HECR-Net: Height-Embedding Context Reassembly Network for Semantic Segmentation in Aerial Images

Wenjie Liu, *Graduate Student Member, IEEE*, Wenkai Zhang, *Member, IEEE*, Xian Sun, *Senior Member, IEEE*,
Zhi Guo, *Member, IEEE*, and Kun Fu, *Member, IEEE*

***Abstract*—Semantic segmentation in aerial images has become an indispensable part in remote sensing image understanding for its extensive application prospects. It is crucial to jointly reason the 2-D appearance along with 3-D information and acquire discriminative global context to achieve better segmentation. However, previous approaches require accurate elevation data (e.g., nDSM and Digital Surface Model (DSM)) as additional inputs to segment semantics, which sorely limits their applications. On the other hand, due to the various forms of objects in complex scenes, the global context is generally dominated by features of salient patterns (e.g., large objects) and tends to smooth inconspicuous patterns (e.g., small stuff and boundaries). In this article, a novel joint framework named height-embedding context reassembly network (HECR-Net) is proposed. First, considering the fact that the corresponding elevation data is insufficient while we still want to exploit the serviceable height information, to alleviate the above data constraint, our method simultaneously predicts semantic labels and height maps from single aerial images by distilling height-aware embeddings implicitly. Second, we introduce a novel context-aware reorganization module to generate a discriminative feature with global context appropriately assigned to each local position. It benefits from both the global context aggregation module for ambiguity eliminating and local feature redistribution module for detailed refinement. Third, we make full use of the learning height-aware embeddings to promote the performance of semantic segmentation via introducing a modality-affinitive propagation block. Finally, without bells and whistles, the segmentation results on ISPRS Vaihingen and Potsdam data set illustrate that the proposed HECR-Net achieves state-of-the-art performance.**

***Index Terms*—Aerial imagery, context-aware reorganization, height-aware embeddings, modality-affinitive, semantic segmentation.**

Manuscript received July 13, 2021; revised August 16, 2021; accepted August 29, 2021. Date of publication September 3, 2021; date of current version September 22, 2021. This work was supported by the National Natural Science Foundation of China under Grant 61725105. *(Corresponding author: Wenkai Zhang.)*

Wenjie Liu, Xian Sun, and Kun Fu are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, with the Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, with the University of Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: liuwenjie18@mails.ucas.ac.cn; sunxian@aircas.ac.cn; kunfuiecas@gmail.com).

Wenkai Zhang and Zhi Guo are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, and also with the Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhangwk@aircas.ac.cn; guozhi@mail.ie.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3109439

I. Introduction

Semantic segmentation is a significant constituent of image interpretation, whose goal is to parse the whole image and assign categories pixel by pixel. In recent years, semantic labeling in aerial images has been introduced widespreadly in many fields, such as disaster prediction, building extraction, and resource exploration.

The tremendous success of convolutional neural networks (CNNs) has born out of the formidable abilities of feature extraction in computer vision tasks [1], such as image classification [2], [3], object recognition and detection [4], [5], and scene segmentation [6]–[8]. In particular, fully convolutional networks (FCNs) [6] have shown prominent improvements when applying them to dense prediction tasks like semantic segmentation and height estimation. Inspired by the idea of FCN [6], an increasing body of research [9]–[12] has been devoted to design FCN-based frameworks for the semantic segmentation in the aerial images. Nevertheless, there are still some technical limitations as a result of the diversity of scenes in the remote sensing scene, which cause mismatched relationships. As shown in Fig. 1, there are three examples from aerial scenarios. The first row reveals misclassifications concerning the inconspicuous classes, where impervious surface is wrongly classified as low vegetation. Moreover, an issue of mismatched relationship where the lawn on the roof is predicted as the low vegetation erroneously is shown in the second row. In the last row, the similar 2-D appearance between tree and low vegetation brings about misclassifications, which makes semantic segmentation more challenging. In short, we mainly consider two obstacles: the diversity of object patterns (e.g., large or small) and the existence of objects with similar spectral characteristics but belonging to different categories (e.g., roads and roofs, trees and lawns).

For the first handicap, semantic segmentation is a pixel-level dense prediction task, and, therefore, not only the dominated salient stuff but also the unremarkable objects ought to be parsed well. Since the traditional FCN extracts feature maps without adequate context, several global context aggregation (GCA) methods [13]–[21] have been proposed to increase the receptive field of the FCN. Here we simply sort out two methods on capturing global context dependencies. The first is multiscale (MS) aggregation method. There are certain studies [13]–[15], [18] to capture global dependencies by adopting global pooling or MS aggregation modules. The other is attention-based aggregation method. Inspired by the idea of nonlocal neural networks [22],

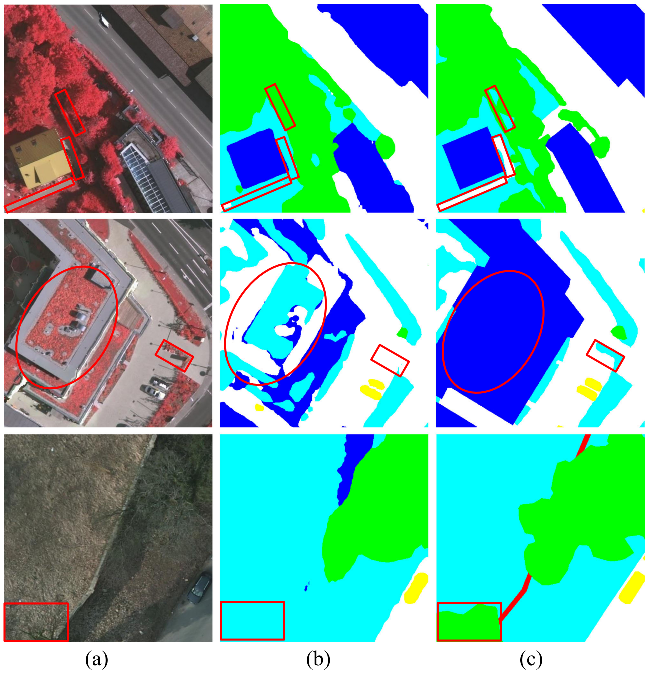


Fig. 1. Example scenes about misclassification when using the present mainstream methods. In the first row, there are several misclassifications concerning the inconspicuous classes where impervious surface is wrongly classified as low vegetation. The middle row indicates an issue of mismatched relationship where the lawn on the roof is predicted as the low vegetation erroneously. In the last row, the similar 2-D appearance between trees and low vegetation brings about misclassifications, which makes semantic segmentation more challenging.

several methods [19], [20], [23]–[27] were proposed to selectively aggregate heterogeneous context by learning spatial attention, channel attention, and class attention. Compared with the standard convolution layer gathering features in a small local area, GCA methods utilize the above two methods over the whole image to consistently improve the performance of the basic FCN. Nevertheless, global contextual information often has smooth representations, which are usually controlled by the patterns of larger objects, while the representation of unobtrusive stuff will be somewhat weakened or even ignored. In order to settle this issue, the discriminative features in local position are desired. Accordingly, we propose a compact and universal context-aware reorganization (CAR) module to acquire global context information over the entire feature map, and then adaptively assign it to each local pixel in the light of the pattern size of each location. The assignment procedure is allocated by a series of active mask maps, which describe the spatial extents of each pattern after local feature redistribution.

For the second handicap, in contrast to natural scene, very high-resolution (VHR) aerial images have more complex spectral properties, in which different categories of objects probably have similar 2-D appearance. It is challenging to parse the semantic regions with similar spectral properties only utilizing optical images (e.g., IRRG). Therefore, some semantic segmentation methods [28]–[33] based on multimodal data have been proposed to leverage additional 3-D elevation data (e.g., DSM, nDSM) to effectively settle ambiguities which are challenging to 2-D appearance solely approaches. The conventional structures of semantic segmentation based on multimodal datas can be summarized as early, late, and multistage fusion, as shown in Fig. 2. However, all of the above methods require the elevation data (DSM here) as an additional input, which is not convenient to be collected from the scene and to be aligned with the optical images. We argue that it is possible to embed the geometric information (height above ground) for semantic segmentation with only monocular image as input. Inspired by [34]–[37], we propose a joint reasoning framework composed of semantic segmentation network and height estimation network to extract 2-D and 3-D features from single IRRG images. The biggest difference between the framework and traditional methods is that we take elevation data as the supervision information to extract height-aware embeddings. Furthermore, we fuse the distilled height-aware embeddings with the semantic features from 2-D appearance via the proposed modality-affinitive propagation (MAP) module, which utilizes the cross-task affinity patterns to guide semantic segmentation. Through the joint training of the above two tasks, the goal of our method is to improve segmentation performance, while taking height-aware embedding into consideration implicitly.

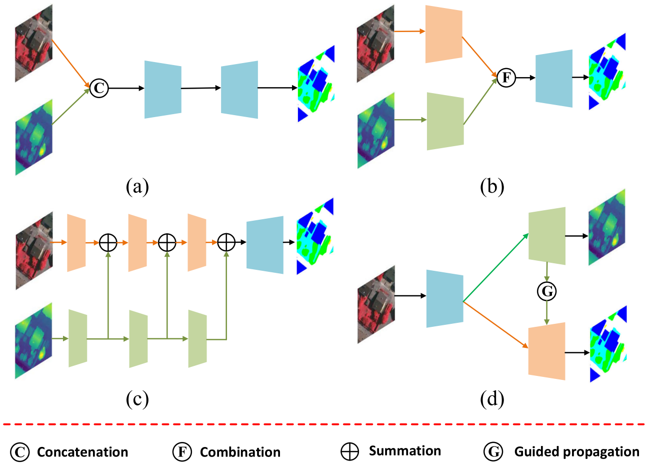


Fig. 2. Fusion structures for semantic segmentation based on multimodal data. The orange color and green color demonstrate IRRG branch and DSM branch, respectively. The quadrilaterals represent encoders and decoders.

To summarize, the main contributions of this article are summarized as follows.

