

JSH-Net: joint semantic segmentation and height estimation using deep convolutional networks from single high-resolution remote sensing imagery

Bin Zhang, Yi Wan, Yongjun Zhang & Yansheng Li

To cite this article: Bin Zhang, Yi Wan, Yongjun Zhang & Yansheng Li (2022) JSH-Net: joint semantic segmentation and height estimation using deep convolutional networks from single high-resolution remote sensing imagery, International Journal of Remote Sensing, 43:17, 6307-6332, DOI: [10.1080/01431161.2022.2135410](https://doi.org/10.1080/01431161.2022.2135410)

To link to this article: <https://doi.org/10.1080/01431161.2022.2135410>



Published online: 09 Nov 2022.



Submit your article to this journal [↗](#)




View related articles [↗](#)



View Crossmark data [↗](#)



JSH-Net: joint semantic segmentation and height estimation using deep convolutional networks from single high-resolution remote sensing imagery

Bin Zhang , Yi Wan, Yongjun Zhang and Yansheng Li

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China

ABSTRACT

Semantic segmentation for high-resolution remote sensing imagery is a pivotal component of land use and land cover categorization, and height estimation is essential for rebuilding the 3D information of an image. Because of the higher intra-class variation and smaller inter-class dissimilarity, these two challenging tasks are generally treated separately. This paper proposes a fully convolutional network that can tackle these problems simultaneously by estimating the land-cover categories and height values of pixels from a single aerial image. To handle these tasks, we develop a multi-task learning architecture (JSH-Net) that employs a shared feature representation and exploits their potential consistency across tasks, resulting in robust features and better prediction accuracy. Specifically, we propose a novel skip connection module that aggregates the contexts from the encoder part to the decoder part, bridging the semantic gap between them. In addition, we propose a progressive refinement strategy to recover detailed information about the objects. Moreover, we also proposed a height estimation branch on the head of the model to utilize shared features. The experiments we conducted on ISPRS 2D Labelling dataset verified that our network provided precise results of semantic segmentation and height estimation from two output branches and outperformed other state-of-the-art approaches.

ARTICLE HISTORY

Received 25 March 2022
Accepted 1 October 2022

KEYWORDS

remote sensing; semantic segmentation; height estimation; deep learning; multi-task learning; convolutional neural network

1. Introduction

High-resolution remote sensing (RS) image interpretation and recognition are among of the most touchy topics in the RS field (Ball, Anderson, and Chan 2017; Ma et al. 2019; Zhang, Zhang, and Du 2016). Among them, one of the most difficult issues in the RS field is the semantic labelling and height estimation of high-resolution RS images (Amirkolae and Arefi 2019; Marmanis et al. 2018). Semantic segmentation of high-resolution RS imagery plays a crucial role in land use and land cover categorization (LULC) (Marcos et al. 2018), building segmentation (Maggiori et al. 2017a), road detection (Mnih 2013), and change detection (Zheng et al. 2021), etc. Height estimation also is necessary for rebuilding the 3D information of an image (Srivastava, Volpi, and Tuia 2017). Due to

higher intra-class variance and smaller inter-class dissimilarity of high-resolution RS imagery, semantic segmentation and height estimation are both challenging tasks.

For semantic segmentation, the specific regions of an image are labelled according to their class based on what is being shown (Thoma 2016). To put it another way, every pixel in the image is labelled with the corresponding class of visual appearance. As a result, semantic segmentation is a pixel-level categorization task. High-resolution RS images with sub-metre ground sampling distance (GSD) have recently been accessible, making it difficult to distinguish things such as roads and buildings based on their spectral signature. Therefore, semantic segmentation requires richer semantic representation as well as contextual information (Sherrah 2016). For height estimation, the goal for each pixel in an image is to assign height values between the ground and objects (Mou and Zhu 2018). In RS, the relative height images are usually named as the normalized digital surface model (nDSM). A similar problem in computer vision is depth estimation. There are several methods for obtaining elevation data in photogrammetry, such as stereo matching for pair-wise images or airborne LiDAR point clouds, which are not always possible. This begs the question of if height values can be estimated from a single image. However, since there are estimating height values from the monocular image is problematic, the task is inherently ambiguous and there is a large source of uncertainty (Eigen, Puhrsch, and Fergus 2014). However, we believe that certain clues, such as object size, perspective change, texture, shading, object occlusion, the effects of atmosphere, and so on, may be used to estimate height values from a single image. (Eigen, Puhrsch, and Fergus 2014). Thus, one of the most important aspects of height estimation is capturing the long-range contextual knowledge to model these cues between objects.

Deep convolutional neural networks (CNN) are currently driving advances in image classification (He et al. 2016; Simonyan and Zisserman 2015) and semantic segmentation (Long, Shelhamer, and Darrell 2015) in computer vision and are achieving state-of-the-art results. Simultaneously, CNNs also are applied for depth estimation (Eigen, Puhrsch, and Fergus 2014). Semantic segmentation models in computer vision have been employed for high-resolution RS imaging, based on the effectiveness of FCN-based semantic segmentation on natural images (Sherrah 2016). RS data is multimodal when compared to natural images due to the unique characteristics of RS. The fusion of multimodal information is now considered the typical scenario in the exploitation of RS semantic segmentation. Thus, many prior works have used elevation data (DSM or nDSM) as the input of the network by using the dual-input network (or the Siamese network) (Audebert, Le Saux, and Lefèvre 2018; Audebert, Saux, and Lefèvre 2016; Marmanis et al. 2018, 2016; Paisitkriangkrai et al. 2016) or stacking together with multispectral imagery (Liu et al. 2018a, 2017a; Maggiori et al. 2017b; Marcos et al. 2018; Nogueira et al. 2019; Volpi and Tuia 2017). However, the following problems remain to be solved for RS semantic segmentation. (1) For objects which present different visual characteristics in high-resolution imagery, such as buildings, it is hard to obtain a correct class. (2) Small objects, such as cars, due to CNN continuously reducing the size of the feature maps, are too small to obtain a precise mask. (3) Last but not least, a key issue remains as far as how multi-source data can be used effectively. For example, some methods use height data as input, which requires a delicate model design to fuse RGB images and height data; furthermore, their models can be used only for images with elevation data only when this data is available.

On the one hand, semantic segmentation extracts the semantic properties of objects from images, on the other hand, height estimation focuses on geometric properties. Both of them require rich contextual information and can be modelled as a pixel-wise labelling problem. With a trainable network, CNNs can extract high-level features in an end-to-end manner. To simulate the relationship between the image and various tasks, we employed CNN to learn the complicated nonlinear mapping. In the past, these two tasks were addressed separately. However, they actually are complementary and consistent. Specifically, semantic segmentation and height estimation both require rich contextual information to extract high-level features. Multi-task learning can train multiple different but common tasks simultaneously by leveraging shared feature representation. As a consequence, we apply multi-task learning to solve them all at once by leveraging their similarities across tasks, resulting in robust features and higher prediction accuracy.

We propose a fully convolutional network based on an encoder-decoder topology to solve the problems encountered by earlier studies. Specifically, to extract the abstract features, the encoder network is a CNN pre-trained on the large image classification dataset, with the last two stages modified by dilated convolution to maintain the shape of the feature map identical and the spatial information preserved. Then, to recover the detailed information, we use a progressive refinement strategy in the decoder network. In this process, the high-level features from the decoder network and corresponding high-resolution features from the encoder network are fused in this method to generate a feature map with high resolution. Unfortunately, combining these two types of feature maps by copying original features or using a simple 1×1 convolution does not provide enough detailed and contextual information. Thus, we propose a dilated pyramid skip connection module to relieve the semantic mismatch between the encoder layers and the decoder layers. Finally, for the semantic segmentation and height estimation tasks, we combine the features from various layers to create a shared feature representation. We used the ISPRS 2D Labelling dataset as experimental data, and the results indicated that our multi-task learning network outperforms existing state-of-the-art approaches.

In summary, the following are the significant contributions of our proposed method:

- (1) To aggregate the contexts from the encoder to the decoder, a dilated pyramid skip connection module is proposed. The semantic gap between the layers is well relieved by making use of contextual information, which makes confusing objects distinguishable.
- (2) A multi-task learning network JSH-Net is proposed for semantic segmentation and height estimation from monocular high-resolution RS images. It achieved state-of-the-art performance on two challenging benchmarks: Vaihingen and Potsdam datasets in ISPRS Semantic Labelling Challenge. For RS image labelling and height estimation, our network establishes a new baseline.

The following is a description of the paper's organization. Recent methods and developments in RS image labelling and height estimation are discussed in [Section 2](#). In [Section 3](#), the suggested multi-task learning network is explained. The experimental data collection and analysis are given in [Section 4](#). Finally, in [Section 5](#), we present our conclusions and future research.

2. Related work

2.1. Semantic segmentation

2.1.1. In the computer vision field

Fully convolution network (FCN) based methods (Badrinarayanan, Kendall, and Cipolla 2017; Chen et al. 2018b; Long, Shelhamer, and Darrell 2015; Ronneberger, Fischer, and Brox 2015; Zhao et al. 2017) can achieve effective feature extraction and end-to-end training and thus have become the most optimal choice for semantic segmentation. In FCN (Long, Shelhamer, and Darrell 2015), the fully connected layers were converted to convolutional layers and the last feature map was upsampled to match the original input size. Further, on the PASCAL VOC and Cityscapes datasets, a series of works based on FCN have achieved state-of-the-art performance. Contextual information played a crucial role in a variety of vision tasks, especially semantic segmentation. To extract more effective contexts and to alleviate the problem of limited receptive fields, dilated convolutions (or atrous convolutions) were widely used (Chen et al. 2017a, 2017b, 2018b). To compensate for too small and coarse output feature maps, many works have used encoder-decoder structure to refine spatial information gradually (Badrinarayanan, Kendall, and Cipolla 2017; Lin et al. 2017; Ronneberger, Fischer, and Brox 2015). Pyramid modules, such as atrous spatial pyramid pooling (ASPP) (Chen et al. 2017a, 2017b, 2018b) and pyramid pooling module (Zhao et al. 2017), were designed to separate multi-scale features and to embed the contextual information. Recently, attention modules also have been introduced in semantic segmentation to model the long-range dependencies in spatial and channel dimensions and then capture useful contexts and extract discriminative features (Fu et al. 2019; Zhang et al. 2018).

2.1.2. In remote sensing field

Some past works combined CNN and hand-crafted features, and they frequently employed post-processing, such as CRF, to refine the final results (Liu et al. 2017b; Paisitkriangkrai et al. 2015). Paisitkriangkrai et al. (2015) proposed a semantic labelling network using CNN features, hand-crafted features, and CRFs as post-processing, but their predictions only classified the centre pixel every time, which led to excessive redundant calculations.

