89.9 | 77.8 | 89.8 | 87.5 |
| DL_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 | 95.5 | 84.9 | 90.2 | 87.4 | 91.1 | 90.2 |

TABLE I

Classification performance on the Vaihingen test set. ‘DSM(s)’ indicates the model using DSM as additional supervision.

Fig. 7. Selected examples of the segmentation results on ISPRS Potsdam dataset.

We show some examples of the semantic labeling results of our GANet model and the compared methods on the Potsdam dataset in Figure 6. As shown in this figure, our GANet model obtains reasonable results on the ISPRS Potsdam dataset, which again validates the effectiveness of GANet method for aerial image classification. A detailed investigation of the qualitative results also shows the advantage of our GANet model for distinguishing objects with similar 2D appearance but different 3D geometric properties.

V. DISCUSSION

In this section, we conduct extensive experiments to validate the effect of the proposed height supervision module and the geometry-aware convolution module. We also investigate the performance of our GANet model under different network
configurations. All performances are reported on the validation set of ISPRS Vaihingen dataset.

### A. Effect of Height Supervision

First, we investigate the effectiveness of height supervision in comparison to methods that directly using DSM images as model input. To achieve this, we remove the height decoder branch as well as the GAC module from our GANet model and use it as a baseline model. This baseline model now becomes a traditional one-branch encoder-decoder network. We report the performance of the baseline model 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 $\lambda$. 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 improvement, which demonstrates the benefits of using height information as side supervision. Specifically, the baseline model obtains an overall accuracy of 90.7% and an average F1 score of 88.2% on the Vaihingen validation set, while our model with the height decoder branch achieves an overall accuracy of 91.3% and an average F1 score of 89.6% when $\lambda$ equals 1. Moreover, Table III also show that different values of $\lambda$ give similar performance. Our GANet model gets the best performance when $\lambda$ equals 1.

Figure 8 shows an example of the semantic segmentation results with and without the height estimation branch. As shown in this figure, by incorporation geometrical information from DSM images, our model successfully identifies building pixels that have a similar 2D appearance with impervious surface pixels. Moreover, benefiting from geometrical information, our model also achieves higher accuracy on vegetation-tree classification by distinguishing them using the height information.

| Method       | Input                        | Imp. surf. | Buildings | Low veg. | Trees | Cars | Overall | Average F1 |
|--------------|------------------------------|------------|-----------|----------|-------|------|---------|-------------|
| SVL_1        | IRRG+nDSM+NDVI               | 83.5       | 91.7      | 72.2     | 63.2  | 62.2 | 77.8    | 74.6        |
| DST_5        | IRRGB+DSM                    | 92.5       | 96.4      | 86.7     | 88.0  | 94.7 | 90.3    | 91.7        |
| UZ_1         | IRRG+nDSM                    | 89.3       | 95.4      | 81.8     | 80.5  | 85.6 | 85.8    | 86.7        |
| 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)                  | 92.8       | 96.7      | 87.3     | 88.5  | 96.1 | 91.1    | 92.6        |

**TABLE III**  
Classification performance on the Potsdam test set. ‘DSM(s)’ indicates the model using DSM as additional supervision.

### B. Effect of GAC Module

Then, we explore the benefits of using our newly proposed GAC module for contextual and geometrical feature fusion. We compare the performance of our model with and without the GAC module. Table IV lists the quantitative performance of our GANet model as well as the comparing methods. The results show that both the element-wise summation fusion strategy and the GAC module can improve the segmentation performance. More importantly, the proposed GAC module performs better than an element-wise summation fusion strategy. This demonstrates the superiority of our GANet model over V-FuseNet [5].

Figure 9 shows an example of the semantic segmentation results with different fusion strategies. As can be seen in Figure 9, 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 make better use of the geometrical information from the height decoder branch. Moreover, our model using the GAC module obtains better performance than its counterpart using an element-wise summation fusion strategy. The improvement mainly comes from a better classification between tree and low vegetation categories.

Furthermore, we compare the final validation performance of our GANet model with V-FuseNet [5] which uses DSM images as model inputs and fuse the feature embeddings from two separate encoder networks. From Table IV one can see that the feature fusion strategy in [5] gets an improvement of 0.9% on the overall accuracy and 1.0% on the average F1 score. While our GANet model achieves an improvement of 1.3% on overall accuracy and 2.5% on the average F1 score by using the height supervision and GAC module. This demonstrates that the proposed height supervision and GAC module can make better use of the geometrical information from DSM images. Note that our GANet model does not require DSM images in the test stage while V-FuseNet [5] needs DSM images as model input both in the training and test stage. This further demonstrates the superiority of our GANet model over V-FuseNet [5].