1) We present a height-embedding context reassembly network (HECR-Net), an end-to-end joint framework that predicts semantic labels and distills height-aware embeddings implicitly, which effectively guides semantic segmentation over the input aerial images.
2) To focus more on the inconspicuous objects, a CAR module is proposed to generate a discriminative feature map. In this procedure, the global context information can be adaptively assigned to each local position.
3) The MAP is proposed to perform cross-modality learning and fusion. A modality-affinitive adaptively combination block is designed for the former while the propagation block is used for the latter.

The rest of this article is organized as follows. Section II briefly introduces the related work on semantic segmentation and height estimation. Next, Section III illustrates the details of our proposed multitask joint reasoning method. In Section IV, the experimental evaluations as well as the corresponding results analysis are provided. Finally, Section V concludes this article.

II. Related Works

This section gives a brief introduction to several previous work: semantic segmentation concerning contextual modeling as well as multimodal fusion, and height estimation.

A. Semantic Segmentation

Compared with hyperspectral images [38], [39] and natural images [15], [18], VHR aerial images bring challenges to the task of semantic segmentation for complex spatial details. Conventional segmentation methods on remote sensing semantic segmentation methods [40], [41] mainly extracted useful low-level and hand-crafted features from input images and then mapped the features into label categories by adopting a supervised classifier. In light of the powerful feature learning and representation abilities of deep learning methods, semantic segmentation has made significant breakthrough. FCNs [6] first discard the full connection layer from the ordinary classification network [2], [42] and replace it with the corresponding convolution layer, so as to achieve dense segmentation of the input image. Since then, FCNs turn to be the most popular baselines for semantic segmentation, and numerous model variants are proposed to improve the segmentation performance. Here, we review only the most related work in terms of handling contextual modeling and multimodal fusion.

Contextual Modeling: In recent years, a large body of literatures have explored contextual modeling, which is crucial to semantic segmentation [13], [15]–[20], [25], [26], [43]. An intuitive idea is to apply new layers to increase receptive field while maintaining a larger spatial resolution. In PSPNet [18], multiple pooling operations were employed to extract the features of different regions to enrich context information at different scales. The DeepLab series [13]–[15], [44] aggregated context information at multiple scales by introducing an atrous spatial pyramid pooling (ASPP). Based on [44], DenseASPP [45] combined the advantages of parallel and cascade atrous convolution to generate MS features in a larger range and acquire the serviceable context information. EncNet [20] and DFN [19] utilized attention mechanism to expand the differences among feature maps and obtain more context information from the perspective of feature channel dimension. In addition, inspired by nonlocal network [22], numerous advanced contextual approaches made full use of the self-similarity manner to gather global spatial information, which achieved the impressive results in scene-understanding tasks. DANet [25] explored spatial and channel relationships from all pixels by means of nonlocal operator [22]. Compact generalized nonlocal [46] considered the global relationships of channel dimensions based on the nonlocal network [22]. CCNet [27] harvested the long-range dependencies via cascading two criss-cross attention modules to economize both memory and computation cost. Different from previous work which focused on global context modeling, in this article, a novel CAR module, which benefits from both global and local information, is proposed to generate a discriminative feature.

Multimodal Fusion: In the last decade, an increasing body of Deep Convolutional Neural Network (DCNN)-based fusion methods [47]–[49] focused on extracting and fusing complementary feature information from multiple modalities have been proposed to enhance the robustness of feature representations. The existing fusion methods can be summarized as early, late, and multistage fusion methods. Early fusion method [see Fig. 2(a)], just as the name implies, is to concatenate multiple modalities directly as a four- or six-channel input along the channel dimension and then feed them into a conventional unimodal network [29]. However, such methods are not conducive to extracting complementary features and capturing cross-modal interdependencies well. A great deal of work has turned to the dual-branch fusion framework with two separated encoders and a single decoder. Late fusion methods [31], [32] [see Fig. 2(b)] train two independent encoders and then combine the modality-specific features in an integration manner (e.g., element-wise summation or concatenation). These methods can distill multimodal features well by a parallel branch, but also bring a lot of parameters and computation. Instead of fusing modality-specific features at early and late stages, multistage fusion methods [50], [51] [see Fig. 2(c)] fuse the features at multiple stages. Although these approaches have got extensive achievements, they require elevation data associated with the optical images as additional inputs. Instead of directly taking elevation data as inputs, we propose a novel height-aware embedding structure based on multitask decoder to extract contextual and geometric features jointly, which further conduces to better semantic segmentation performance.

B. Height Estimation

The methods on height estimation in remote sensing scene are closely involved in the work concerning the monocular depth estimation in the field of scene reconstruction. There are plenty of existing approaches for monocular depth estimation, which can be approximately summarized into two types: CNN-based approaches [34], [52]–[55] and some hybrid approaches incorporating CNNs with probabilistic graphic models [56]–[58]. A MS neural network was proposed to recurse the depth in [34], which was composed of coarse-scale and fine-scale network. The former was responsible for a coarse prediction of the global depth, while the latter was based on the former for a fine prediction of the depth in the local area. Eigen and Fergus [52] designed a general MS network, which could be applied to the task of semantic segmentation, surface normal estimation, and depth estimation. By sharing the backbone network among the multiple tasks, the structure simplified the implementation of multitask system, which greatly reduced the network parameters and improved the network efficiency. In the hybrid approaches, Wang *et al.* [59] predicted the depth maps and segmentation results jointly from a single image by proposing a joint framework combined with a hierarchical Conditional Random Field

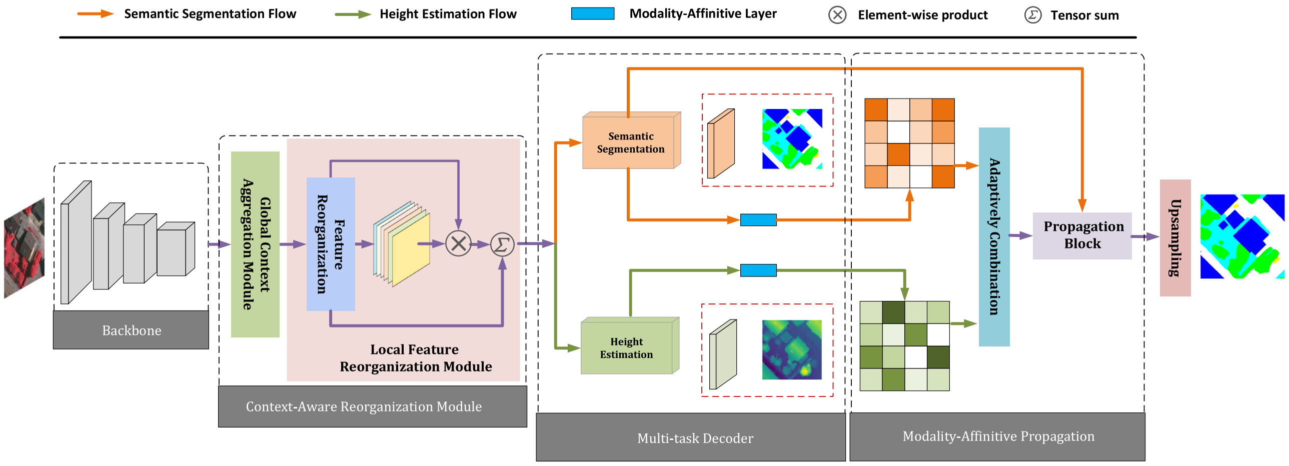


Fig. 3. Overview of the proposed height-embedding context reassembly network (HECR-Net). It consists of a backbone, context-aware reorganization (CAR) module, multitask decoder, and modality-affinitive propagation (MAP) block.

(CRF). The purpose of hierarchical CRF was to optimize the final results.

The researches on height estimation in remote sensing is relatively few. Early efforts employed the shadow shape to estimate the height of the artificial satellite image with only several control points [60] or a sparse Digital Terrain Model (DTM) [61]. Following the CNN-based depth estimation methods, Mou and Zhu [36] predicted height map by introducing an encoder–decoder network encompassing residual learning and a skip connection. In addition, a unified framework for estimating height and segmenting semantics from single aerial images was proposed in [62]. The joint framework was optimized by a multitask loss function. Through the method, the performance of semantic segmentation can benefit from the geometric information extracted from the task height estimation. Nevertheless, the sharing and fusion of complementary information between different tasks (e.g., semantic segmentation and height estimation) are not well explored.

III. Methodology

This part first briefly introduces our framework topology and our main contributions inspired by our design criteria. Then, CAR module is proposed to appropriately handle the out-off-balance spread of context from salient and inconspicuous objects in global context aggregation (GCA) module. Then, we introduce the details of height-aware embeddings via a multitask model and modality-affinitive propagation module to introduce the corresponding height information to assist semantic segmentation. Finally, the proposed framework is trained jointly in an end-to-end manner by employing a multitask loss function.

A. Overview

Contextual and geometric information are critical to semantic segmentation in VHR aerial images, which are widely explored in various methods [12], [31], [33]. In this section, we propose a unified network to learn both contextual and geometric features, and fuse the complementary information to enable height-embedding semantic labeling. As shown in Fig. 3, the whole network framework follows the encoder–decoder design principle of multitask outputs. And the whole network includes four components: shared backbone, CAR module, multitask decoder, and MAP block.

Concretely, given an image $I \in \mathbb{R}^{3\times H\times W}$, where 3, H, and W indicate the IRRG or RGB channels, height, and width of I, respectively, our network architecture first passes I through a shared backbone ResNet-101 pretrained over the ImageNet data set [63], following the majority of the previous works [18], [25], [27]. It is worth mentioning that, we utilize atrous convolutions to maintain the spatial resolution of output feature to 1/8 of the original image by removing the last two downsampling operations in stage-3 and stage-4. We employ a hard parameter-sharing mechanism between the following two tasks. In other words, we adopt a backbone network to embed the feature representations of multiple tasks into the same semantic space. In addition, CAR module aims to model the long-range spatial relationship and then adaptively distribute the long-range dependencies according to the pattern size of each pixel position, which will be illustrated in Section III-B. Then in order to extract both contextual and geometric information from I, we feed the output features of CAR module to two task-specific decoders, where the upper stream segments semantics while the lower stream distills the height-embeddings by predicting height maps. Then, MAP block first learns two affinity matrices to acquire the pair-wise relationships in respective tasks and then reorganize the matrix with the other affinity matrix adaptively to aggregate the task-complementary information. After that, we transfer the recombinant affinitive modalities back to the feature maps via a propagation block so as to realize height-aware semantic segmentation. Finally, we upscale the feature maps by means of a upsampling block to predict the final results with higher resolution. The whole network can be optimized in an end-to-end manner by a joint objective function and the details of the CAR, height-aware embedding, modality-affinitive propagation, and multitask objective function will be illustrated in the following sections.