To use multi-sensor data as input, many researchers have combined both the image and the DSM data to offer more information to improve performance. Some methods adopted the Siamese network for working on two different inputs, e.g. visible images and DSM, see Figure 1(a). Sherrah (2016) proposed a network, in which the aerial images were used as input of a pretrained VGG network, and the DSM data was used as another FCN trained from scratch, then the feature maps from these two networks were concatenated to predict the label. Marmanis et al. (2016) presented a Siamese network to integrate the images and the DSM data. Their improved version suggested a network that included edge detection and semantic segmentation. However, their model is complicated and requires phased training (Marmanis et al. 2018). Audebert, Saux, and Lefèvre (2016) presented a variant encoder-decoder model with a multi-kernel layer for merging predictions from multiple scales. Then, they created a new network by using FuseNet to

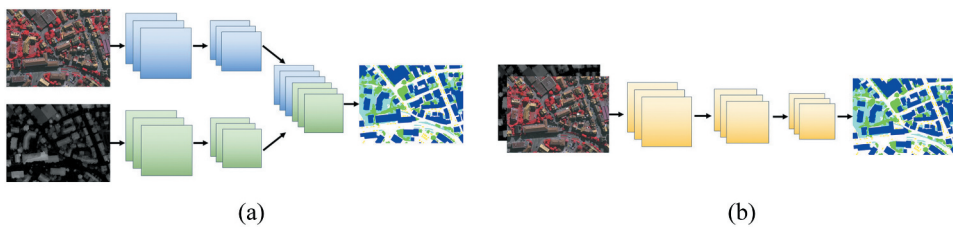


Figure 1. (a). Using multi-sensor data by Siamese network fashion. (b) Using multi-sensor data by multi-channel input network.

accomplish semantic labelling of multi-modal multi-scale RS data (Audebert, Le Saux, and Lefèvre 2018).

Other methods exploited multi-modal data by stacking them together as the input of the network, see Figure 1(b). Volpi and Tuia (2017) presented an encoder-decoder network, in which the deconvolutions were used to upsample feature maps to the original input size. Maggiori et al. (2017b) proposed a CNN framework, which combined different layers to obtain hypercolumn features. Liu et al. (2017a) proposed an encoder-decoder network to predict semantic labels, and their subsequent work then introduced a novel edge loss function to increase the segmentation accuracy at the edge (Liu et al. 2018a). Marcos et al. (2018) presented a CNN to enhance the rotation equivariance of the neural network. Nogueira et al. (2019) proposed a FCN, which was trained with different sizes of images to capture multi-scale features and extract context information.

In addition, some researchers who only used image data as input also obtained remarkable results. Wang et al. (2017) proposed a gated segmentation network for adaptive information propagation between different levels of feature maps progressively. Chen et al. (2018a) introduced two semantic segmentation frameworks with dense residual connection modules. Bui et al. (2018) proposed a neural network based on FCN and neural search network architecture. Liu et al. (2018b) proposed a FCN, which successively aggregated contexts from large to small scale. Both Bai et al. (2021) and Li, Lei, and Kuang (2021) proposed a module to extract multi-scale contextual features to improve the accuracy of semantic segmentation. Wang et al. (2021) used a new transformer as the backbone to capture long-term dependencies.

2.2. Height estimation

In the computer vision field, depth estimation is most related to height estimation. Before the deep learning methods brought the breakthroughs, depth estimation from a single image was generally formulated with a probabilistic graphical model (Liu, Salzmann, and He 2014; Saxena, Chung, and Ng 2005, 2008). Eigen, Puhrsch, and Fergus (2014) solved this task for the first time by stacking two deep networks. In their extended work, a single multi-scale CNN was proposed to address three different computer vision tasks (Eigen and Fergus 2015).

Few studies have focused on the height estimation of a single high-resolution RS image until recently. Srivastava, Volpi, and Tuia (2017) first estimated the land-cover types and height values of pixels simultaneously from a single RS image by utilizing multi-task

learning. Our work in this paper extended their work to further prove the superiority of multi-task learning. Ghamisi and Yokoya (2018) used a generative adversarial network to predict the high values from a single image. Mou and Zhu (2018) proposed an encoder-decoder network to learn the connection between the single RS images and height data. Amirkolaei and Arefi (2019) proposed an encoder-decoder CNN to estimate the height values from a single image. Liu et al. (2021) proposed a height-embedding context reassembly network to predict semantic labels and height values.

3. The proposed method

In this section, the presented multi-task learning network architecture (denoted as JSH-Net) for semantic segmentation and height estimation from the single high-resolution RS images is illustrated. Our proposed network has two basic components: an encoder network and a decoder network. See Figure 2.

3.1. Encoder network

The encoder network was built using a deep neural network pre-trained on ImageNet. Generally, the CNN models used for the classification task are not suitable for the dense prediction task. To use these models, the final fully-connected layers are deprecated and previous convolution layers are used for extracting the high-level abstract features. In this paper, for fair comparison with other methods, we used VGG (Simonyan and Zisserman 2015) and ResNet (He et al. 2016) as the backbone for our encoder network. The output feature map of CNN is 32 times smaller than the input resolution size in the conventional image classification task, which is harmful for semantic segmentation. Due to the successive pooling and convolutions with the striding operation, detailed information related to the objects is missing. For example, the width of a car usually is about 20 pixels in the Vaihingen dataset, which is invisible after reducing it 32 times. Thus, state-of-art networks for semantic segmentation usually have a down-sampling rate of 8, which benefits from dilated convolution. For example, in DeepLabV3 the backbone adopted different atrous rates by the multi-grid method (Chen et al. 2017b). To preserve small and thin objects in images and alleviate the grid effect, we used the hybrid dilated convolution in the

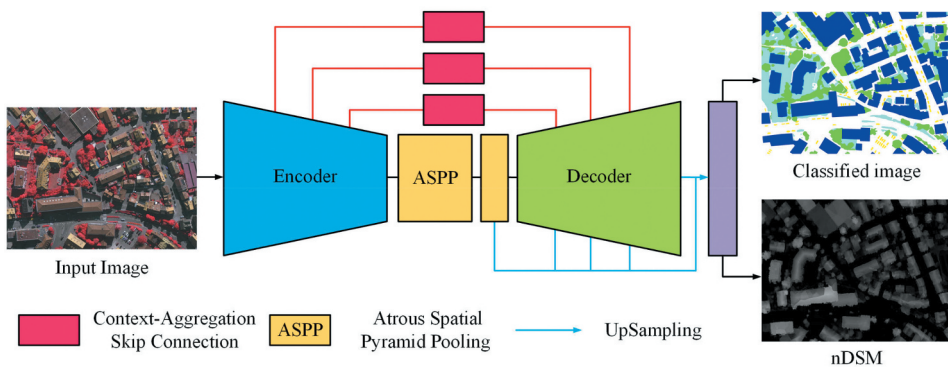


Figure 2. The general network structure for joint semantic segmentation and height estimation.

encoder network. Specifically, for the fourth stage in ResNet, we grouped every 4 blocks and changed their dilation rates to 1, 2, 5, and 9, respectively. For the fifth stage, we set the dilation rates to 5, 9, and 17. Therefore, the output stride of our encoder was 8.

After the encoder network, we used the atrous spatial pyramid pooling (ASPP) module to capture rich context information. Then, a 1×1 convolution was employed to reduce the channel dimension. The output denotes $F_{ASPP} \in R^{C' \times H/8 \times W/8}$.

3.2. Context-aggregation skip connection (CASC)

In U-Net, the skip connection was first proposed to connect the encoder part and the corresponding decoder part to recover detailed information (Ronneberger, Fischer, and Brox 2015). Since then, this structure generally has been used in many related fields. A skip connection is usually implemented by copying features directly or using a 1×1 convolution. However, this is too simple to capture enough context and detailed information. Since encoder features are low-level and their corresponding decoder features are high-level, thus there is a semantic mismatch. A simple skip connection can hinder the network from extracting the context information correctly.

To handle this problem, inspired by the ASPP module, we proposed a novel skip connection architecture CASC to alleviate the gap between the encoder and the decoder by capturing the multi-scale features. As illustrated in Figure 3(a), given feature maps $f \in R^{C \times h \times w}$ from the encoder, we first fed it into a 1×1 convolution to generate new features $f_1 \in R^{C'/4 \times h \times w}$. Then, we fed f into a small parallel network with three paths, which included three average pooling operations with different kernel sizes and three 3×3 dilated convolutions with various dilation rates. Specifically, for average pooling operations, the kernel sizes were set to s_1, s_2, s_3 , and the stride was set to 1. For the dilated convolution operations, the dilation rates were set to d_1, d_2 , and d_3 . After that, we had three new feature maps $\{f_2, f_3, f_4\} \in R^{C'/4 \times h \times w}$. Finally, we aggregated these feature maps f_1, f_2, f_3 , and f_4 by a concatenation operation to obtain the final feature map $f' \in R^{C' \times h \times w}$. Compared to copy features directly in the original skip connection, our context aggregation skip connection module has three extra paths to capture context information to minimize the gap between the encoder and decoder

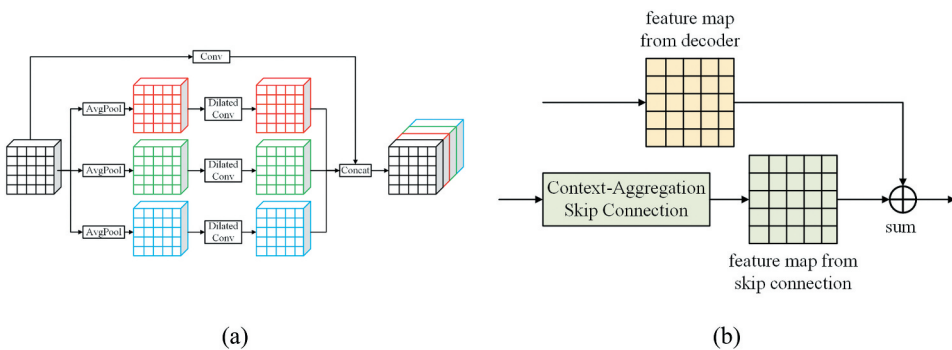


Figure 3. (a) Context-aggregation skip connection. (b) Block in decoder part.

network. In our approach, the s_1, s_2 and s_3 are set to 3, 5, 7, and $d_1, d_2,$ and d_3 are set to 4, 8, 12.

3.3. Decoder network

The decoder network is used for restoring feature maps to the input size and mapping features from the feature domain to the label domain. To restore the edge information accurately, we adopted a direct progressive upsampling approach. As illustrated in Figure 3(b), we summed the feature maps from the skip connection and the output of the last decoder network for simplicity.

At the same time, the feature maps from the ASPP module F_{ASPP} and decoder part ($\{F_2, F_3, F_4\} \in R^{C \times H/8 \times W/8}$) were interpolated to stride 4 to match the size of F_1 . After that, F_{ASPP}, F_4, F_3, F_2 and F_1 were aggregated to generate the final feature map $F' \in R^{5C \times H/4 \times W/4}$. To conduct semantic segmentation and height estimation simultaneously, we used two 1×1 convolutions to generate a class response map and to regress the height values. Finally, two output feature maps were interpolated to the input size. The output feature maps denote $F_{cls} \in R^{C \times H \times W}$ and $F_{height} \in R^{1 \times H \times W}$, respectively.