Moreover, we also investigate the effectiveness of a multi-scale test strategy. As shown in Table IV, our GANet model
enjoys a slight performance boost on both overall accuracy and average F1 score.

| Method                        | Overall | Average F1 |
|-------------------------------|---------|------------|
| SegNet [5]                    | 90.2    | 89.3       |
| V-FuseNet [5]                 | 91.1    | 90.3       |
| 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    | 91.0       |

**Table IV**

Effect of GAC module. Performances are reported on the Vaihingen validation set. 'Sum Fusion' denotes our model using element-wise summation for feature fusion. 'MS Test' denotes multiscale test.

### C. Effect of Network Depth

Moreover, we investigate the performance of our GANet model with different network depth. Table V shows the performance of our GANet model using ResNet-50, ResNet-101 and ResNet-152. Detailed configurations of these architectures can be found in [48]. As shown in Table V, our GANet model obtains the best performance when using the ResNet-101 architecture. Our model with ResNet-50 architecture leads to inferior performance due to limited feature extraction abilities, while our model with ResNet-152 architecture suffers from the over-fitting problem and the performance decreases.

| Method  | Overall | Average F1 |
|---------|---------|------------|
| ResNet-50 | 91.5    | 89.8       |
| ResNet-101 | **92.3** | **91.0**   |
| ResNet-152 | 92.0    | 90.6       |

**Table V**

Semantic segmentation performance on ISPRS Vaihingen validation set with different network depths.

### D. Height Estimation Performance

To demonstrate that our GANet model can learn geometrical features and predict height maps, we visualized the predicted height maps on the ISPRS Vaihingen validation set in Figure 10. As shown in Figure 10, our GANet model obtains satisfying height estimation performance on the ISPRS Vaihingen dataset. Specifically, our GANet model predicts larger height values for building and tree pixels while it predicts smaller height values for those pixels from the impervious surface, low vegetation, and car categories. Note that, our GANet model only predicts the relative height values in the range of $[0,1]$, while the ground truth height maps show real-value altitude. For example, there's a height lifting from bottom to top in
the ground truth DSM image in Figure 10. One can use post-processing techniques (e.g., [45]) to merge the predicted height image patches and generate absolute height maps.

VI. CONCLUSIONS

In this paper, we introduce a geometry-aware convolutional neural network to approach the problem of semantic segmentation of remote sensing images. Our model benefits from the 3D geometric information via implicit height estimation. Unlike previous methods that mostly use a single decoder network to predict pixel-wise semantic labels, in our model, a newly designed height decoder branch is developed to predict the height map under the supervision of DSM images. The height decoder branch is trained to be capable of distilling 3D geometric features from 2D contextual/appearance features. In this way, our model does not require DSM as model input and still benefits from the helpful 3D geometric information. Moreover, we develop a new geometry-aware convolution module to combine the learned 3D geometric features and 2D contextual features from two decoder branches. With the fused feature embeddings, our model can produce geometry-aware segmentation with enhanced performance. In the training stage, our model uses DSM images as side supervision to enforce geometric feature distillation, while in the inference stage, it does not need DSM data and can directly produce the segmentation labels in an end-to-end fashion. Experiments on ISPRS Vaihingen and Potsdam datasets demonstrate the effectiveness of our proposed method for aerial image classification. Our proposed method achieves remarkable performance on these two datasets without using any hand-crafted features or post-processing.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Q. Zou, L. Ni, T. Zhang, and Q. Wang, “Deep learning based feature selection for remote sensing scene classification,” IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 11, pp. 2331–2335, 2015.
[2] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, “When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns,” IEEE transactions on geoscience and remote sensing, vol. 56, no. 5, pp. 2811–2821, 2018.
[3] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, “Convolutional neural networks for large-scale remote-sensing image classification,” IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 2, pp. 645–657, 2016.
[4] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, “Classification with an edge: improving semantic image segmentation with boundary detection,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 135, pp. 158–172, 2018.
[5] N. Audebert, B. Le Saux, and S. Lefèvre, “Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 140, pp. 26–32, 2018.
Fig. 10. Height estimation on ISPRS Vaihingen dataset.

[6] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, “Vehicle detection in satellite images by hybrid deep convolutional neural networks,” IEEE Geoscience and Remote Sensing Letters, vol. 11, no. 10, pp. 1797–1801, 2014.

[7] Y. Hu, X. Li, N. Zhou, L. Yang, L. Peng, and S. Xiao, “A sample update-based convolutional neural network framework for object detection in large-area remote sensing images,” IEEE Geoscience and Remote Sensing Letters, vol. 16, no. 6, pp. 947–951, 2019.