B. Context-Aware Reorganization

Fig. 3 illustrates the overall topology of CAR module, which captures contextual information, especially in the long range over the entire feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ learned from a shared backbone network, and then adaptively distribute the long-range dependencies to each pixel position of the output feature.

Global Context Aggregation: As discussed in Section II-A, GCA module is of great importance for scene parsing [13], [15], [17], [18], [20], [25], [26]. GCA intended for modeling long-range dependencies over local feature representations can be modeled as

$$\mathbf{F}_{gca} = \frac{1}{C(\mathbf{X})} \sum_{i}^{N} \sum_{\forall j \in \theta(i)} \mathbf{G}(x_i, x_j) g(x_j) \qquad (1)$$

where x_i and is the feature response at pixel *i* of the input entire feature map $\mathbf{X}$, and $\forall j \in \theta(i)$ is the collection that enumerates all possible pixels related to x_i. $G(x_i, x_j)$ can be any learned parameters or pairwise function computing representing the relationship between x_i and all x_j. For brevity, g is set to the form of a unary linear embedding: $g(x_j) = W_j x_j$, where W_j is a learnable weight matrix. Here we set normalization factor as $C(\mathbf{X}) = N$, in which $N = H \times W$ indicates the number of pixels in $\mathbf{X}$.

It is worth mentioning that, any of the methods described in Section II-A can be used as GCA module for discussion. In this article, we utilize the compact generalized nonlocal (CGNL) [46] module downsampled by a factor of 2 to model long-range dependencies with lightweight computation and memory. In particular, CGNL augments the nonlocal operation in differentiating fine-grained object regions by taking different channels information into account simultaneously, and then calculates the global feature statistics to form $\mathbf{F}_{gca}$. We additionally give a contrastive analysis among the state-of-the-art of GCA module in Section IV-D.

Local Feature Redistribution: Despite the GCA modules can capture contextual information more efficiently, they are easily partial to features from large object regions containing more sample pixels, for they collect global statistics of features in large receptive fields. Therefore, the global context obtained by each pixel position tends to smooth the object regions, which contain small patterns. To tackle this problem, local feature redistribution (LFR) module, as a spatial operator, can recalculate the spatial size of the object regions based on the feature map $\mathbf{F}_{gca}[:, :, c]$ to adaptively distribute $\mathbf{F}_{gca}$ to each pixel position. Next, we will elaborate the specific process of LFR.

Given an input feature $\mathbf{F}_{gca} \in \mathbb{R}^{C\times H\times W}$, we first adopt a set of 3×3 convolution layers to shrink the spatial extent of the feature to $C \times \frac{H}{s} \times \frac{N}{s}$, where s is the downsampled ratio (supposing s is an integer). Inspired by [64], we propose a local feature reorganization module that predicts learnable weights for each pixel p according to its contextual information and then reassembles the features in a predefined neighborhood to generate a novel feature $\mathbf{F}_{lfr} \in \mathbb{R}^{C\times H\times W}$.

The entire reorganization block consists of channel compressor, weight learning, and feature reorganization, as shown in Fig. 4. Here, we suppose that the size of the learning weight is $r \times r$, where the larger the weight, the larger the receptive field and the larger the computation. If we want to use different weights for each position of the output feature, we need to predict the size of the weight to be $H \times W \times r \times r$.

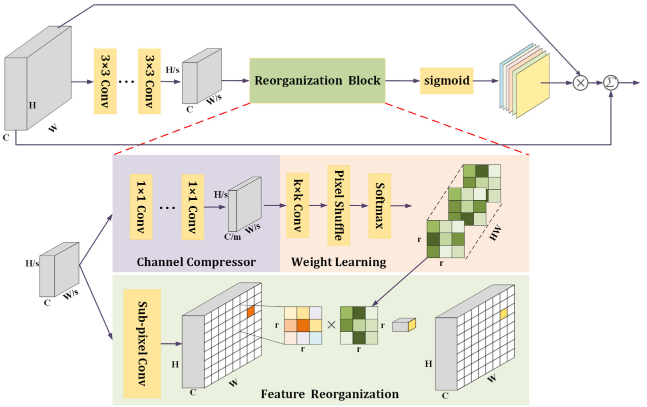


Fig. 4. Full pipeline of the proposed for local feature redistribution (LFR) module, where reorganization block consists of three important parts, i.e., channel compressor, weight learning, as well as feature reorganization.

Algorithm 1: Local Feature Redistribution.

```
Input: The input feature: $\mathbf{F}$; scale: $s$; the size of $\mathbf{F}$ :
    $(H_{in}, W_{in})$; the $k \times k$ subregion of $\mathbf{F}$ centered at the
    source position $(i, j)$: $N(\mathbf{F}(i, j), k)$; the weight prediction
    function: $\mathbf{\Gamma}$
Output: The reassembled feature: $\mathbf{F}'$
Channel compressor for $\mathbf{F}$
Calculate the size of
    $\mathbf{F}'$: $H_{\text{out}} = H_{\text{in}} \times s, \mathbf{W}_{\text{out}} = \mathbf{W}_{\text{in}} \times s$
Predict normalized weights
    $\mathbf{W}(i,' j') = \mathbf{\Gamma}(N(\mathbf{F}(i, j), k))$
Sub-pixel convolution for $\mathbf{F}$, then reassembling feature
for each $i' \in [0, H_{\text{out}}]$ do
    for each $j' \in [0, W_{\text{out}}]$ do
        Source position $i = \lfloor i'/s \rfloor, j = \lfloor j'/s \rfloor$
        Weight $\mathbf{W}(i,' j') = \mathbf{\Gamma}(N(\mathbf{F}(i, j), k))$
        $\mathbf{F}'(i,' j') = \mathbf{F}(i, j) \cdot \mathbf{W}(i,' j')$
    end for
end for
```

The pseudocode of our algorithm can be shown in Algorithm 1. We first compress the feature channels from C to C_m by adopting a 1×1 convolution layer, which can reduce the parameters and computational cost without harming the performance. Then, we utilize a $k \times k$ (5 by default) convolution layer to capture the contextual information to predict normalized weights, where the number of input channels is C_m while the number of output channels is $r^2 \times s^2$. After that, we expand the predicted weights in the first step along the channel dimension to obtain the weight with the size of $H \times W \times r \times r$ by pixel shuffling. Finally, we utilize softmax function to normalize the weights. The weight-learning process $\mathbf{\Gamma}$ is formulated as

$$\mathbf{W}_{p'} = \mathbf{\Gamma}(N(\mathbf{F}_p, k)) \qquad (2)$$

where W'_p denotes the learnable weight for each target position $p' = (i,' j')$ in the output $\mathbf{F}'$, and $N(\mathbf{F}_p, k)$ is the $k \times k$ subregion of the input feature map $\mathbf{F}$ centered at the source position p. For each position $p' = (i,' j')$, there is a corresponding location $p = (i, j)$, where $i = \lfloor i'/s \rfloor, j = \lfloor j'/s \rfloor$ and s is the downsampled ratio. The weight-learning process is in charge of generating the weights in a way of local information perception. After that, we use softmax function to normalize the weight.

For every position in the output $\mathbf{F}'$, we feed it back to the input $\mathbf{F}$ and take out the corresponding square region $N(\mathbf{F}_\mathbf{p}, r)$ centered at $p = (i, j)$. Specifically, we use the input feature through the subpixel convolution layer [65] to get the feature map with the same dimension as the output feature. Then, the final local eigenvalue is obtained by dot product of the region $\mathbf{N}(\mathbf{F}_p, r)$ and the predicted weight $\mathbf{W}_{p'}$. Note that the same location in different channels shares the weights. The features reorganization step can be generated as follows:

$$\mathbf{F'}_{p'} = \mathbf{\Theta}(N(\mathbf{F}_\mathbf{p}, r), \mathbf{W}_{p'}) \tag{3}$$

$$\mathbf{F}_{lfr} = \sum_{p'} \mathbf{F'}_{p'} = \sum_{p'} \sum_{a=-m}^{m} \sum_{b=-m}^{m} \mathbf{F}_{(i+a,j+b)} W_{p'_{(a,b)}} \tag{4}$$

where $m = \lfloor r/2 \rfloor$, and the function $\mathbf{\Theta}$ is a weighted sum operation. In this way, the position p' is upsampled and reassembled according to the context of each local position in the region of $N(\mathbf{F}_p, r)$ rather than the distance of positions. Therefore, the corresponding position information from the local region can get more attention.