3.4. Loss function

To train a network to handle semantic segmentation and height estimation simultaneously, our loss function includes two parts: \mathcal{L}_{seg} and \mathcal{L}_{height} . For segmentation, we employ the usual cross-entropy loss. The sum of the L1 and L2 losses is used to estimate height. An additional scaling parameter, λ , has been incorporated into our total loss. For simplicity, we set its value to 1.

$$\mathcal{L}_{total}(I, G_s, G_h, \theta) = \mathcal{L}_{seg} + \lambda \cdot \mathcal{L}_{height} \quad (1)$$

$$\mathcal{L}_{seg}(I, G_s, \theta) = \frac{1}{n} \sum_{i \in I} -\log p_i = \frac{1}{n} \sum_{i \in I} -\log \frac{e^{F_{cls}^{G_s^i}}}{\sum_{k=1}^C e^{F_{cls}^k}} \quad (2)$$

$$\mathcal{L}_{height}(I, G_h, \theta) = \frac{1}{n} \sum_{i \in I} \left(|F_{height} - G_h| + |F_{height} - G_h|^2 \right) \quad (3)$$

Where I, G_s and G_h denote input image, ground truth segmentation mask, and height map, correspondingly; θ denotes weights of the network; p_i denotes probability when the class of pixel i belongs to G_s^i ; C denotes the number of classes.

4. Experiments and analysis

4.1. Dataset

To validate the performance of the proposed multi-task learning network, we tested it on the Vaihingen dataset and Potsdam dataset. The two datasets were classified into six of the most common land cover classes: impervious surfaces, building, low vegetation, tree,

car, and clutter/background. In semantic segmentation, there must be some error in the categories of manual labelling, especially in the adjacent edges between categories. To reduce the impact of uncertainty in the classification of the edge during the evaluation, the benchmark also provided eroded label images where the edges of the objects were eroded 3 pixels. Those boundaries were ignored during the inference.

4.1.1. *Vaihingen dataset*

This dataset consists of 33 images of various sizes, each consisting of an image cropped from a large aerial true orthoimage. The dataset includes three-band images which correspond to the near-infrared (IR), red (R), green (G) bands, and DSM. The normalized DSM (nDSM) data also was provided for our experiments. The GSD of both, the TOP and the nDSM, was 9 cm. Following the works of (Liu et al. 2018a, 2017a; Maggiori et al. 2017b; Marcos et al. 2018; Sherrah 2016; Volpi and Tuia 2017), 11 images were utilized in the training set, while 5 images were used in the validation set. A test set of the remaining 17 images was created.

4.1.2. *Potsdam dataset*

This dataset consists 38 images of the same size. Every image is an aerial true orthoimage with four-band that corresponds to the near-infrared (IR), red (R), green (G), and blue (B) bands. The dataset also provided DSM and nDSMs. The GSD of both was 5 cm for all patches. We selected 16 images as the training set and 8 images as the validation set. The remaining 14 images were used as a test set.

It is worth noting that only the IRRG and nDSM images were utilized for training in the Vaihingen and Potsdam datasets. For evaluation of the test dataset, all the training and validation data are used as the training set. Table 1 summarizes the detailed information about the Vaihingen dataset and Potsdam datasets, and Figure 4 shows the number of pixels in each class in both datasets. It can be seen that the number of pixels of 'car' and 'clutter' is quite small compared to the other classes both in Vaihingen and Potsdam datasets, which leads to sampling imbalances between the categories, making it difficult to identify them correctly.

4.2. *Evaluation metrics*

To validate the performance of different methods for semantic segmentation, four indicators were used, including per-class F1-score, mean F1-score (mF1), overall accuracy,

Table 1. Detailed information on the ISPRS 2D semantic labelling challenge dataset.

	Vaihingen	Potsdam
Total images	33	38
Image size	around 2500×2000	6000×6000
GSD	9cm	5cm
Bands	IR, R, G, nDSM	IR, R, G, B, nDSM
Training images	1, 3, 5, 7, 13, 17, 21, 23, 26, 32, 37	2_10, 2_12, 3_10, 3_11, 4_11, 4_12, 5_10, 5_12, 6_8, 6_9, 6_10, 6_11, 7_7, 7_9, 7_11, 7_12
Validation images	11, 15, 28, 30, 34	2_11, 3_12, 4_10, 5_11, 6_7, 6_12, 7_8, 7_10

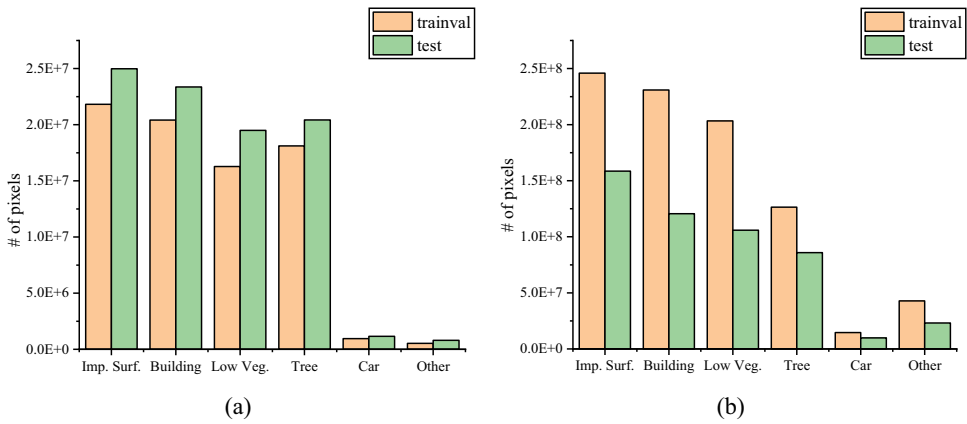


Figure 4. (a). Class distribution in the Vaihingen dataset. (b). Class distribution in the Potsdam dataset.

and mean intersection over union (mIoU). For height estimation, we used MAE and RMSE as the criteria.

$$\text{precision} = \frac{TP}{TP + FP} \quad \text{recall} = \frac{TP}{TP + FN} \quad (4)$$

$$F_1 = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (5)$$

These values can be calculated by a pixel-based accumulated confusion matrix, which is simply the sum of all the individual confusion matrices. The overall accuracy is derived by the normalization of the trace from the accumulated confusion matrix. The IoU is a well-known metric for determining how similar two sets are. The IoU is defined as the intersection size divided by the union size of two sets. By averaging the per-class IoU, the mIoU may be calculated.

$$\text{IoU} = \frac{|R_g \cap R_p|}{|R_g \cup R_p|} = \frac{|R_g \cap R_p|}{|R_g| + |R_p| - |R_g \cap R_p|} \quad (6)$$

where R_g and R_p indicate the label and prediction pixels, respectively, and $|\cdot|$ denotes the number of pixels in the set.

4.3. Implementation details

The following data augmentation approach was employed to alleviate the over-fitting problem. First, we used a random set of brightness, contrast, and saturation disturbances. The training images were then randomly sampled by flipping them horizontally or vertically and rotating them randomly. Finally, the sample patches were randomly cropped to 512×512 pixels. Besides, the dropout also was used to alleviate this problem. We employed stochastic gradient descent with a batch size of 8. The weight decay and momentum were set to 0.0001 and 0.9, respectively, and the adjustment strategy for the learning rate was the poly method. The learning rate was set at 0.01 and the models were

trained for 80,000 iterations in total. Due to the limitations of GPU memory, RS imagery is generally known to be too huge to send into a network. Thus, the test images were cropped to 512×512 patches with 20% overlap in our experiments. To achieve better performance, we used multi-scale inference, including horizontal flip, vertical flip, and scaling. Then, we averaged the output of each prediction to generate the final output. To evaluate our method more accurately, each experiment was conducted three times to calculate the mean values.

For the baseline, we used the VGG-16 as the backbone, which was pre-trained on the ImageNet dataset. Furthermore, the dilated convolution was used to make the output stride 8. Finally, a convolutional layer was added at the end of the network for classification.

4.4. The experiments on validation set

The validation set was used to conduct ablation studies to assess the effectiveness of the proposed network.

The segmentation results of the ablation study for the CASC module on the Vaihingen validation set were presented in Table 2. In this experiment, we performed four simplified versions of CASC modules. The first module operated on the channel dimension of the encoder features through a 1×1 convolution, which increased the accuracy by 2.14 mIoU. The second module contains three parallel 3×3 convolutions in addition to a 1×1 convolution, and the mIoU reaches 81.16. The third and fourth modules add pooling operations with different kernel sizes and convolutions with different dilation rates based on the second module, respectively. We observed that adding pooling or dilated convolutions further improved the accuracy. If our proposed CASC module was used, the final accuracy could reach 81.61 mIoU, which was 2.89 higher than the baseline.

As shown in Table 3, adding the ASPP module with VGG as the backbone brought about 2.13 and 2.94 improvements in mF1 and mIoU, respectively. This verified that the pyramid module can extract multi-scale features to enhance context information. As we can see, the CASC module also can capture context information to decrease the mismatch between the encoder and the decoder. In addition, the module also can restore edge information accurately using low-level features, which leads to 89.61 in mF1 and 81.57 in mIoU. When combining ASPP and CASC at the same time, the performance can be further improved, with mF1 up to 89.77 and mIoU up to 81.80. In these experiments, we show that the performance can be improved by making use of context information.

Then, we added a branch on the head of the decoder network to predict height values to conduct joint training, which further boosts the performance of segmentation. In detail, combining HEB and ASPP yielded a result of 81.78 in mIoU, which brought 0.24

Table 2. Segmentation results of the ablation study for the CASC module on the Vaihingen validation set.

Method	dilation rates	pool kernel size	Surf.	Building	Veg.	Tree	Car	mIoU
baseline			90.88	94.88	79.36	87.86	84.28	78.72
	-	-	91.33	95.40	79.56	88.54	91.05	80.86
	1,1,1	-	91.44	95.55	79.82	88.49	91.48	81.16
	1,1,1	3,5,7	91.57	95.50	80.00	88.62	91.49	81.28
	4,8,12	-	91.97	95.58	79.78	88.56	91.42	81.34
CASC	4,8,12	3,5,7	92.12	95.92	80.18	88.75	91.19	81.61

Table 3. Segmentation results on the Vaihingen validation set. **ASPP**: atrous spatial pyramid pooling, **CASC**: context-aggregation skip connection, **HEB**: height estimation branch.