[8] V. Mnih, “Machine learning for aerial image labeling,” Ph.D. dissertation, University of Toronto (Canada), 2013.

[9] S. Saito, T. Yamashita, and Y. Aoki, “Multiple object extraction from aerial imagery with convolutional neural networks,” Electronic Imaging, vol. 2016, no. 10, pp. 1–9, 2016.

[10] R. Alshbibi, P. R. Marpu, W. L. Woon, and M. Dalla Mura, “Simultaneous extraction of roads and buildings in remote imagery with convolutional neural networks,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 130, pp. 139–149, 2017.

[11] X. Li, X. Yao, and Y. Fang, “Building-a-nets: Robust building extraction from high-resolution remote sensing images with adversarial networks,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 11, no. 10, pp. 3680–3687, 2018.

[12] F. P. Luus, B. P. Salmon, F. Van den Bergh, and B. T. J. Maharaj, “Multiview deep learning for land-use classification,” IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 12, pp. 2448–2452, 2015.

[13] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, “Land use classification in remote sensing images by convolutional neural networks,” arXiv preprint arXiv:1508.00092, 2015.

[14] Z. Yang, W. Jiang, B. Xu, Q. Zhu, S. Jiang, and W. Huang, “A convolutional neural network-based 3D semantic labeling method for ALS point clouds,” Remote Sensing, vol. 9, no. 9, p. 936, 2017.

[15] R. Zhao, M. Pang, and J. Wang, “Classifying airborne lidar point clouds via deep features learned by a multi-scale convolutional neural network,” International Journal of Geographical Information Science, vol. 32, no. 5, pp. 960–979, 2018.

[16] C. Wen, L. Yang, L. Peng, and T. Chi, “Directionally constrained fully convolutional neural network for airborne lidar point cloud classification,” arXiv preprint arXiv:1908.06673, 2019.

[17] J. Sherrah, “Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery,” arXiv preprint arXiv:1606.02585, 2016.

[18] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 12, pp. 2481–2495, 2017.

[19] L. Zhou, C. Zhang, and M. Wu, “D-linknet: Linknet with pretrained encoder and dilated convolution for high-resolution satellite imagery road extraction.” in CVPR Workshops, 2018, pp. 182–186.

[20] X. Wei, K. Fu, X. Gao, M. Yan, X. Sun, K. Chen, and H. Sun, “Semantic pixel labelling in remote sensing images using a deep convolutional encoder-decoder model,” Remote Sensing Letters, vol. 9, no. 3, pp. 199–209, 2018.

[21] J. Yuan, “Learning building extraction in aerial scenes with convolutional networks,” IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 11, pp. 2793–2798, 2017.

[22] D. Chai, S. Newsum, and J. Huang, “Aerial image semantic segmentation using dcn predicted distance maps,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 161, pp. 309–322, 2020.

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

[24] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep ordinal regression network for monocular depth estimation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2002–2011.

[25] L. Mou and X. X. Zhu, “Im2height: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network,” arXiv preprint arXiv:1802.10249, 2018.

[26] P. Ghamisi and N. Yokoya, “Im2dsm: Height simulation from single imagery using conditional generative adversarial net,” IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 5, pp. 794–798, 2018.

[27] V. Mnih and G. E. Hinton, “Learning to detect roads in high-resolution aerial images,” in European Conference on Computer Vision. Springer, 2010, pp. 210–227.

[28] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, “Building detection in very high resolution multispectral data with deep learning features,” in 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE, 2015, pp. 1873–1876.

[29] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[30] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, “Semantic segmentation of aerial images with an ensemble of cnns,” ISPRS Journal of Photogrammetry and Spatial Information Sciences, vol. 3, p. 473, 2016.

[31] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, “ Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order crfs,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 76–85.

[32] M. Volpi and D. Tuia, “Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images,” ISPRS journal of photogrammetry and remote sensing, vol. 144, pp. 48–60, 2018.

[33] Y. Shao, G. N. Taff, and S. J. Walsh, “Shadow detection and building-height estimation using ikonos data,” International journal of remote sensing, vol. 32, no. 22, pp. 6929–6944, 2011.

[34] A. Comber, M. Umezaki, R. Zhou, Y. Ding, Y. Li, H. Fu, H. Jiang, and A. Teklesbury, “Using shadows in high-resolution imagery to determine building height,” Remote sensing letters, vol. 3, no. 7, pp. 551–556, 2012.