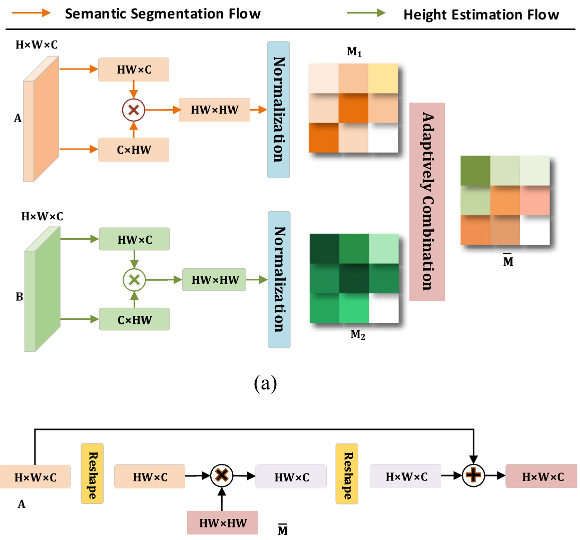
After that, we apply a sigmoid function σ to recalculate the spatial extents of each pattern as an adaptive weight $\mathbf{S}$. Then, we utilize S to weight the global feature maps $\mathbf{F}_{gca}$ and carry out an identity map operation to acquire the final feature $\mathbf{F}_{car} \in \mathbb{R}^{C \times H \times W}$ as follows:

$$\begin{aligned} \mathbf{F}_{car} &= \mathbf{S} \otimes \mathbf{F}_{gca} + \mathbf{F}_{gca} \\ &= \sigma(\mathbf{F}_{lfr}) \otimes \mathbf{F}_{gca} + \mathbf{F}_{gca} \end{aligned} \tag{5}$$

where $\otimes$ is the Hadamard product, an element-wise multiplication. From the view of $\mathbf{F}_{gca}$, the introduction of the local weight S encourages the global feature maps $\mathbf{F}_{gca}$ suffering from coarse feature representations to distill more details from $\mathbf{F}_{lfr}$. As a result, CAR module is able to distribute the global information to each position conditionally, which could be regarded as an explicit contributing factor for modeling long-range dependencies.

C. Height-Aware Embedding Via a Multitask Model

Over the past few years, some researchers [28], [29], [31], [32] have proved that the introduction of geometry information can further enhance the performance of semantic labeling by designing multimodal convolutional neural network. However, this framework explicitly requires aligned height annotation as inputs, which impedes its usage in many practical applications due to its vast GPU memory occupation and prohibitive computational cost. Here, we design a multitask dense prediction framework to distill the height-aware embedding and perform semantic segmentation simultaneously from the single optical images to learn 2-D and 3-D information jointly. Particularly, given an input image $\mathbf{X} \in \mathbb{R}^{3 \times H \times W}$, the traditional semantic segmentation methods train a model $\mathbf{M}_\theta$ that distills contextual information from ground-truth $\mathbf{S}$ by minimizing the segmentation loss, formalized as a fully optimization problem:



(a)

(b)

Fig. 5. The details of Modality-Affinitive Adaptively Combination block and Propagation block. ⊗ indicates the matrix multiplication. ⊕ indicates the element-wise sum. (a) Modality-Affinitive Adaptively Combination block. Here we adopt dot-product for computing similarities. (b) Propagation Block.

$$\min_{\ell} \mathbf{E}[\ell(\mathbf{M}_\theta(\mathbf{X}), \mathbf{S})] \tag{6}$$

where $\mathbf{E}[\cdot]$ denotes statistical expectation and $\ell(\cdot)$ is a segmentation loss function, such as cross-entropy loss. Inspired by the formulation, the proposed multitask model M can be optimized as follows:

$$\min_{\ell_1, \ell_2} \mathbf{E}[\ell_1(\mathbf{M}_\zeta(\mathbf{X}), \mathbf{H}) + \ell_2(\mathbf{M}_\eta(\mathbf{X}), \mathbf{S})] \tag{7}$$

in which $\mathbf{E}[\ell_1(\cdot)]$ is the elevation estimation term. Here H is the elevation data providing geometric information and $\ell_1(\cdot)$ is a regression loss to train height estimation network $\mathbf{M}_\zeta$. The second term $\mathbf{E}[\ell_2(\cdot)]$, similar to the above formulation, is a semantic segmentation term intended for extracting the 2-D semantic information, in which $\mathbf{M}_\zeta$ shares weights with $\mathbf{M}_\eta$ for the shared encoder partially.

D. Modality-Affinitive Propagation

Here we elaborate the proposed MAP method aiming to leverage the complementarity of semantic information and geometric information to improve semantic segmentation. First, in order to represent the pairwise similarities for each task, we learn a semantic affinity matrix and a geometric affinity matrix, respectively, by two affinity layers, as shown in Fig. 5(a). Assume that the output feature map of the last layer in multitask decoder is $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, so the pairwise similarities computation are done by matrix multiplication and then receive the affinity matrix $\mathbf{M} = \mathbf{F}\mathbf{F}^{\intercal} \in \mathbb{R}^{HW\times HW}$.

In addition, since the value of each row in the affinity matrix M indicates the pairwise similarities between one position and the other positions and in order to reduce the influence of scale, a softmax operation is performed on each row of M to conduct normalization. Note that there is not an additional auxiliary loss to supervise M, because such supervision is difficult to define for some tasks and can add extra memory overhead. After that, we insert adaptively combination blocks into networks to aggregate the cross-task relatively complemented information. Assuming semantic segmentation and height estimation as T_1, T_2, respectively, as well as the respective affinity matrices as $\mathbf{M}_1, \mathbf{M}_2$, next we need to obtain two weighting factors γ_i (where $i = 1, 2$; $\gamma_1 + \gamma_2 = 1$) to adaptively integrate the matrices as

$$M = \gamma_1 \cdot \mathbf{M}_1 + \gamma_2 \cdot \mathbf{M}_2 \tag{8}$$

where the weighting factor γ_1 is a trainable multidimensional matrix, which can be bound to the model and becomes a learnable parameter in the model. We obtain the factor γ_1 by converting a fixed untrainable tensor into a trainable parameter. In this way, γ_1 and γ_2 constantly modify their values in the process of learning to achieve the optimal effect. Finally, we obtain the combined affinity matrix $\overline{M}$ by propagating cross-task affinitive modalities and then we spread such matrix as guidance to improve the accuracy of semantic labeling. As illustrated in Fig. 5(b), the fusion process designed by some convolution units is performed on the original semantic feature. We perform propagation by multiplying the affinity matrix $\overline{M}$ by the original feature. Moreover, similar to a residual connection [42], we add the original semantic feature to the result in an element-wise summation manner to maintain its initial behavior.

E. Multitask Objective Function

The proposed network is featured by dual-task loss functions, i.e., the *lassification loss* and the *regression loss*. Nevertheless, distribution of the class labels in the existing remote sensing data sets, e.g., *Vaihingen* [66] and *Potsdam* [67], is extremely imbalanced, because the pixels belonging to the different class labels are different. As demonstrated in Fig. 6, the distribution of the above two data sets is dramatically imbalanced, which will bias the training process toward those dominant samples and bring about low segmentation performance for the small samples. In order to address the category imbalance issues, we follow focal loss [68] to take the weighted cross-entropy loss function as our segmentation optimization function

$$L_s = -\sum_i \sum_c w_i \times \ell_i \times log(p_i, c) \tag{9}$$

in which i indexes the position and w_i set by the inverse category frequency indicates the balance factor for position i to address the class imbalance. $c \in [1, 2, \ldots, C]$ indicates the category. ℓ_i is the semantic label of position i and (p_i, c) is the predicted probability of position i belonging to category c.

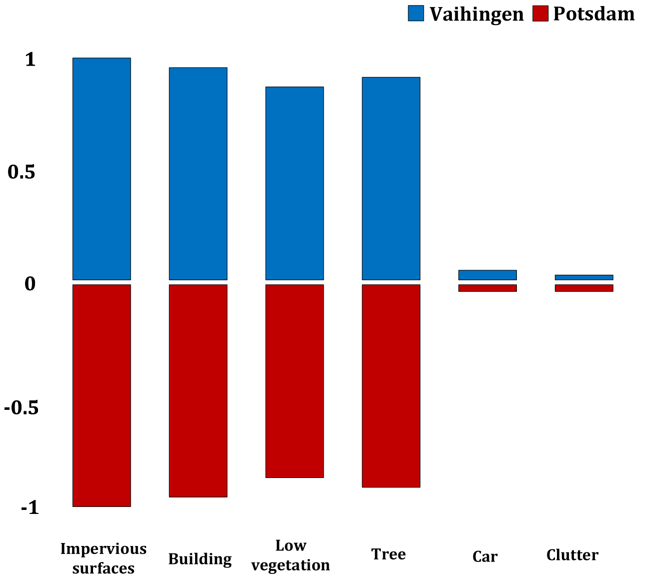


Fig. 6. The statistical distribution of semantic categories on the training set of Vaihingen (top) and Potsdam (bottom). Horizontal axis represents the semantic categories, while vertical axis indicates the relative proportion of pixels.

For height estimation task, we utilize smooth L1 loss as height supervision formulated as

$$L_h = \sum_i^n \begin{cases} 0.5(h_i - H_i)^2, & \text{if } |h_i - H_i| \le 1 \\ |h_i - H_i| - 0.5 & \text{otherwise} \end{cases} \tag{10}$$

where i indexes the position and n indicates the total number of positions. Here h_i and H_i represent the the final output height and ground-truth height at position i separately.

Following PSPNet [18], we add an auxiliary supervision L_a after the third stage of the ResNet-101 [42] to help optimize the learning process. Finally, we define the joint objective function as

$$L = L_s + \lambda_1 L_a + \lambda_2 L_h \tag{11}$$

where λ_1, λ_2 are set 0.5, 1 by default to balance the objective function. In Section IV-D, we perform the ablation studies for the influence generated by the choice of super parameter λ_1 and λ_2 on segmentation performance.

IV. Experiments

In this part, we conduct sufficient experiments on Vaihingen data set [66] and Potsdam data set [67] to evaluate the effectiveness of our proposed framework for semantic segmentation.

A. Datatsets

Vaihingen: There are 33 images in the Vaihingen data set in total, where the training set embodies 16 images and the remaining 17 tiles are used to evaluate our proposed network following the previous works [12], [29], [69]. Each aerial image is provided with orthophoto images, semantic labels, and digital surface models (DSM and nDSM). Each image has a ground sampling distance of 9 cm and three channels of near-infrared, red, as well as green.

Potsdam: The Potsdam data set involves 38 image tiles in all, where the training set contains 24 images and the remaining 14 tiles are used to evaluate our proposed network. Each aerial image is composed of four channels (near-infrared, red, green, and blue) and a spatial resolution of 5 cm/pixel.