Method	Surf.	Building	Veg.	Tree	Car	mF1	Acc	mIoU
ADL (Paisitkriangkrai et al. 2015)	89.10	94.30	77.36	86.25	71.91	83.78	86.89	-
RotEqNet (Marcos et al. 2018)	89.50	94.80	77.50	86.50	72.60	84.18	87.50	-
CNN-FPL (Volpi and Tuia 2017)	-	-	-	-	-	83.58	87.83	-
DST 2 (Sherrah 2016)	90.41	94.73	78.25	87.25	75.57	85.24	87.90	-
ERN (Liu et al. 2018a)	91.48	95.11	79.42	88.18	89.00	88.64	88.88	-
MLP (Maggiori et al. 2017b)	91.69	95.24	79.44	88.12	78.42	86.58	88.92	-
HCANet (Bai et al. 2021)	92.20	95.55	80.66	88.92	87.36	88.94	89.71	-
Baseline VGG16	90.96	94.81	79.48	88.04	85.64	87.78±0.06	88.68±0.07	78.60±0.10
VGG16+ASPP	91.85	95.76	80.62	88.78	91.03	89.61±0.10	89.61±0.13	81.54±0.17
VGG16+CASC	91.91	95.73	80.10	88.72	91.58	89.61±0.08	89.54±0.12	81.57±0.15
VGG16+ASPP+CASC	92.09	95.86	80.81	89.01	91.07	89.77±0.03	89.81±0.06	81.80±0.04
VGG16+ASPP+HEB	92.14	95.98	80.86	88.89	90.91	89.76±0.10	89.83±0.05	81.78±0.15
VGG16+CASC+HEB	91.82	95.72	80.64	88.90	91.73	89.76±0.05	89.65±0.05	81.79±0.08
VGG16+ASPP+CASC+HEB	92.14	95.92	80.75	89.00	91.34	89.83±0.04	89.84±0.09	81.91±0.07
Ours ResNet 50	92.57	96.39	80.76	88.75	91.76	90.05±0.06	90.00±0.11	82.30±0.10
Ours ResNet 101	92.66	96.46	81.25	88.87	92.04	90.26±0.08	90.17±0.09	82.63±0.16

improvement. Similarly, combined CASCs and HEB can reach 81.79 mIoU. Further, combining ASPP, CASC and HEB simultaneously can reach 81.91 mIoU and 89.83 mF1. This confirmed joint training improved prediction accuracy for the semantic segmentation task by using a shared representation to exploit commonalities across the tasks.

We also employed the ResNet-50 and ResNet-101 as the backbone. As can be seen, our proposed network based on ResNet-50 and ResNet-101 performed better than the network based on the VGG-16 network and had 0.39 and 0.72 improvements in mIoU, respectively. It is noted that the F1 score of the car class achieved 92.04 with a large margin compared to the network based on VGG16.

Beyond that, we also compared our method with other state-of-the-art approaches on the Vaihingen validation dataset for a comprehensive evaluation, including ADL, RotEqNet, CNN-FPL, DST 2, ERN, MLP, and HCANet. The numerical results are shown in Table 3, which shows a considerable improvement between other methods and ours. It is demonstrated that our network outperformed the other methods in terms of the mF1 and the overall accuracy, which indicated the effectiveness of our method. Besides, our method also remarkably surpassed other methods for the class of the car.

In addition, the results on the Potsdam validation dataset are included in Table 4. The mF1, overall accuracy, and mIoU were improved by 1.45, 1.62, and 2.42 respectively after

Table 4. Segmentation results on the Potsdam validation set. **ASPP**: atrous spatial pyramid pooling, **CASC**: context-aggregation skip connection, **HEB**: height estimation branch.

Method	Surf.	Building	Veg.	Tree	Car	Clutter	mF1	Acc	mIoU
Baseline VGG16	90.56	94.07	83.51	85.30	94.22	70.25	89.53±0.13	87.97±0.10	81.34±0.21
VGG16+ASPP	92.26	95.42	85.32	86.16	95.74	75.41	90.98±0.07	89.59±0.10	83.76±0.10
VGG16+CASC	91.99	95.32	85.22	85.99	95.87	74.47	90.87±0.03	89.43±0.03	83.59±0.04
VGG16+ASPP+CASC	92.30	95.56	85.38	86.07	95.88	77.16	91.04±0.10	89.73±0.11	83.86±0.17
VGG16+ASPP+HEB	92.27	95.67	85.34	85.89	95.95	77.02	91.03±0.07	89.70±0.14	83.85±0.11
VGG16+CASC+HEB	92.13	95.48	85.28	86.13	95.81	76.09	90.97±0.02	89.62±0.09	83.74±0.04
VGG16+ASPP+CASC+HEB	92.28	95.57	85.60	86.13	95.97	77.52	91.13±0.06	89.84±0.06	84.02±0.09
Ours ResNet 50	92.89	95.88	85.37	86.34	96.60	76.66	91.41±0.04	90.00±0.09	84.54±0.07
Ours ResNet 101	92.57	96.01	85.63	86.33	96.72	76.97	91.48±0.02	90.05±0.05	84.60±0.05

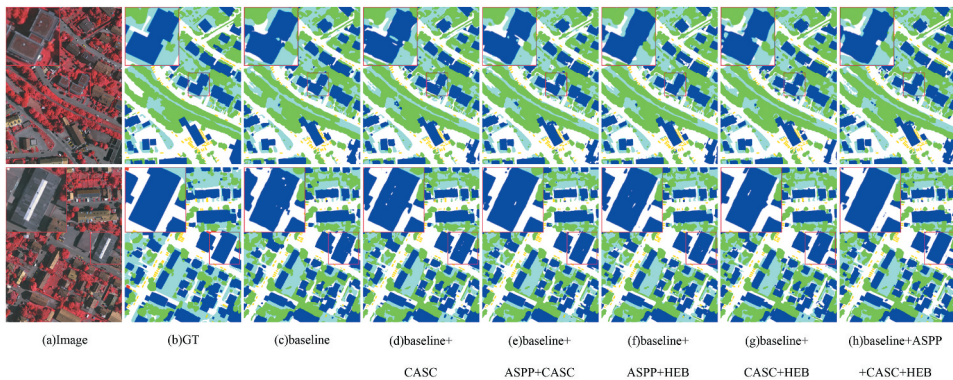


Figure 5. Qualitative comparison with different modules on ISPRS Vaihingen validation set.

adding the ASPP module. It shows an increase of 1.46 mIoU when adding the CASC module. Add combining ASPP and CASC also boosted performance, which can reach 91.04 mF1 and 83.86 mIoU. For the class of clutter, the mF1 was improved by 6.91 with a large margin compared with the baseline. This also shows that the performance can be improved by making use of context information in images.

When predicting the land-cover types and height values of pixels simultaneously, the accuracy can be promoted further. Similar to the results on the Vaihingen dataset, combining ASPP, CASC, and HEB modules at the same time, the highest accuracy can be achieved. That demonstrated that more robust features can be learned by using multi-task learning. We also tried to utilize ResNet-50 and ResNet-101 as the backbone and the result exhibited the best overall performance. Compared to the results on the Vaihingen dataset, we noted that the performance of the car class was higher. We argued that the images of the Potsdam dataset have higher GSD, thus cars occupy more pixels, which makes cars easier to classify.

Furthermore, the visualization results of different modules on the Vaihingen validation set are shown in Figure 5. Combining the CASC and HEB modules produces the most comprehensive segmentation for the entire building, further validating the effectiveness of our proposed module.

4.5. Comparison with state-of-the-arts on the Vaihingen and Potsdam dataset

In this section, the state-of-the-art approaches and our proposed method were compared on ISPRS Vaihingen and Potsdam datasets.

4.5.1. Vaihingen dataset

We evaluated our approach on the Vaihingen test set and the evaluation is shown in Table 5 alongside other state-of-the-art approaches. In general, our method achieved 91.4% overall accuracy, which exceeded most of the other methods. Our method performed well in all the given classes. More specifically, although the pixels of the car and clutter classes only accounted for a small part of the total image pixels, our results achieved better performance, resulting in 90.4 and 63.5 on the F1 score. Some methods

Table 5. Segmentation results on the Vaihingen test set.

Method	Surf.	Building	Veg.	Tree	Car	Clutter	mF1	Acc
UZ_1 (Volpi and Tuia 2017)	89.2	92.5	81.6	86.9	57.3	4.6	81.5	87.3
ADL_3 (Paisitkriangkrai et al. 2015)	89.5	93.2	82.3	88.2	63.3	-	83.3	88.0
CVEO (Chen et al. 2018a)	90.5	92.4	81.7	88.5	79.4	38.3	86.5	88.3
Ano2 (Zhang et al. 2017)	90.4	93.0	81.4	88.6	74.5	47.5	85.6	88.4
DLR_2 (Marmanis et al. 2016)	90.3	92.3	82.5	89.5	76.3	-	86.2	88.5
VNU4 (Bui et al. 2018)	91.2	93.6	81.5	88.2	77.7	45.1	86.4	89.0
DST_2 (Sherrah 2016)	90.5	93.7	83.4	89.2	72.6	-	85.9	89.1
UFMG_5 (Nogueira et al. 2019)	91.0	94.6	82.7	88.9	82.5	-	87.9	89.3
ONE_7 (Audebert, Saux, and Lefèvre 2016)	91.0	94.5	84.4	89.9	77.8	-	87.5	89.8
RIT_7 (Piramanayagam et al. 2018)	91.7	95.2	83.5	89.2	82.8	-	88.5	89.9
V-FuseNet (Audebert, Le Saux, and Lefèvre 2018)	91.0	94.4	84.5	89.9	86.3	-	89.2	90.0
DLR_9 (Marmanis et al. 2018)	92.4	95.2	83.9	89.9	81.2	-	88.5	90.3
GSN3 (Wang et al. 2017)	92.2	95.1	83.7	89.9	82.4	48.7	88.7	90.3
NLPR2 (Sun et al. 2017)	92.6	95.3	84.7	90.0	81.0	53.2	88.7	90.7
CASIA2 (Liu et al. 2018b)	93.2	96.0	84.7	89.9	86.7	50.4	90.1	91.1
DDCM-Net (Liu et al. 2020)	92.7	95.3	83.3	89.4	88.3	-	89.8	90.4
EaNet (Zheng et al. 2020)	93.4	96.2	85.6	90.5	88.3	-	90.8	91.2
CCANet (Deng et al. 2021)	93.3	94.3	82.0	88.6	86.6	-	89.0	91.1
BANet (Wang et al. 2021)	92.2	95.2	83.8	89.9	86.8	-	89.6	90.5
HCANet (Bai et al. 2021)	92.5	95.0	84.2	89.4	84.0	-	89.0	90.3
MACANet (Li, Lei, and Kuang 2021)	88.4	91.6	77.8	85.6	78.5	-	84.4	-
HECR-Net (Liu et al. 2021)	93.6	95.5	85.8	90.4	89.1	-	90.9	91.5
JSH-Net	93.3	96.3	85.0	90.0	90.4	63.5	91.0±0.11	91.4±0.03

used post-processing such as CRF to refine their prediction (Marmanis et al. 2016; Paisitkriangkrai et al. 2015; Sherrah 2016). However, our method did not use any post-processing. In addition, some works combined both image and elevation data as input to offer more information to get better performance (Audebert, Le Saux, and Lefèvre 2018; Audebert, Saux, and Lefèvre 2016; Marmanis et al. 2018, 2016; Nogueira et al. 2019; Piramanayagam et al. 2018; Volpi and Tuia 2017). In our case, we used nDSM as the output to estimate height values, which made the network learn more robust features and leads to stronger regularization. Although HCANet and MACANet propose a skip connection module, our network can achieve higher segmentation accuracy. In addition, HECR-Net proposed the same idea as ours (using multi-task learning to predict segmentation maps and relative height maps simultaneously). However, our method is more accurate in the building and vehicle classes. BANet uses a new lightweight Transformer backbone. However, our method achieves higher performance than theirs.