[35] T. Kim, T. Javzandulam, and T.-Y. Lee, “Semiautomatic reconstruction of building height and footprints from single satellite images,” in 2007 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2007, pp. 4737–4740.

[36] Z. Chen, Q. Qin, L. Lin, Q. Liu, and W. Zhan, “Dem densification using perspective shape from shading through multispectral imagery,” IEEE Geoscience and Remote sensing letters, vol. 10, no. 1, pp. 145–149, 2012.
[37] M. A. Rajabi and J. R. Blais, “Optimization of dtm interpolation using sfs with single satellite imagery,” The Journal of Supercomputing, vol. 28, no. 2, pp. 193–213, 2004. 2

[38] D. Eigen and R. Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2650–2658. 3

[39] Y. Kuznetsova, T. Stuckler, and B. Leibe, “Semi-supervised deep learning for monocular depth map prediction,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6647–6655. 3

[40] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, “Deeper depth prediction with fully convolutional residual networks,” in 2016 Fourth international conference on 3D vision (3DV). IEEE, 2016, pp. 239–248. 3

[41] J. Li, R. Klein, and A. Yao, “A two-streamed network for estimating fine-scale depth maps from single rgb images,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3372–3380. 3

[42] F. Liu, C. Shen, G. Lin, and I. Reid, “Learning depth from single monocular images using deep convolutional neural fields,” IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 10, pp. 2024–2039, 2015. 3

[43] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, “Towards unified depth and semantic prediction from a single image,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2800–2809. 3

[44] R. Garg, V. K. BG, G. Carneiro, and I. Reid, “Unsupervised cnn for single view depth estimation: Geometry to the rescue,” in European Conference on Computer Vision. Springer, 2016, pp. 740–756. 3

[45] H. A. Amirkolaee and H. Arefi, “Height estimation from single aerial images using a deep convolutional encoder-decoder network,” ISPRS journal of photogrammetry and remote sensing, vol. 149, pp. 50–66, 2019. 3, 11

[46] S. Srivastava, M. Volpi, and D. Tuia, “Joint height estimation and semantic labeling of monocular aerial images with cnns,” in 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), July 2017, pp. 5173–5176. 3

[47] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818. 4

[48] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778. 4, 10

[49] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” International journal of computer vision, vol. 111, no. 1, pp. 98–136, 2015. 4

[50] W. Wang and U. Neumann, “Depth-aware cnn for rgb-d segmentation,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 135–150. 5

[51] Y. Chen, T. Mensink, and E. Gavves, “3d neighborhood convolution: Learning depth-aware features for rgb-d and rgbd semantic segmentation,” in 2019 International Conference on 3D Vision (3DV). IEEE, 2019, pp. 173–182. 5

[52] J. Jiao, Y. Wei, Z. Jie, H. Shi, R. W. Lau, and T. S. Huang, “Geometry-aware distillation for indoor semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2869–2878. 5

[53] J. Niemeyer, F. Rottensteiner, and U. Soergel, “Contextual classification of lidar data and building object detection in urban areas,” ISPRS journal of photogrammetry and remote sensing, vol. 87, pp. 152–165, 2014. 6

[54] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, “Context encoding for semantic segmentation,” in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 7151–7160. 6

[55] M. Gerke, “Use of the stair vision library within the isprs 2d semantic labeling benchmark (vaihingen),” 2014. 7

[56] S. Gould, O. Russakovsky, I. Goodfellow, P. Baumstark, A. Ng, and D. Koller, “The stair vision library (v2. 4),” 2009. 7

[57] M. Volpi and D. Tuia, “Dense semantic labeling of subdecimeter resolution images with convolutional neural networks,” IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 2, pp. 881–893, 2017. 7

[58] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Van Den Hengel, “Semantic labeling of aerial and satellite imagery,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 9, no. 7, pp. 2868–2881, 2016. 7

[59] N. Audebert, B. Le Saux, and S. Lefèvre, “Semantic segmentation of earth observation data using multimodal and multi-scale deep networks,” in Asian conference on computer vision. Springer, 2016, pp. 180–196. 7

[60] S. Piramanyagam, W. Schwartzkopf, F. Koehler, and E. Saber, “Classification of remote sensed images using random forests and deep learning framework,” in Image and Signal Processing for Remote Sensing XXII, vol. 10004. International Society for Optics and Photonics, 2016, p. 100040L. 7

[61] 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 Sensing, vol. 11, no. 13, p. 1617, 2019. 7

[62] 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 journal of photogrammetry and remote sensing, vol. 145, pp. 78–95, 2018. 7

[63] 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 Journal of Photogrammetry and Remote Sensing, vol. 156, pp. 1–13, 2019. 7