B. Evaluation Metrics

We adopt overall accuracy (OA), mean IoU (intersection over union), and F_1 score as evaluation metrics to evaluate the segmentation performance. OA and IoU are two widely used metrics, which are calculated separately as

$$\mathrm{IoU} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}} + N_{\mathrm{FN}}} \tag{12}$$

$$\mathrm{OA} = \frac{N_{\mathrm{TP}} + N_{\mathrm{TN}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}} + N_{\mathrm{FN}} + N_{\mathrm{TN}}} \tag{13}$$

in which N_{TP}, N_{TN}, N_{FP}, and N_{FN} are the number of true positives, true negatives, false positives, and false negatives, respectively. Note that we count the F_1 score for all categories except background as follows:

$$F_1 = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{14}$$

where precision $= \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}}$, recall $= \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}}$.

Note that, in order to reduce the influence of boundary uncertainty, the evaluation is performed by semantic labels with eroded object boundaries. Moreover, the OA value is calculated for all classes including clutter/background to make a synthetical comparison.

C. Implementation Details

We adopt FCN [6] and ResNet-101 [42] as our baseline separately for further comparison experiments. All our network frameworks are implemented in PyTorch 1.2 and performed over two Tesla P100 GPUs. Following the previous works [27], [44], [70], we train our network by adopting a poly-learning rate schedule in which the initial learning rate is set to 0.01. After each iteration, the learning rate is multiplied by $1 - (\frac{\mathrm{iter}}{\mathrm{max\_iter}})^{\mathrm{power}}$, with power $= 0.9$. To synchronize the mean and standard deviation of multiple branches, we adopt InPlace-ABNSync [71] after each convolutional layer. Our model is optimized with a stochastic gradient descent for 50 epoches. And we utilize the momentum of 0.9 and the weight decay of 0.0005, respectively. Limited to GPU memory, the training images in both Vaihingen and Potsdam data sets are cropped into 512×512 by applying random cropping, random scaling, and random horizontal flipping. In the test stage, we adopt MS inputs to enhance the testing performance for MS issues in the aerial images.

D. Experiments on Vaihingen Data Set

1) Comparisons With Baseline: In order to evaluate the performance of each component in the proposed network, we carry out many experiments with different settings on the Vaihingen data set, and the numerical results are reported in Table I. Note that we take FCN (VGG-16 and ResNet-101, respectively) as our Baseline-S, which includes semantic-only without height-aware embedding.

TABLE I
ABLATION STUDY OF THE PROPOSED MODULES WITH DIFFERENT BACKBONES ON THE VAIHINGEN TEST SET. BASELINE-S INCLUDES SEMANTIC-ONLY WITHOUT HEIGHT-AWARE EMBEDDING; CAR INDICATES CONTEXT-AWARE REORGANIZATION MODULE; HE INDICATES HEIGHT-AWARE EMBEDDING; MAP INDICATES MODALITY-AFFINITIVE PROPAGATION

Method	**OA(%)**	**mIoU(%)**
Baseline-S (Backbone VGG-16)	86.51	72.69
+ CAR	89.53	79.54
+ CAR + HE	89.68	79.93
+ CAR + HE + MAP	**90.04**	**80.61**
Baseline-S (Backbone ResNet-101)	89.98	79.53
+ CAR	90.77	82.47
+ CAR + HE	90.81	82.55
+ CAR + HE + MAP	**90.94**	**82.80**

As illustrated in Table I, compared with the baseline FCN (VGG-16), employing CAR module achieves a result of 89.53% in OA and 79.54% in mIoU, which brings 3.02% and 6.85% improvement, respectively. Another discovery is that, introducing the geometric information from elevation data gains a larger improvement, which reveals the significance of jointly reasoning 2-D contextual and 3-D geometric information. On the basis of the above, introducing the height-aware embedding (HE) from height estimation branch to the semantic segmentation branch enables a boost over Baseline-S by 3.17%, 7.24% in OA and mIoU, separately. Furthermore, employing the proposed MAP method achieves the best performance, which outbalances the Baseline-S by 3.53% in OA and 7.92% in mIoU, respectively.

We further explore the influence of different backbones to our method. In comparison with the baseline, when we apply a deeper framework (ResNet-101), the network with CAR module, HE, and MAP together increases the OA and mIoU by 0.96% and 3.27%, respectively. Results show that each component we proposed can bring great benefits to scene segmentation.

2) Ablation Study for Context-Aware Reorganization Module: In the proposed HECR-Net, CAR module composed of GCA module and LFR module is employed to distribute the global context to each position conditionally. Table II first reports the results using only four GCA modules, i.e., ASPP [15], nonlocal block (NLB) in [22], pyramid pooling module (PPM) [18], and CGNL in [46] to the Baseline-S. As shown in Table II, all GCA methods show a better performance compared with the Baseline-S. Meanwhile, we report the related results by adding our proposed LFR module to the GCA modules in Table III. Directly exploiting LFR module alone outperforms the baseline by 90.21% in OA, which reveals that features from the dilation backbone have the similar problem as features from GCA modules. Comparing Tables II and III, we can clearly see that the performance of our network is further increased by incorporating LFR module in conjunction with

TABLE II
COMPARISON OF DIFFERENT GLOBAL CONTEXT AGGREGATION (GCA) MODULES USING RESNET-101 AS BACKBONE

Model	OA(%)	mIoU(%)
Baseline-S	89.98	79.53
+ NLB [22]	90.23	81.15
+ PPM [18]	90.41	81.32
+ ASPP [15]	90.53	81.44
+ CGNL [46]	**90.62**	**82.11**

The bold entities indicate the model with best performance.

TABLE III
ABLATION STUDY FOR LOCAL FEATURE REDISTRIBUTION (LFR) MODULE APPLIED ON DIFFERENT GCA MODULES USING RESNET-101 AS BACKBONE

Model	OA(%)	$\Delta\alpha$	$\Delta\beta$
Baseline-S	89.98	-	-
+ LFR	90.21	0.23 ↑	-
+ NLB + LFR	90.46	0.48 ↑	**0.23** ↑
+ PPM + LFR	90.53	0.55 ↑	0.12 ↑
+ ASPP + LFR	90.61	0.63 ↑	0.08 ↑
+ CGNL + LFR	**90.77**	**0.79** ↑	0.15 ↑

$\Delta\alpha$ indicates the results difference comparing with Baseline-S, and $\Delta\beta$ means the results difference between using CAR module (GCA module + LFR module) and the corresponding GCA module.

TABLE IV
ABLATION STUDY FOR DIFFERENT INTEGRATION MANNERS OF GCA AND LFR ADOPTING RESNET-101 AS BACKBONE

Model	OA(%)	mIoU(%)
Baseline-S	89.98	79.53
+ Parallel	90.43	81.35
+ LFR-GCA	90.56	81.86
+ GCA-LFR	**90.77**	**82.47**

different GCA modules, which further proves the effectiveness of the proposed LFR module. Another interesting observation is that LFR module greatly improves the performance of the attention-based aggregation methods, but less for MS aggregation methods, which is mainly attributed to the methods of modeling long-range dependences between them. As shown in Table III, comparing with baseline, “+ CGNL + LFR” achieves the result of 90.77% in OA. Note that, we only select CGNL as our GCA module in the follow-up experiments.

Considering LFR module can also promote the performance of the baseline, this work further studies the impact of different arrangements methods of the GCA module and LFR module. Table IV shows the performance comparisons of the baseline and three different integration manners, where “+ Parallel” denotes concatenating the output features of GCA as well as LFR module and “+ LFR-GCA” indicates performing LFR before performing GCA. As shown in Table IV, “+ GCA-LFR” attains the result, i.e., 90.77% in OA and 82.47% in mIoU, while “+ Parallel” and “+ LFR-GCA” achieve 90.43% and 90.56% in OA and 81.35% and 81.86% in mIoU, respectively. The result shows that concatenating integration patterns and performing LFR before performing GCA lead to an unsatisfactory result. The reason may be that LFR module has not extracted global features yet and the features extracted from LFR module are not insensitive to the regions inside large objects.

TABLE V
COMPARISON WITH ELEVATION INCORPORATION METHODS ON THE VAIHINGEN TEST SET. “SUP” INDICATES SUPERVISION; “D” INDICATES DSM; AND “GT” INDICATES GROUND-TRUTH OF SEMANTIC SEGMENTATION

Model	Input	Sup	OA(%)	mIoU(%)
Baseline-S	IRRG	GT	89.98	79.53
Baseline-SD	IRRG+D	GT	88.87	78.10
Baseline-S-D	IRRG+D	GT	90.48	81.97
Baseline-S+HE	IRRG	GT+D	**90.72**	**82.32**

3) Effect of Height-Aware Embedding: In this part, we mainly investigate the influence of incorporating height information in three different manners on semantic segmentation performance. First, Baseline-SD indicates an image-level fusion method, which directly concatenates the IRRG images and elevation data from the channel dimension. Then, Baseline-S-D is the conventional feature-level fusion method, which feeds the IRRG images and elevation data to two different backbones, separately. The proposed Baseline-S+HE leverages the geometry information for semantic segmentation by learning height-aware embedding without explicitly requiring elevation data as inputs.

The results in Table V show that there is no guarantee of the improvement by introducing additional geographic elevation data. On the contrary, the performance of Baseline-SD is inferior to the performance of Baseline-S. The possible reason is that the pretraining weight is usually based on three-channel data and is not suitable for four-channel data to train. And this simple and direct image-level fusion is not only detrimental to extracting features by backbone, but brings redundant features. By contrast, Baseline-S-D and Baseline-S+HE reveal the effectiveness of inferring 2-D contextual and 3-D geometric information jointly. Especially, Baseline-S+HE yields a considerable amplification of 0.74% and 2.79% in the OA and mIoU, respectively. In summary, the aforementioned experimental results show that the proposed height-aware embedding can effectively distill the geographic information as the auxiliary information and the performance of semantic segmentation is further improved.