A confusion matrix is shown in Table 6. It is easy to see that the class most likely to be misclassified is the clutter (background) class. Because the number of clutter category samples was very small and contained a wide variety of object categories, it was difficult

Table 6. Confusion matrix on the Vaihingen test set.

reference \ predicted	Surf.	Building	Veg.	Tree	Car	Clutter
Surf.	0.937	0.022	0.043	0.008	0.057	0.012
Building	0.024	0.966	0.011	0.001	0.007	0.007
Veg.	0.030	0.009	0.843	0.100	0.003	0.004
Tree	0.008	0.001	0.083	0.910	0.001	0.003
Car	0.003	0.001	0.000	0.000	0.878	0.002
Clutter	0.009	0.007	0.000	0.000	0.008	0.464
Precision	0.927	0.960	0.860	0.893	0.921	0.942
Recall	0.937	0.966	0.843	0.910	0.878	0.464
F1	0.932	0.963	0.851	0.901	0.899	0.622

for the networks to distinguish this class. The other two confusing categories were low vegetation and tree. We think it is difficult to distinguish these two classes simply by relying on the texture information in the image. To improve recognition accuracy, more extra information is needed.

Figure 6 displays some of the qualitative results on large patches of the Vaihingen test set. The results of other methods were obtained from the website. As can be seen, our method produced more accurate and finer structures compared to the other methods. For big buildings (first, third, and fourth images), the other methods' results often produced incomplete segmentation results. These manmade structures were irregular

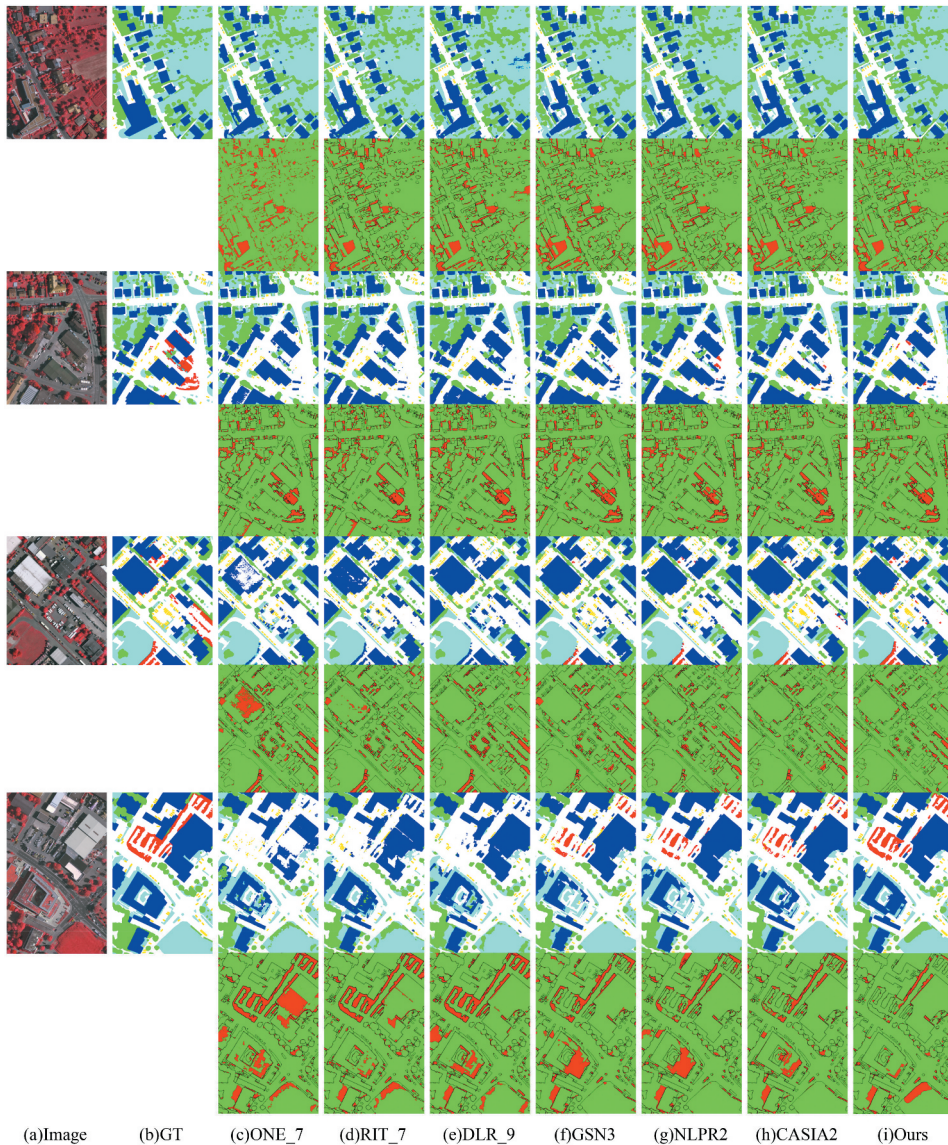


Figure 6. Qualitative comparison with other competitors' methods on small patches of ISPRS Vaihingen challenge TEST SET.

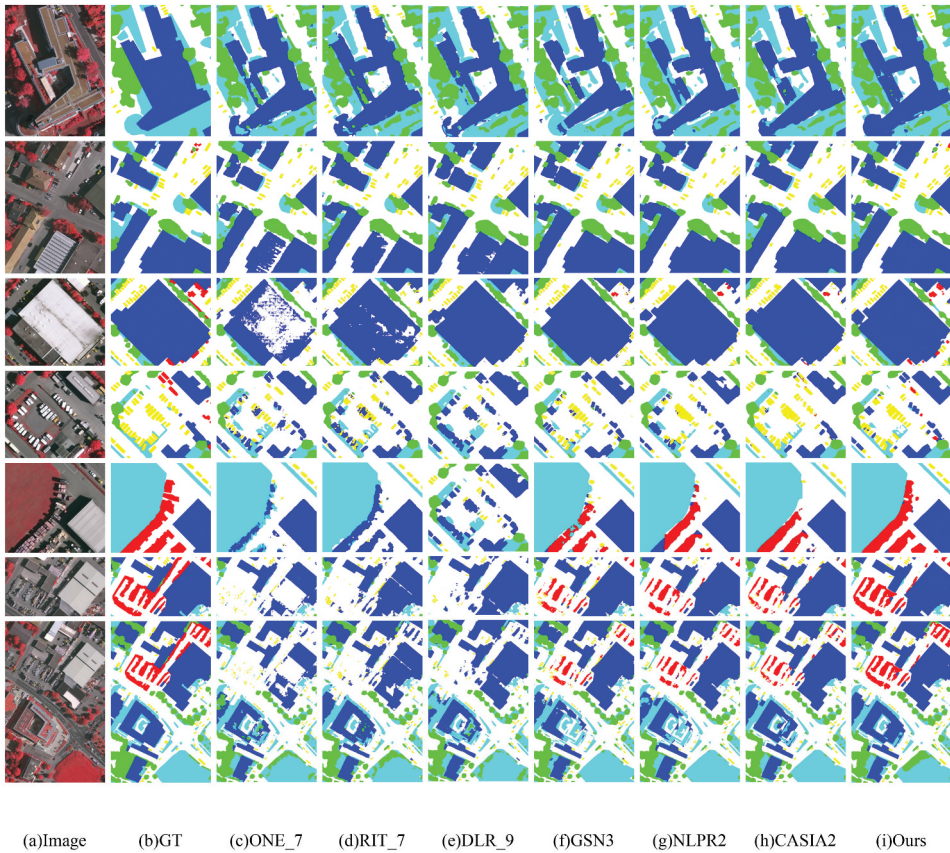


Figure 7. Qualitative comparison with other competitors' methods on small patches of ISPRS Vaihingen challenge TEST SET.

in shape, the roof material was very different from the other buildings and some buildings had vegetation on the roof, which made these buildings confusing, see the first, second, third, and seventh small patches in Figure 7, especially the seventh image. For fine-structured objects, such as cars and clutter, the other methods tended to obtain inaccurate recognition. For example, there are many cars in the parking lot in the fourth image. The other methods tended to identify cars as buildings. However, our method achieved coherent labelling for fine-structured objects. In addition, as seen in the fifth and sixth images, our approach for the clutter class was more robust for the intricate scenarios.

4.5.2. Potsdam dataset

Similarly, we showed the evaluation predictions on the Potsdam test set in Table 7 and compared with the other state-of-the-art approaches, from which a conclusion was drawn similar to that for the Vaihingen dataset. Our method achieved 92.0% overall accuracy and 93.9 mF1. For the six classes, our results also demonstrated the best performance. Similarly, the classes of impervious surfaces, building, car, and clutter achieved better performance compared to the other methods. Moreover, we observed a relatively small increase in the Potsdam dataset in the vehicle and background categories compared to

Table 7. Segmentation results on the Potsdam test set.

Method	Surf.	Building	Veg.	Tree	Car	Clutter	mF1	Acc
UZ_1 (Volpi and Tuia 2017)	89.3	95.4	81.8	80.5	86.5	29.3	86.7	85.8
UFMG_4 (Nogueira et al. 2019)	90.8	95.6	84.4	84.3	92.4	49.5	89.5	87.9
RIT_L7 (Liu et al. 2017b)	91.2	94.6	85.1	85.1	92.8	46.8	89.8	88.4
CVEO (Chen et al. 2018a)	91.2	94.5	86.4	87.4	95.4	40.2	91.0	89.0
DST_5 (Sherrah 2016)	92.5	96.4	86.7	88.0	94.7	56.2	91.7	90.3
RIT4 (Piramanayagam et al. 2018)	92.6	97.0	86.9	87.4	95.2	54.4	91.8	90.3
V-FuseNet (Audebert, Le Saux, and Lefèvre 2018)	92.7	96.3	87.3	88.5	95.4	-	92.0	90.6
CASIA2 (Liu et al. 2018b)	93.3	97.0	87.7	88.4	96.2	59.1	92.5	91.1
DDCM-Net (Liu et al. 2020)	92.9	96.9	87.7	89.4	94.9	-	92.4	90.8
BANet (Wang et al. 2021)	93.3	96.7	87.4	89.1	96.0	-	92.5	91.1
HCANet (Bai et al. 2021)	93.1	96.6	87.0	88.5	96.1	61.2	92.3	90.8
MACANet (Li, Lei, and Kuang 2021)	90.6	94.3	83.5	84.0	90.0	-	88.5	-
HECR-Net (Liu et al. 2021)	93.8	97.4	88.7	89.2	95.4	-	92.9	91.8
JSH-Net	94.3	97.7	88.5	89.1	97.2	63.2	93.9±0.05	92.0±0.06

Table 8. Confusion matrix on the Potsdam test set.

reference \ predicted	Surf.	Building	Veg.	Tree	Car	Clutter
Surf.	0.952	0.009	0.036	0.022	0.003	0.042
Building	0.007	0.982	0.003	0.004	0.000	0.021
Veg.	0.017	0.003	0.910	0.061	0.000	0.049
Tree	0.010	0.002	0.081	0.878	0.012	0.010
Car	0.000	0.000	0.000	0.001	0.975	0.002
Clutter	0.031	0.017	0.025	0.005	0.018	0.521
Precision	0.935	0.970	0.862	0.904	0.966	0.808
Recall	0.952	0.982	0.910	0.878	0.975	0.521
F1	0.943	0.976	0.885	0.891	0.971	0.633

the Vaihingen dataset. We suspect this was attributable to the lower intra-class variance of the Potsdam dataset. We also show a confusion matrix in Table 8. The two most confusing categories were low vegetation and tree, which is also consistent with the Vaihingen dataset. We argue that the low vegetation and tree classes have a smaller inter-class variance, which makes their recognition accuracy relatively low.

We also visualized some of the results from the large whole patches and small patches on the test set of the Potsdam dataset in Figures 8 and 9. Compared to the other methods, our results contained more accurate detailed information. For buildings, our approach was more robust to building diversity. For example, the other methods did not accurately identify the buildings in the second image of Figure 9, and there were flaws in their results. Although the results of our models also had a few flaws, our method achieved a relatively more consistent segmentation result and precise edges. For fine-grain objects, our results achieved clear and precise results. In the fifth image of Figure 9, it is noted that a grid pattern exists in the lower vegetation in the results of DST5 and CASIA2 while our results did not. However, our results in some of the complicated scenes struggled with the clutter class, such as the first and fifth images in Figure 8, which we believe that more contextual information or more training data would improve the results.

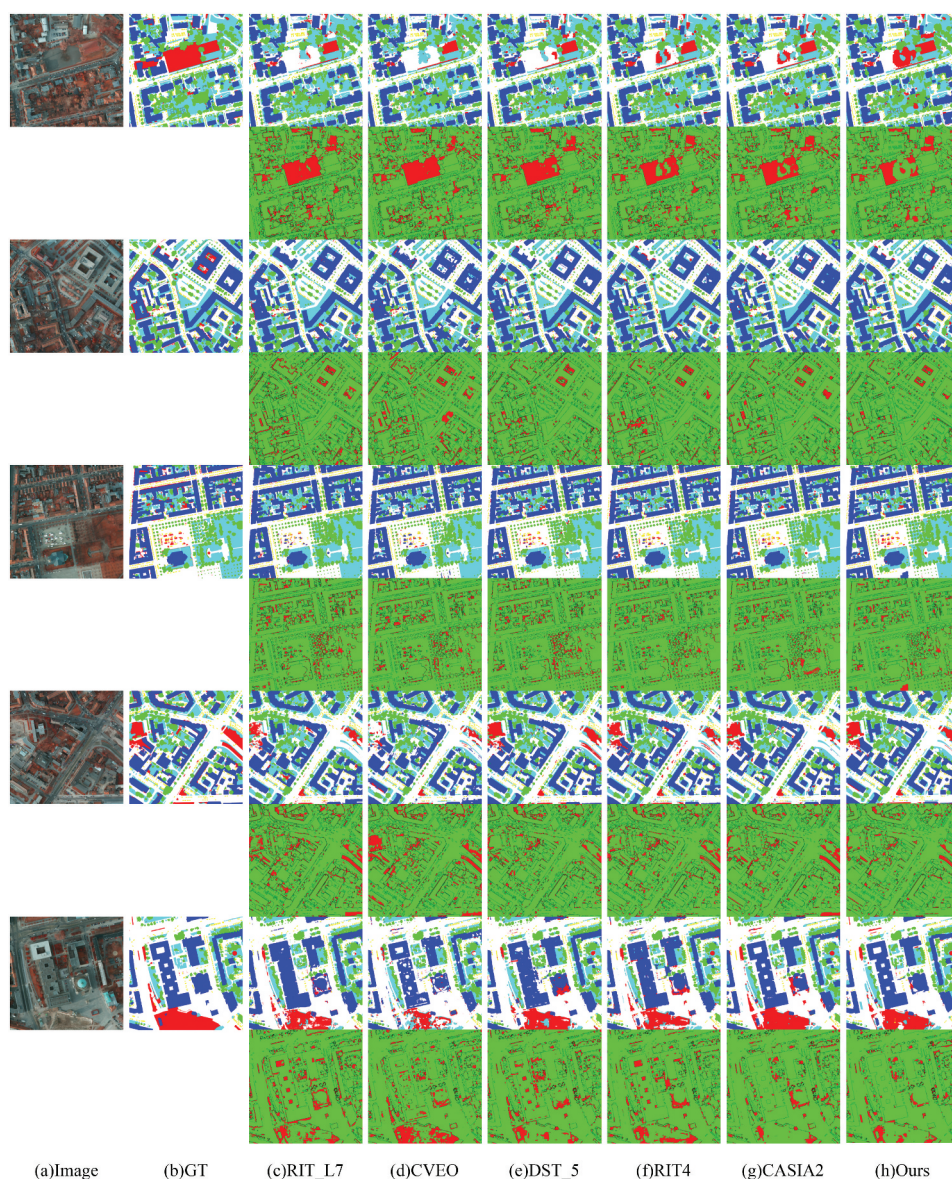


Figure 8. Qualitative comparison with other competitors' methods on large tiles of ISPRS Potsdam challenge TEST SET.

4.6. The results of height estimation on Vaihingen and Potsdam dataset

The quantitative results of height estimation were reported in Table 9 and the qualitative results were shown in Figure 10. Because the benchmark did not provide height estimation results, we only compared our results with the ground truth. Our model reached 0.037 and 0.067 for MAE and RMSE on the Vaihingen dataset, respectively. For a simple scene, height estimation can get good performance. We believe that semantic segmentation and height estimation are complementary and consistent

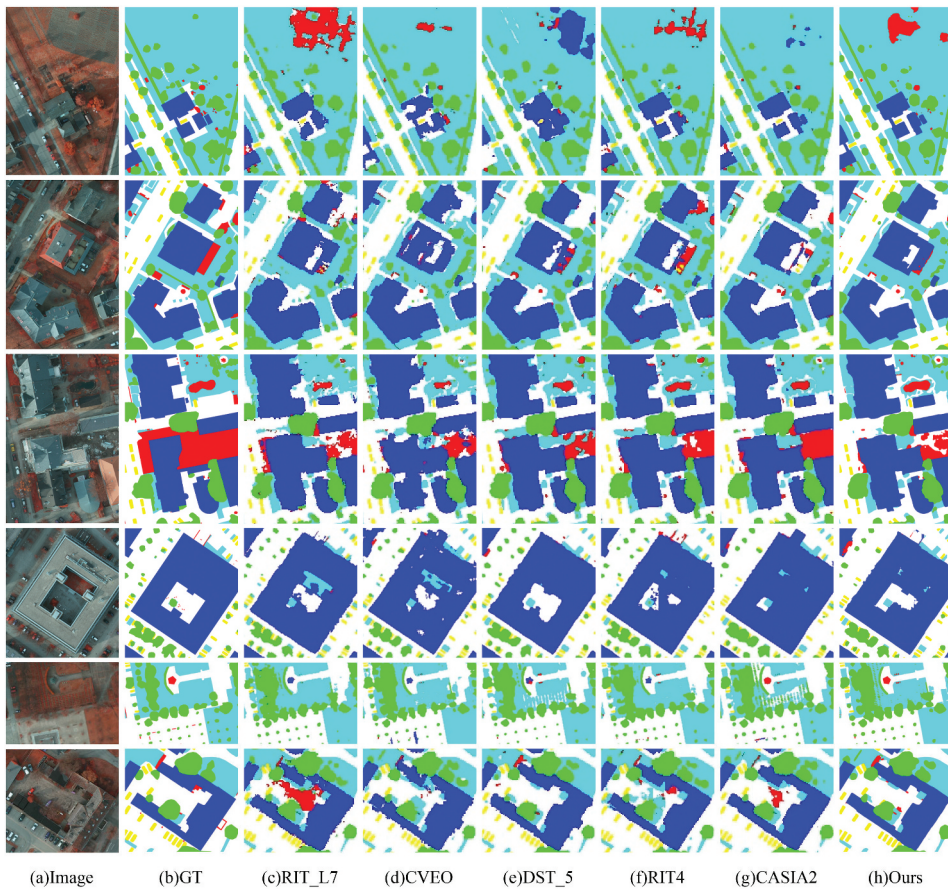


Figure 9. Qualitative comparison with other competitors' methods on small patches of ISPRS Vaihingen challenge TEST SET.

Table 9. The results of height estimation on the Vaihingen and Potsdam test set.

dataset	MAE	RMSE
Vaihingen	0.037	0.067
Potsdam	0.047	0.098

and predicting only the semantic class can cause model over-fitting. However, multi-task learning can improve the generalization capability of the model due to shared features, which can be thought of as a kind of regularization. For example, in the first and fourth images of Figure 6, because there is the vegetation on the roof of the big buildings, the other methods identified the buildings as low vegetation. But in our case, the big buildings were recognized correctly, which means that height estimation can conduce to semantic segmentation. The same situation also occurred in the Potsdam dataset. In addition, the generated nDSM contained some flaws because of the restriction of the algorithm. For example, in the area of the big building, the

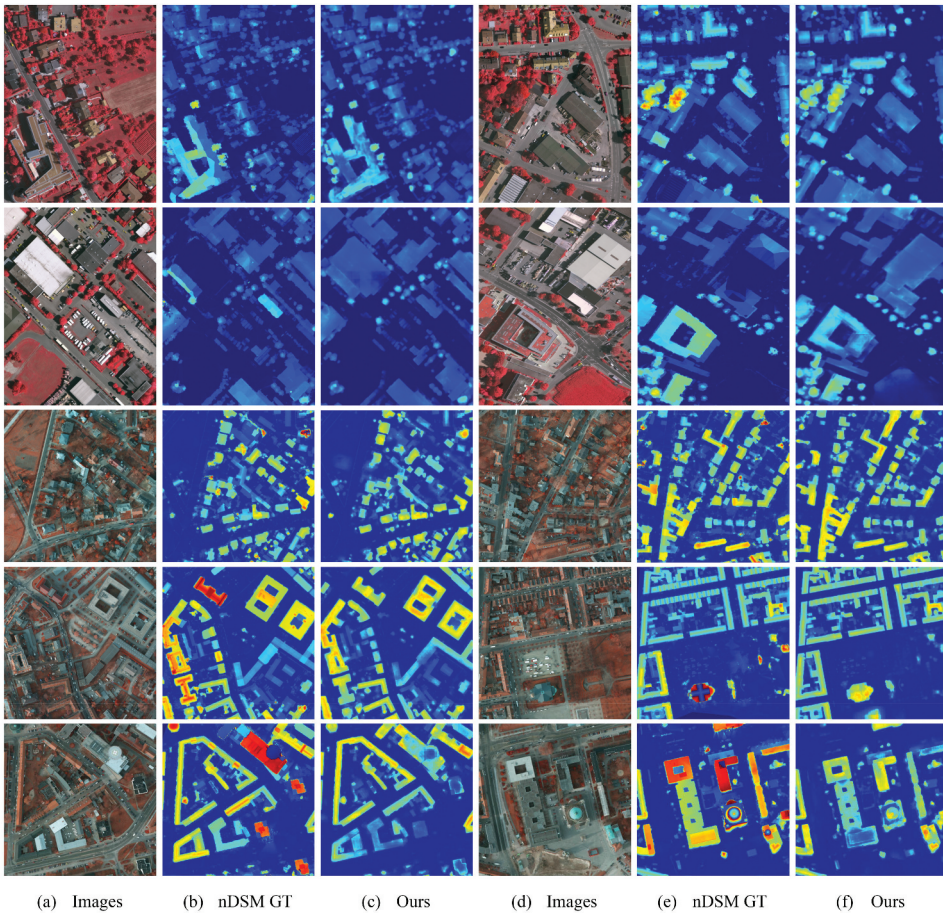


Figure 10. Qualitative comparison for results of height estimation.

elevation in ground truth is missing in the second row of Figure 10, which will lead to the wrong supervised signal to train the model.