We further compare the proposed Baseline-S+HE with Baseline-S, Baseline-SD, and Baseline-S-D in terms of efficiency, including parameters, computation cost (MACs), and execution time. For a fair comparison, we compare the efficiency of the above methods under the same setting, i.e., $output\_stride = 8$. As shown in Table VI, compared with the Baseline-S and Baseline-SD, the proposed Baseline-S+HE increases less computation, memory, and execution time, which can be still tolerable in real scenes. Meanwhile, Baseline-S-D almost doubles the parameters, computation, and execution time with less performance improvement. It is obvious that we have more efficient parameter utilization in the Baseline-S+HE. Fig. 7

TABLE VI
EFFICIENCY COMPARISON WITH ELEVATION INCORPORATION METHODS ON THE VAIHINGEN TEST SET DURING INFERENCE STAGE

Method	Params(M)	MACs(G)	Execution time(s)
Baseline-S	45.149	204.552	5.267
Baseline-SD	45.152	204.758	5.765
Baseline-S-D	101.626	437.341	11.954
Baseline-S+HE	51.146	219.924	6.070

Execution time refers to the time of inferring an image of size [512 × 512].

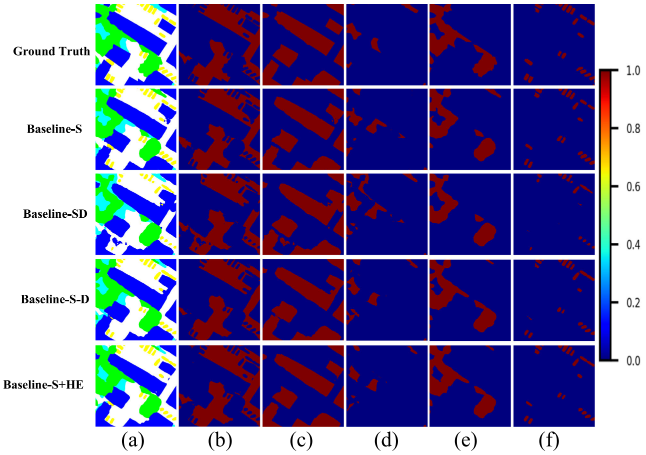


Fig. 7. Feature maps of Ground-truth, Baseline-S, Baseline-SD, Baseline-S-D and Baseline-S+HE. The larger the pixel value, the stronger the response. (a) Predictions for all categories. (b) Impervious surfaces. (c) Building. (d) Low vegetation. (e) Tree. (f) Car. (g) Clutter/background. Greater values of pixels indicate stronger responses.

TABLE VII
COMPARISON WITH DIFFERENT FUSION STRATEGIES ON THE VAIHINGEN TEST SET

Model	OA(%)	mIoU(%)
Baseline-S+HE w/o Feature Fusion	90.31	81.19
Baseline-S+HE w/ SF	90.43	81.52
Baseline-S+HE w/ GAC [72]	90.59	81.87
Baseline-S+HE w/ MAP	**90.72**	**82.32**

"SF" indicates element-wise summation for feature fusion.

qualitatively shows the advantages of the proposed Baseline-S+HE.

4) Effect of MAP Module: In this part, we qualitatively analyze the effect of utilizing the proposed MAP module on the fusion of semantic and geometrical features. As shown in Table VII, we conduct three groups of experiments with different fusion strategies and compare them with the Baseline-S+HE. The results show that all strategies can improve the segmentation performance, where MAP is better than other methods in OA and mIoU especially. This is because that, after obtaining semantic affinity and geometric affinity, our MAP module can effectively integrate the cross-task information to improve segmentation performance.

As shown in Fig. 8, we provide qualitative comparisons between different fusion strategies. First, Baseline-S+HE without feature fusion is easy to lead to incorrect classification between different categories, such as low vegetation and tree, buildings, and impervious surfaces. The three methods with feature fusion can successfully make corrections between the above similar categories in appearance by introducing the geometric information from height estimation branches. In contrast, our proposed MAP module achieves better segmentation performance than the other two fusion methods in the classification of low vegetation and trees categories.

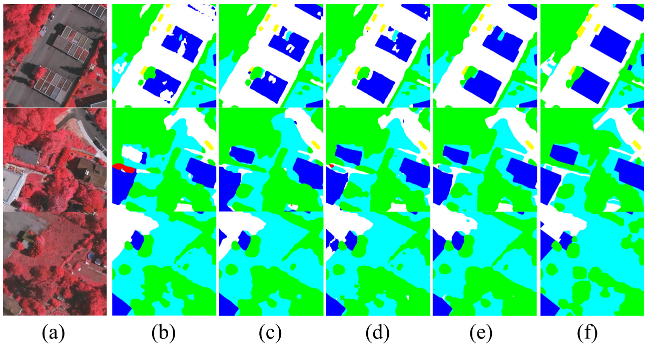


Fig. 8. Qualitative comparisons between different fusion strategies. (a) IRRG images. (b) Without feature fusion. (c) Element-wise summation. (d) GAC. (e) MAP. (f) Ground-truth.

5) Ablation Study for Multitask Objective Function: In this part, we conduce a large number of experiments to verify the effect of the proposed auxiliary loss functions and height estimation module.

Effect of Auxiliary Loss: We explore the optimal value of λ_1, which denotes the weight parameter of the auxiliary loss functions. Concretely, we set the weight parameter of height supervisory loss function (i.e., λ_2) as 1, and the results for the choice of λ_1 are reported in Table IX, which shows that the choice of $\lambda_1 = 0.5$ achieves the best performance. In addition, five experiments are carried out for each parameter value, and the average value was taken as the final result to avoid the effect of the errors in the process of experimental training.

Effect of Height Supervision: This work further analyzes the effect of the height supervisory with different configurations (i.e., λ_2). It is worth noting that none of the models include the MAP module in this section. The results in Table X show that, by incorporating the proposed height-aware embedding branch for the network training, semantic segmentation performance is improved significantly and the model is not particularly sensitive to parameter selection. Particularly, when $\lambda_2 = 0$, the model denotes the baseline network without height estimation and achieves a result of 90.77% in OA and 82.47% in mIoU. Meanwhile, the proposed model with height estimation gets the best performance and outperforms the baseline by in OA and in mIoU, when λ_2 equals 1.

6) Comparison With State-of-the-Art: Following [25], [26], [70], some common strategies are adopted to improve the performance. We first apply random left–right flipping and random scaling for data augmentation (DA). In addition, we average the segmentation results of five scales $\{1.5, 1.25, 1.0, 0.75, 0.5\}$ for MS inference. Meanwhile, we employ a hybrid dilated

TABLE VIII
EXPERIMENTAL RESULTS ON VAIHINGEN TEST SET

Model	**Input**	**Imp. surf.**	**Build.**	**Low veg.**	**Tree**	**Car**	**mean F_1**	**OA**	**mIoU**
FCN [6]	IRRG	88.67	92.83	76.32	86.67	74.21	83.74	86.51	72.69
UZ_1 [69]	IRRG+nDSM	89.20	92.50	81.60	86.90	57.30	81.50	87.30	-
RoteEqNet [29]	IRRG+DSM	89.50	94.80	77.50	86.50	72.60	84.18	87.50	-
S-RA-FCN [12]	IRRG	91.47	94.97	80.63	88.57	87.05	88.54	89.23	79.76
UFMG_4 [73]	-	91.10	94.50	82.90	88.80	81.30	87.72	89.40	-
ONE 7 [74]	IRRG+DSM+NDSM	91.00	94.50	84.40	89.90	77.80	87.50	89.80	-
V-FuseNet [31]	IRRG+DSM	92.00	94.40	84.50	89.90	86.30	89.42	90.00	-
DLR_10 [75]	IRRG+DSM+Edge	92.40	95.20	83.90	89.90	81.20	88.52	90.30	-
TreeUNet [76]	IRRG+DSM	92.50	94.90	83.60	89.60	85.90	89.30	90.40	-
DANet [25]	IRRG	91.63	95.02	83.25	88.87	87.16	89.19	90.44	81.32
DeepLabV3+ [15]	IRRG	92.38	95.17	84.29	89.52	86.47	89.57	90.56	81.47
PSPNet [18]	IRRG	92.79	95.46	84.51	89.94	88.61	90.26	90.85	82.58
BKHN11	IRRG+DSM+nDSM	92.90	**96.00**	84.60	89.90	88.60	90.40	91.00	-
CASIA2 [77]	IRRG	93.20	**96.00**	84.70	89.90	86.70	90.10	91.10	-
GANet [72]	IRRG	93.10	95.50	84.90	90.20	87.40	90.20	91.10	-
HECR-Net	IRRG	**93.64**	95.53	**85.78**	**90.37**	**89.09**	**90.89**	**91.45**	**83.47**

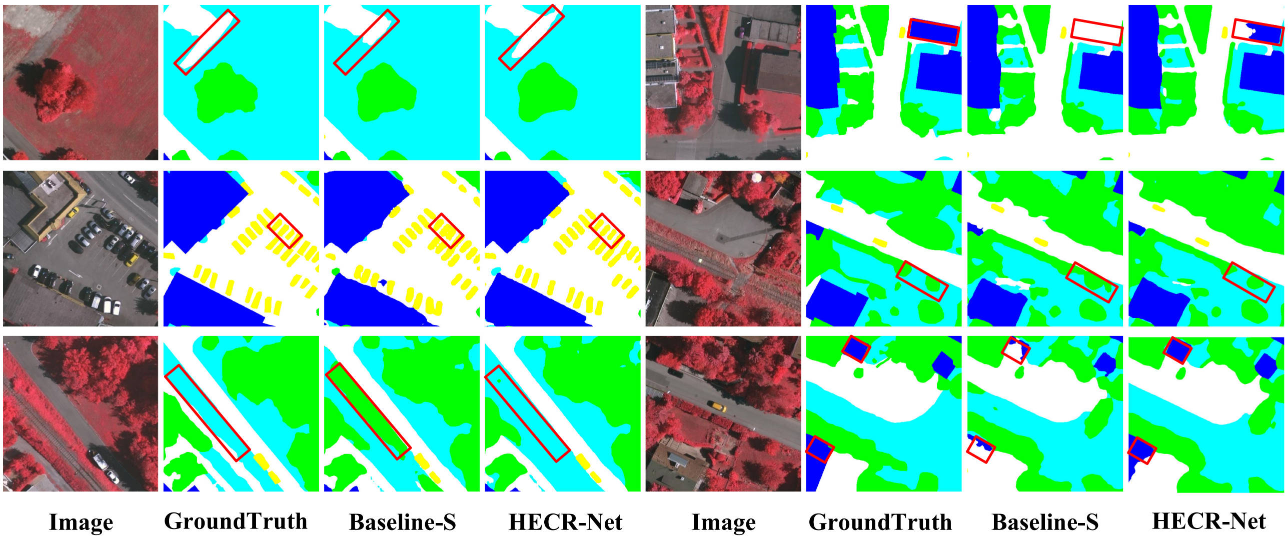


Fig. 9. Visualization results of HECR-Net on Vaihingen test set. Impervious surfaces: white; buildings: blue; low vegetation: cyan; trees: green; cars: yellow.

TABLE IX
COMPARISONS OF DIFFERENT WEIGHT PARAMETERS λ_1

λ_1	0	0.3	0.4	**0.5**	0.6	0.7
OA(%)	90.63	90.81	90.89	**90.94**	90.86	90.79
mIoU(%)	82.15	82.48	82.68	**82.80**	82.61	82.53

TABLE X
EXPERIMENTAL RESULTS OF DIFFERENT WEIGHT PARAMETERS λ_2

λ_2	0	0.5	**1**	1.5	2
OA(%)	90.77	90.83	**90.94**	90.69	90.51
mIoU(%)	82.47	82.56	**82.80**	82.33	82.24

convolution bottleneck with multigrid (MG) structure. As shown in Table XII, all the above strategies improve the performance significantly. When we adopt all the strategies, the proposed HECR-Net improves the segmentation performance by almost 0.51% in OA and 0.67% in mIoU.

We compare our model with the current methods on Vaihingen test set, and the numerical results are reported in Table VIII, where our approach still maintains the highest mean $\boldsymbol{F_1}$, OA, and mIoU. In addition, we also report the segmentation results on each category in Table VIII. It is obvious that our method is superior to other methods in most classes, which confirms the effectiveness of our proposed modules again.

7) Visualization Results: Fig. 9 showcases several qualitative results of our model and baseline on the ISPRS Vaihingen data

TABLE XI
EXPERIMENTAL RESULTS ON POTSDAM TEST SET

Model	Input	Imp. surf.	Build.	Low veg.	Tree	Car	mean F_1	OA	mIoU
FCN [6]	RGB	88.61	93.29	83.29	79.83	93.02	87.61	85.59	78.34
UZ_1 [69]	IRRG+nDSM	89.30	95.40	81.80	80.50	86.50	86.70	85.80	-
S-RA-FCN [12]	RGB	91.33	94.70	86.81	83.47	94.52	90.17	88.59	82.38
V-FuseNet [31]	IRRGB+DSM+nDSM	92.70	96.30	87.30	88.50	95.40	92.04	90.60	-
TSMTA [78]	RGB	92.91	97.13	87.03	87.26	95.16	91.90	90.64	-
TreeUNet [76]	IRRGB+DSM+nDSM	93.10	97.30	86.60	87.10	95.80	91.98	90.70	-
DeepLabV3+ [15]	RGB	92.95	95.88	87.62	88.15	96.02	92.12	90.88	84.32
CASIA2 [77]	IRRGB	93.40	96.80	87.60	88.30	96.10	92.44	91.00	-
PSPNet [18]	RGB	93.36	96.97	87.75	88.50	95.42	92.40	91.08	84.88
BKHN_3	IRRGB+DSM+nDSM	93.30	97.20	88.00	88.50	96.00	92.60	91.10	-
GANet [72]	IRRG	92.80	96.70	87.30	88.50	96.10	92.60	91.10	-
CCNet [27]	RGB	93.58	96.77	86.87	88.59	**96.24**	92.41	91.47	85.65
HUSTW4 [79]	IRRG+DSM+nDSM	93.60	**97.60**	88.50	88.80	94.60	92.62	91.60	-
SWJ_2	IRRG	**94.40**	97.40	87.80	87.60	94.70	92.38	91.70	-
HECR-Net	RGB	93.75	97.44	**88.68**	**89.19**	95.40	**92.89**	**91.84**	**87.08**

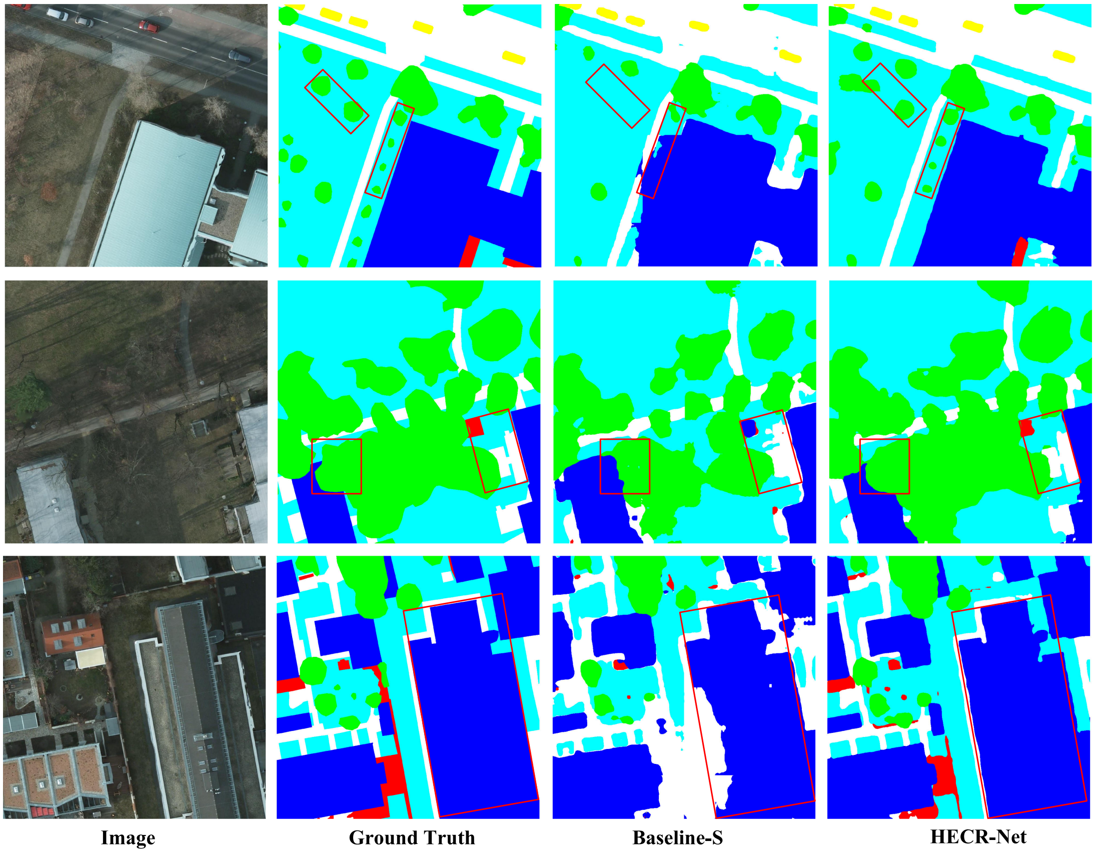


Fig. 10. Visualization results of HECR-Net on Potsdam test set. Impervious surfaces: white; buildings: blue; low vegetation: cyan; trees: green; cars: yellow.

sets for semantic segmentation. In order to show the results of segmentation more clearly, we choose 512 × 512 patches. The results show that the segmentation result is obviously better than that of baseline network in the regions with similar color or shadows marked with red solid box. The proposed HECR-Net predicts more accurate segmentation maps, which commendably demonstrates the effectiveness of inferring 2-D contextual and 3-D geometric information jointly.

E. Experiments on Potsdam Data Set

In order to evaluate the effectiveness of HECR-Net, we also carry out experiments on the large-scale ISPRS Potsdam data

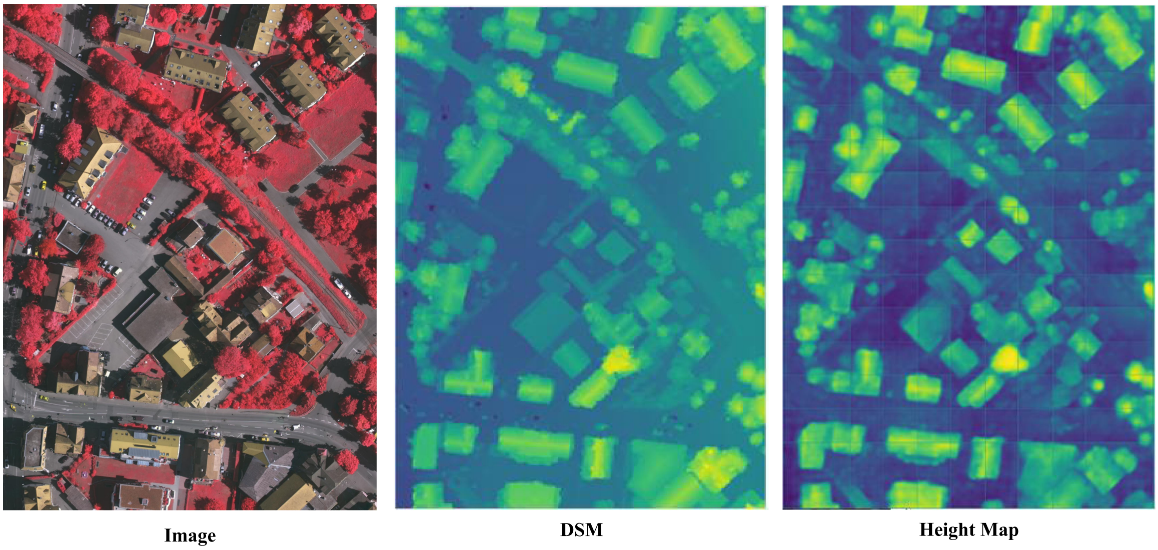


Fig. 11. Visualization results of height estimation on ISPRS Vaihingen data set.