4.7. The analysis of the proposed network

4.7.1. Computational cost analysis

To evaluate the computational cost of our method and other state-of-the-art methods, we compared the number of parameters, computational cost (FLOPs), inference time, and mIoU on the Vaihingen validation set. The cost and inference time were calculated when the input image size was 512×512 and the GPU was NVIDIA GTX 3090. As shown in Table 10, when using VGG16 as the backbone, our method has higher mIoU than FCN and UNet, but its parameters and computational cost are relatively high. However, it is worthwhile that the proposed method leverages multi-task learning to improve semantic segmentation results. When using ResNet50 as the backbone, the proposed method still achieves higher mIoU than PSPNet, DeeplabV3, and DANet. Moreover, compared to

Table 10. Comparison of computational cost with state-of-the-art methods.

Models	Backbones	Params (Million)	FLOPs (Giga)	Inference time (ms)	mIoU
FCN (Long, Shelhamer, and Darrell 2015)	VGG16	16.2	80.73	10.67	74.29
UNet (Ronneberger, Fischer, and Brox 2015)	VGG16	24.6	113.69	14.33	81.61
PSPNet (Zhao et al. 2017)	ResNet50	46.6	178.45	30.07	82.04
DeepLab V3 (Chen et al. 2017b)	ResNet50	65.74	269.65	35.65	81.76
DANet (Fu et al. 2019)	ResNet50	47.46	199.07	31.89	81.86
JSH-Net	VGG16	26.95	152.58	18.29	81.97
JSH-Net	ResNet50	51.52	213.59	35.37	82.22

DeepLabV3, the proposed method requires a small number of parameters, less computational cost, and less inference time.

4.7.2. Comparison with the state-of-the-art methods

Compared with the methods using only image information networks, our approach differs from others in that we take into account the mismatch between encoder and decoder features and propose a new skip connection module that aggregates contextual information, alleviating the divide in its corresponding features and improving the ability of the network to segment complex features. Compared with other methods that directly input elevation information as auxiliary data into the network, the main advantage of our approach is that our network does not narrow the application of the network. Inputting elevation information as auxiliary data into the network requires that elevation information must be available for use in tests or real-world situations. In contrast, our network is tested under the same conditions as other networks that use image information.

Some networks share our approach of using height-assisted data as output, such as Srivastava, Volpi, and Tuia (2017). However, our approach differs from others in that our method improves the network's ability to handle complex scenes by introducing a skip-join module that aggregates contextual information and validates multi-task learning effectiveness. Compared with HECR-Net (Liu et al. 2021) and CI-Net (Gao et al. 2021), which explore the use of predicted height maps to improve semantic segmentation results further, CI-Net designs a scene understanding module and a feature complementation module based on a self-attentive mechanism. Our approach is similar in that both use a single image to predict the semantic map and the depth map (or height map) via a multi-task network. Our approach differs because our network aggregates features at different scales to train discriminative features. Our approach is complementary to both methods.

4.7.3. Limitations of the proposed method

The above experiments show that the proposed method improves the accuracy of semantic segmentation, especially for buildings and cars. However, the proposed method still suffers from some limitations as follows:

- (1) We observe that the segmentation results are not good enough when dealing with some complex buildings or buildings affected by shadows (see the first four rows of Figure 11). A possible solution to this problem is to fuse images and elevation data to further improve building segmentation accuracy.

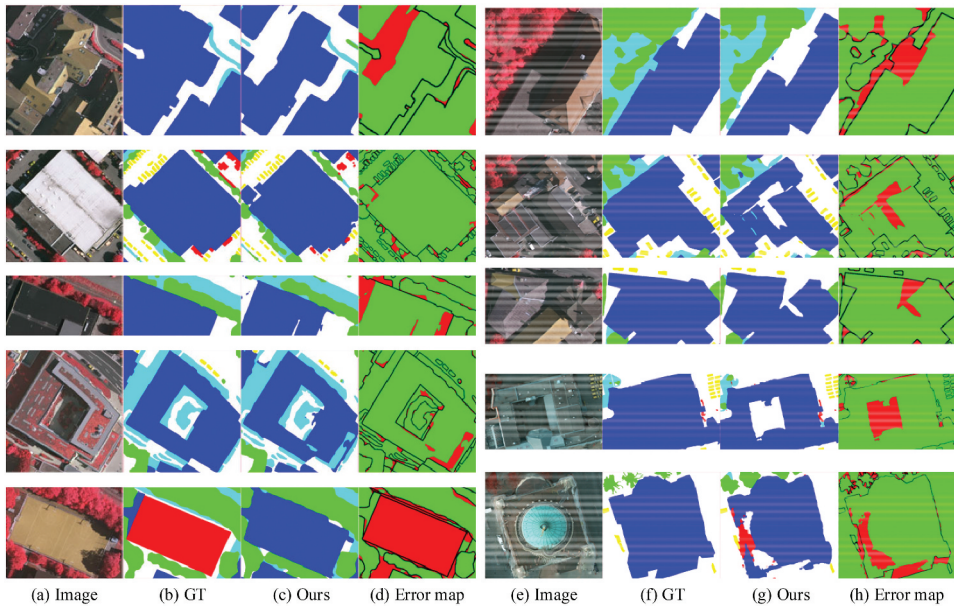


Figure 11. Segmentation results of complex objects.

- (2) The correctness of segmentation is greatly reduced when encountering samples with different distributions or unseen categories (last row of Figure 11). In the future, this can be mitigated by domain adaptation methods using a large number of unlabelled samples.
- (3) The proposed network requires a large number of high-precision labelled samples. In practical applications, the available annotated samples may be very limited. A way to overcome this problem is to incorporate prior knowledge in the field of remote sensing into deep networks, such as using remote sensing knowledge graphs to assist semantic segmentation, and another way is to use a large number of unlabelled samples to improve the accuracy of the network through semi-supervised learning.

5. Conclusion

We proposed a multi-task learning convolutional network for semantic segmentation and height estimation in this research. In detail, a dilated pyramid skip connection module was proposed, which can aggregate context information from the encoder part and alleviate the semantic mismatch between the encoder network and decoder network. Its main objective is to increase the ability of the network to recognize objects of different scales and sizes. The experimental results on two ISPRS datasets show that using the CASC module can increase the recognition ability of complex objects (such as buildings with shadows) and small objects (cars). Furthermore, the progressive refinement strategy was proposed to recover detailed information about objects. Moreover, we also proposed a height estimation branch on the head of the model to utilize shared features and exploit their potential similarity across the tasks, resulting in robust features and higher

prediction accuracy. Our experiments verified that our approach can achieve state-of-the-art results on the ISPRS benchmarks by multi-task learning, which demonstrated that our method is effective for high-quality semantic segmentation and height estimation. In the future, we intend to further address the limitations in the proposed network and explore the application of multi-task learning in RS imagery.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 42030102, Grant 42192583, Grant 42001406, and Grant 62102268; in part by the Fund for Innovative Research Groups of the Hubei Natural Science Foundation under Grant 2020CFA003; and in part by the China Postdoctoral Science Foundation under Grant 2020M672416; and in part by the Major special projects of Guizhou[2022]001.

ORCID

Bin Zhang  <http://orcid.org/0000-0001-9545-2760>

References

- Amirkolaee, H. A., and H. Arefi. 2019. "Height Estimation from Single Aerial Images Using a Deep Convolutional Encoder-Decoder Network." *ISPRS Journal of Photogrammetry and Remote Sensing* 149: 50–66. doi:10.1016/j.isprsjprs.2019.01.013.
- Audebert, N., B. Le Saux, and S. Lefèvre. 2018. "Beyond RGB: Very High Resolution Urban Remote Sensing with Multimodal Deep Networks." *ISPRS Journal of Photogrammetry and Remote Sensing* 140: 20–32. doi:10.1016/j.isprsjprs.2017.11.011.
- Audebert, N., B. L. Saux, and S. Lefèvre, 2016. "Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-Scale Deep Networks." Asian conference on computer vision, Taipei, Taiwan. Springer, pp. 180–196.
- Badrinarayanan, V., A. Kendall, and R. Cipolla. 2017. "Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (12): 2481–2495. doi:10.1109/TPAMI.2016.2644615.
- Bai, H., J. Cheng, X. Huang, S. Liu, and C. Deng. 2021. "HCANet: A Hierarchical Context Aggregation Network for Semantic Segmentation of High-Resolution Remote Sensing Images." *IEEE Geoscience and Remote Sensing Letters* 19: 1–5. doi:10.1109/LGRS.2021.3063799.
- Ball, J. E., D. T. Anderson, and C. S. Chan Sr. 2017. "Comprehensive Survey of Deep Learning in Remote Sensing: Theories, Tools, and Challenges for the Community." *Journal of Applied Remote Sensing* 11 (04): 042609. doi:10.1117/1.JRS.11.042609.
- Bui, D.-T., T.-D. Tran, T.-T. Nguyen, Q.-L. Tran, and D.-V. Nguyen. 2018. "Aerial Image Semantic Segmentation Using Neural Search Network Architecture." International Conference on Multi-disciplinary Trends in Artificial Intelligence, Hanoi, Vietnam. Springer, pp. 113–124.
- Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2017a. "Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (4): 834–848. doi:10.1109/TPAMI.2017.2699184.