TABLE XII
COMPARISONS AMONG DATA AUGMENTATION (DA), MULTISCALE (MS), AND MULTIGRID (MG)

Method	DA	MS	MG	OA(%)	mIoU(%)
HECR-Net				90.94	82.80
HECR-Net	✓			91.15	83.08
HECR-Net	✓	✓		91.26	83.32
HECR-Net	✓	✓	✓	**91.45**	**83.47**

set. We adopt the same experimental setup as on Vaihingen data set. The results compared with state-of-the-arts are reported in Table XI. Note that the compared methods may vary from Table VIII to Table XI. Remarkably, the proposed HECR-Net (ResNet-101) achieves 91.84% in OA and 87.08% in mIoU. The experimental results on Potsdam data set again verify the effectiveness of HECR-Net, with better performance than the compared methods. Notably, there are some low-quality DSMs in Potsdam data set as a result of the capture device, which probably influences the auxiliary effect from the height-aware embeddings.

In addition, Fig. 10 demonstrates some qualitative classification results of our HERC-Net and Baseline-S on Potsdam test set. As shown in Fig. 10, HECR-Net produces desirable segmentation results on distinguishing the objects with similar 2-D appearance (e.g., roads and roofs, trees and low vegetations).

F. Height Estimation Performance

In this article, we focus on jointly extracting 2-D contextual and 3-D geometric features from a single optical image, and fusing the above features to solve the challenges in semantic segmentation. Here we qualitatively analyze the ability of the proposed network to distill 3-D height features.

Fig. 11 shows the qualitative height estimation results of the proposed HECR-Net to demonstrate the capability of learning 3-D geometrical features on the ISPRS Vaihingen test set. We choose DSM as the supervision for height estimation to get height-aware embeddings. Note that, we utilize a sigmoid function to normalize the predicted values of height maps to $[0, 1]$. In addition, because the aerial images must be cropped into smaller images for memory constraints, our HECR-Net produces a surface with gap after integrating the estimated height images.

V. CONCLUSION

In this article, we present a joint reasoning network for dense prediction tasks in the complex scenes, namely HECR-Net. The biggest innovation of the proposed HECR-Net is to decouple the single prediction task into semantic segmentation and height estimation. Furthermore, we introduce a CAR module embedded with a GCA module and a local feature redistribution module. This CAR module is specifically responsible for generating a discriminative feature with global information appropriately assigned to each local position. Additionally, to extract 3-D height information, a multitask decoder is trained under the supervision of semantic labels and elevation data (e.g., DSM and nDSM). What is more, modality-affinitive propagation block fuses the distilled 3-D height features and 2-D contextual features to improve the performance of semantic segmentation. A large number of experiments on ISPRS Vaihingen and Potsdam data set certify that the proposed HECR-Net achieves remarkable performance. In the future work, we will study the optimization of multitask joint training and further study the effectiveness of the CAR module in the task of height estimation, where both global and local information are important.

REFERENCES

[1] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, 2018.

[2] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[3] Z. Lu, X. Jiang, and A. Kot, "Deep coupled ResNet for low-resolution face recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 526–530, Apr. 2018.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[5] Z. Liu, G. Lin, S. Yang, J. Feng, W. Lin, and W. L. Goh, "Learning Markov clustering networks for scene text detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6936–6944.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[7] B. Shuai, Z. Zuo, B. Wang, and G. Wang, "Scene segmentation with DAG-recurrent neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1480–1493, Jun. 2018.

[8] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1925–1934.

[9] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.

[10] 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 Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

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

[12] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7557–7569, Nov. 2020.

[13] C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. Int. Conf. Learn. Representations*, 2015.

[14] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017, *arXiv:1706.05587*.

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

[16] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters-improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4353–4361.

[17] W.-C. Hung *et al.*, "Scene parsing with global context embedding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2631–2639.

[18] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[19] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1857–1866.

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

[21] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, 2021.

[22] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[23] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, BMVA, Press, 2018, p. 285.

[24] H. Zhao *et al.*, "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 267–283.

[25] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[26] Y. Yuan and J. Wang, "OCNet: Object context network for scene parsing," *CoRR*, vol. abs/1809.00916, 2018, *arXiv:1809.00916*.

[27] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 603–612.

[28] R. Qin and W. Fang, "A hierarchical building detection method for very high resolution remotely sensed images combined with DSM using graph cut optimization," *Photogrammetric Eng. Remote Sens.*, vol. 80, no. 9, pp. 873–883, 2014.

[29] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 96–107, 2018.

[30] L. Mou and X. X. Zhu, "RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images," *CoRR*, vol. abs/1805.02091, 2018, *arXiv:1805.02091*.

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

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

[33] Z. Cao *et al.*, "End-to-end DSM fusion networks for semantic segmentation in high-resolution aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1766–1770, Nov. 2019, doi: 10.1109/LGRS.2019.2907009.

[34] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014, pp. 2366–2374.

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

[36] L. Mou and X. X. Zhu, "IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," *CoRR*, vol. abs/1802.10249, 2018.

[37] P. Ghamisi and N. Yokoya, "IMG2DSM: Height simulation from single imagery using conditional generative adversarial net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 794–798, May 2018.

[38] G. Cheng, Z. Li, J. Han, X. Yao, and L. Guo, "Exploring hierarchical convolutional features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6712–6722, Nov. 2018.

[39] P. Zhou, J. Han, G. Cheng, and B. Zhang, "Learning compact and discriminative stacked autoencoder for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4823–4833, Jul. 2019.

[40] P. Tokarczyk, J. D. Wegner, S. Walk, and K. Schindler, "Features, color spaces, and boosting: New insights on semantic classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 280–295, Jan. 2015.

[41] M. D. Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[43] H. Liu *et al.*, "An end-to-end network for panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6172–6181.

[44] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[45] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3684–3692.

[46] K. Yue, M. Sun, Y. Yuan, F. Zhou, E. Ding, and F. Xu, "Compact generalized non-local network," *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 2018, pp. 6511–6520.

[47] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 681–687.

[48] D.-K. Kim, D. Maturana, M. Uenoyama, and S. Scherer, "Season-invariant semantic segmentation with a deep multimodal network," in *Field and Service Robotics*. Berlin, Germany: Springer, 2018, pp. 255–270.

[49] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, and L. Lin, "LSTM-CF: Unifying context modeling and fusion with LSTMs for RGB-D scene labeling," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 541–557.

[50] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 213–228.

[51] A. Valada, R. Mohan, and W. Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *Int. J. Comput. Vis.*, vol. 128, pp. 1–47, 2019.

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

[53] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 740–756.

[54] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6647–6655.

[55] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5684–5693.

[56] B. Li, C. Shen, Y. Dai, A. V. D. Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1119–1127.

[57] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.

[58] A. Roy and S. Todorovic, "Monocular depth estimation using neural regression forest," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2016, pp. 5506–5514.

[59] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2800–2809.

[60] Z. Chen, Q. Qin, L. Lin, Q. Liu, and W. Zhan, "DEM densification using perspective shape from shading through multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 1, pp. 145–149, Jan. 2013.

[61] M. A. Rajabi and J. R. Blais, "Optimization of DTM interpolation using SFS with single satellite imagery," *J. Supercomput.*, vol. 28, no. 2, pp. 193–213, 2004.

[62] S. Srivastava, M. Volpi, and D. Tuia, "Joint height estimation and semantic labeling of monocular aerial images with CNNs," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 5173–5176.

[63] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[64] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, "CARAFE: Content-aware reassembly of features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3007–3016.

[65] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.

[66] "ISPRS.2D semantic labeling contest-Vaihingen," [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html

[67] "ISPRS.2D semantic labeling contest-Potsdam," [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

[68] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[69] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 881–893, Feb. 2017.

[70] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9167–9176.

[71] S. Rota Bulò, L. Porzi, and P. Kontschieder, "In-place activated batchnorm for memory-optimized training of DNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5639–5647.

[72] X. Li, C. Wen, L. Wang, and Y. Fang, "Geometry-aware segmentation of remote sensing images via joint height estimation," *IEEE Geoscience and Remote Sensing Letters*, IEEE, 2021.

[73] K. Nogueira, M. D. Mura, J. Chanussot, W. R. Schwartz, and J. A. dos Santos, "Dynamic multicontext segmentation of remote sensing images based on convolutional networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7503–7520, Oct. 2019.

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

[75] 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 J. Photogrammetry Remote Sens.*, vol. 135, pp. 158–172, 2018.

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

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

[78] L. Ding, J. Zhang, and L. Bruzzone, "Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5367–5376, Aug. 2020.

[79] Y. Sun, Y. Tian, and Y. Xu, "Problems of encoder–decoder frameworks for high-resolution remote sensing image segmentation: Structural stereotype and insufficient learning," *Neurocomputing*, vol. 330, pp. 297–304, 2019.

Wenjie Liu (Graduate Student Member, IEEE) received the B.Sc. degree in communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2018. He is currently working toward the Ph.D. degree in signal and information processing with the University of Chinese Academy of Sciences, Beijing, China, and the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing.

His research interests include computer vision and remote sensing image processing, especially on RGB-D semantic segmentation.

Wenkai Zhang (Member, IEEE) received the B.Sc. degree in electronic information engineering from the China University of Petroleum, Qingdao, China, in 2013, and the M.Sc. and Ph.D. degrees in Electronic Information Engineering from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2018.

He is currently an Assistant Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision and remote sensing image analysis.

Xian Sun received the B.Sc. degree in electronic information engineering from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees in Electronic Information Engineering from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2009.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

Zhi Guo received the B.Sc. degree in electrical engineering automation from Tsinghua University, Beijing, China, in 1998, and the M.Sc. and Ph.D. degrees in electronic information engineering from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2003.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

Kun Fu received the B.Sc., M.Sc., and Ph.D. degrees in electronic information engineering from the National University of Defense Technology, Changsha, China, in 1995, 1999, and 2002, respectively.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, remote sensing image understanding, geospatial data mining, and visualization.