- Chen, L.-C., G. Papandreou, F. Schroff, and H. Adam. 2017b. "Rethinking Atrous Convolution for Semantic Image Segmentation." *arXiv preprint arXiv:1706.05587*.
- Chen, G., X. Zhang, Q. Wang, F. Dai, Y. Gong, and K. Zhu. 2018a. "Symmetrical Dense-Shortcut Deep Fully Convolutional Networks for Semantic Segmentation of Very-High-Resolution Remote Sensing Images." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11 (5): 1633–1644. doi:10.1109/JSTARS.2018.2810320.
- Chen, L.-C., Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. 2018b. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation." Proceedings of the European conference on computer vision (ECCV), Munich, Germany, pp. 801–818.
- Deng, G., Z. Wu, C. Wang, M. Xu, and Y. Zhong. 2021. "CCANet: Class-Constraint Coarse-To-Fine Attentional Deep Network for Subdecimeter Aerial Image Semantic Segmentation." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–20.
- Eigen, D., and R. Fergus. 2015. "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture." Proceedings of the IEEE international conference on computer vision, Santiago, Chile, pp. 2650–2658.
- Eigen, D., C. Puhrsch, and R. Fergus. 2014. "Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network." *Advances in Neural Information Processing Systems* 27: 2366–2374.
- Fu, J., J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. 2019. "Dual Attention Network for Scene Segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Long Beach, CA, USA, pp. 3146–3154.
- Gao, T., W. Wei, Z. Cai, Z. Fan, S. Xie, X. Wang, and Q. Yu. 2021. "CI-Net: Contextual Information for Joint Semantic Segmentation and Depth Estimation." *arXiv preprint arXiv:2107.13800*. doi:10.1007/s10489-022-03401-x.
- Ghamisi, P., and N. Yokoya. 2018. "IMG2DSM: Height Simulation from Single Imagery Using Conditional Generative Adversarial Net." *IEEE Geoscience and Remote Sensing Letters* 15 (5): 794–798. doi:10.1109/LGRS.2018.2806945.
- He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep Residual Learning for Image Recognition." Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, pp. 770–778.
- Li, X., L. Lei, and G. Kuang. 2021. "Multilevel Adaptive-Scale Context Aggregating Network for Semantic Segmentation in High-Resolution Remote Sensing Images." *IEEE Geoscience and Remote Sensing Letters* 19: 1–5. doi:10.1109/LGRS.2021.3091284.
- Lin, G., A. Milan, C. Shen, and I. Reid. 2017. "Refinenet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, pp. 1925–1934.
- Liu, S., W. Ding, C. Liu, Y. Liu, Y. Wang, and H. Li. 2018a. "ERN: Edge Loss Reinforced Semantic Segmentation Network for Remote Sensing Images." *Remote Sensing* 10 (9): 1339. doi:10.3390/rs10091339.
- Liu, Y., B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan. 2018b. "Semantic Labeling in Very High Resolution Images via a Self-Cascaded Convolutional Neural Network." *ISPRS Journal of Photogrammetry and Remote Sensing* 145: 78–95. doi:10.1016/j.isprsjprs.2017.12.007.
- Liu, Q., M. Kampffmeyer, R. Jenssen, and A.-B. Salberg. 2020. "Dense Dilated Convolutions' Merging Network for Land Cover Classification." *IEEE Transactions on Geoscience and Remote Sensing* 58 (9): 6309–6320. doi:10.1109/TGRS.2020.2976658.
- Liu, Y., D. Minh Nguyen, N. Deligiannis, W. Ding, and A. Munteanu. 2017a. "Hourglass-Shapenetwork Based Semantic Segmentation for High Resolution Aerial Imagery." *Remote Sensing* 9 (6): 522. doi:10.3390/rs9060522.
- Liu, Y., S. Piramanayagam, S. T. Monteiro, and E. Saber. 2017b. "Dense Semantic Labeling of Very-High-Resolution Aerial Imagery and Lidar with Fully-Convolutional Neural Networks and Higher-Order CRFs." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, pp. 76–85.
- Liu, M., M. Salzmann, and X. He. 2014. "Discrete-Continuous Depth Estimation from a Single Image." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, pp. 716–723.

- Liu, W., W. Zhang, X. Sun, Z. Guo, and K. Fu. 2021. "HECR-Net: Height-Embedding Context Reassembly Network for Semantic Segmentation in Aerial Images." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14: 9117–9131. doi:10.1109/JSTARS.2021.3109439.
- Long, J., E. Shelhamer, and T. Darrell. 2015. "Fully Convolutional Networks for Semantic Segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, pp. 3431–3440.
- Maggiori, E., Y. Tarabalka, G. Charpiat, and P. Alliez. 2017a. "Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark." *IEEE International Geoscience and Remote Sensing Symposium*, Fort Worth, TX, USA, pp. 3226–3229.
- Maggiori, E., Y. Tarabalka, G. Charpiat, and P. Alliez. 2017b. "High-Resolution Aerial Image Labeling with Convolutional Neural Networks." *IEEE Transactions on Geoscience and Remote Sensing* 55 (12): 7092–7103. doi:10.1109/TGRS.2017.2740362.
- Ma, L., Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson. 2019. "Deep Learning in Remote Sensing Applications: A Meta-Analysis and Review." *ISPRS Journal of Photogrammetry and Remote Sensing* 152: 166–177. doi:10.1016/j.isprsjprs.2019.04.015.
- Marcos, D., M. Volpi, B. Kellenberger, and D. Tuia. 2018. "Land Cover Mapping at Very High Resolution with Rotation Equivariant CNNs: Towards Small Yet Accurate Models." *ISPRS Journal of Photogrammetry and Remote Sensing* 145: 96–107. doi:10.1016/j.isprsjprs.2018.01.021.
- Marmanis, D., K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla. 2018. "Classification with an Edge: Improving Semantic Image Segmentation with Boundary Detection." *ISPRS Journal of Photogrammetry and Remote Sensing* 135: 158–172. doi:10.1016/j.isprsjprs.2017.11.009.
- Marmanis, D., J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla. 2016. "Semantic Segmentation of Aerial Images with an Ensemble of CNNs." *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 3: 473–480. doi:10.5194/isprs-annals-III-3-473-2016.
- Mnih, V. 2013. *Machine Learning for Aerial Image Labeling*. Canada: University of Toronto.
- Mou, L., and X. X. Zhu. 2018. "IM2HEIGHT: Height Estimation from Single Monocular Imagery via Fully Residual Convolutional-Deconvolutional Network." *arXiv preprint arXiv:180210249*.
- Nogueira, K., M. Dalla Mura, J. Chanussot, W. R. Schwartz, and J. A. Dos Santos. 2019. "Dynamic Multicontext Segmentation of Remote Sensing Images Based on Convolutional Networks." *IEEE Transactions on Geoscience and Remote Sensing* 57 (10): 7503–7520. doi:10.1109/TGRS.2019.2913861.
- Paisitkriangkrai, S., J. Sherrah, P. Janney, and V.-D. Hengel. 2015. "Effective Semantic Pixel Labelling with Convolutional Networks and Conditional Random Fields." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, pp. 36–43.
- Paisitkriangkrai, S., J. Sherrah, P. Janney, and A. Van Den Hengel. 2016. "Semantic Labeling of Aerial and Satellite Imagery." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9 (7): 2868–2881. doi:10.1109/JSTARS.2016.2582921.
- Piramanayagam, S., E. Saber, W. Schwartzkopf, and F. W. Koehler. 2018. "Supervised Classification of Multisensor Remotely Sensed Images Using a Deep Learning Framework." *Remote Sensing* 10 (9): 1429. doi:10.3390/rs10091429.
- Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, Springer, pp. 234–241.
- Saxena, A., S. Chung, and A. Ng. 2005. "Learning Depth from Single Monocular Images." *Advances in Neural Information Processing Systems* 18: 1161–1168.
- Saxena, A., S. H. Chung, and A. Y. Ng. 2008. "3-D Depth Reconstruction from a Single Still Image." *International Journal of Computer Vision* 76 (1): 53–69. doi:10.1007/s11263-007-0071-y.
- Sherrah, J. 2016. "Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery." *arXiv preprint arXiv:160602585*.

- Simonyan, K., and A. Zisserman. 2015. "Very Deep Convolutional Networks for Large-Scale Image Recognition." 3rd International Conference on Learning Representations, San Diego, CA, USA, ICLR.
- Srivastava, S., M. Volpi, and D. Tuia. 2017. "Joint Height Estimation and Semantic Labeling of Monocular Aerial Images with CNNs." *IEEE International Geoscience and Remote Sensing Symposium*, Fort Worth, TX, USA, IEEE, pp. 5173–5176.
- Sun, X., S. Shen, X. Lin, and Z. Hu. 2017. "Semantic Labeling of High-Resolution Aerial Images Using an Ensemble of Fully Convolutional Networks." *Journal of Applied Remote Sensing* 11 (04): 042617. doi:10.1117/1.JRS.11.042617.
- Thoma, M. 2016. "A Survey of Semantic Segmentation." *arXiv preprint arXiv:160206541*.
- Volpi, M., and D. Tuia. 2017. "Dense Semantic Labeling of Subdecimeter Resolution Images with Convolutional Neural Networks." *IEEE Transactions on Geoscience and Remote Sensing* 55 (2): 881–893. doi:10.1109/TGRS.2016.2616585.
- Wang, L., R. Li, D. Wang, C. Duan, T. Wang, and X. Meng. 2021. "Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images." *Remote Sensing* 13 (16): 3065. doi:10.3390/rs13163065.
- Wang, H., Y. Wang, Q. Zhang, S. Xiang, and C. Pan. 2017. "Gated Convolutional Neural Network for Semantic Segmentation in High-Resolution Images." *Remote Sensing* 9 (5): 446. doi:10.3390/rs9050446.
- Zhang, H., K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. 2018. "Context Encoding for Semantic Segmentation." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 7151–7160.
- Zhang, M., X. Hu, L. Zhao, Y. Lv, M. Luo, and S. Pang. 2017. "Learning Dual Multi-Scale Manifold Ranking for Semantic Segmentation of High-Resolution Images." *Remote Sensing* 9 (5): 500. doi:10.3390/rs9050500.
- Zhang, L., L. Zhang, and B. Du. 2016. "Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art." *IEEE Geoscience and Remote Sensing Magazine* 4 (2): 22–40. doi:10.1109/MGRS.2016.2540798.
- Zhao, H., J. Shi, X. Qi, X. Wang, and J. Jia. 2017. "Pyramid Scene Parsing Network." Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, pp. 2881–2890.
- Zheng, X., L. Huan, G.-S. Xia, and J. Gong. 2020. "Parsing Very High Resolution Urban Scene Images by Learning Deep ConvNets with Edge-Aware Loss." *ISPRS Journal of Photogrammetry and Remote Sensing* 170: 15–28. doi:10.1016/j.isprsjprs.2020.09.019.
- Zheng, Z., Y. Wan, Y. Zhang, S. Xiang, D. Peng, and B. Zhang. 2021. "CLNet: Cross-Layer Convolutional Neural Network for Change Detection in Optical Remote Sensing Imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 175: 247–267. doi:10.1016/j.isprsjprs.2021.03.005.