Attention-Guided Label Refinement Network for Semantic Segmentation of Very High Resolution Aerial Orthoimages

Jianfeng Huang, Xinchang Zhang, Ying Sun, and Qinchuan Xin, *Member, IEEE*

***Abstract*—The recent applications of fully convolutional networks (FCNs) have shown to improve the semantic segmentation of very high resolution (VHR) remote-sensing images because of the excellent feature representation and end-to-end pixel labeling capabilities. While many FCN-based methods concatenate features from multilevel encoding stages to refine the coarse labeling results, the semantic gap between features of different levels and the selection of representative features are often overlooked, leading to the generation of redundant information and unexpected classification results. In this article, we propose an attention-guided label refinement network (ALRNet) for improved semantic labeling of VHR images. ALRNet follows the paradigm of the encoder–decoder architecture, which progressively refines the coarse labeling maps of different scales by using the channelwise attention mechanism. A novel attention-guided feature fusion module based on the squeeze-and-excitation module is designed to fuse higher level and lower level features. In this way, the semantic gaps among features of different levels are declined, and the category discrimination of each pixel in the lower level features is strengthened, which is helpful for subsequent label refinement. ALRNet is tested on three public datasets, including two ISRPS 2-D labeling datasets and the Wuhan University aerial building dataset. Results demonstrated that ALRNet had shown promising segmentation performance in comparison with state-of-the-art deep learning networks. The source code of ALRNet is made publicly available for further studies.**

Manuscript received November 29, 2020; revised February 21, 2021; accepted April 2, 2021. Date of publication April 22, 2021; date of current version May 12, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0604300, Grant 2017YFA0604400, and Grant 2018YFB2100702; in part by the National Natural Science Foundation of China under Grant 42071441, Grant 41801351, and Grant 41875122; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2020A1515110441, and in part by the Innovation Group Project of Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) under Grant 311020008. *(Corresponding author: Ying Sun.)*

Jianfeng Huang is with the Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), School of Atmospheric Sciences, Sun Yat-sen University, Zhuhai 519082, China, and also with the Guangdong Province Key Laboratory for Climate Change and Natural Disaster Studies, School of Atmospheric Sciences, Sun Yat-sen University, Guangzhou 510275, China (e-mail: huangjf9@mail3.sysu.edu.cn).

Xinchang Zhang is with the School of Geography and Remote Sensing, Guangzhou University, Guangzhou 510006, China, and also with the College of Environment and Planning, Henan University, Kaifeng 475004, China (e-mail: eeszxc@mail.sysu.edu.cn).

Ying Sun and Qinchuan Xin are with the Guangdong Key Laboratory for Urbanization and Geo-simulation and the School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China (e-mail: sunying23@mail.sysu.edu.cn; xinqinchuan@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2021.3073935

***Index Terms*—Attention mechanism, convolutional neural networks (CNNs), deep learning, semantic segmentation, urban object extraction, very high spatial resolution images.**

I. Introduction

THE development of remote-sensing technologies for the earth observation has significantly increased the accessibility to very high spatial resolution (VHR) images [1], which opens up new horizons for a better understanding of our changing world. Semantic segmentation that assigns a semantic label to each pixel in an image is one of the fundamental approaches to analyze remote-sensing data [2], and plays an essential role in diverse applications, such as land cover/land use interpretation [3], disaster analysis, urban planning [4], and environment monitoring. Developing automatic and reliable algorithms of semantic segmentation is now a research frontier in the field of remote sensing [5].

Many efforts have been made in the past few decades to develop accurate semantic segmentation methods, including machine-learning-based methods [6], [7] and object-based analysis methods [8]. Nevertheless, accurate semantic labeling of VHR images is challenging for reasons. On the one hand, the high intraclass and low interclass spectral variation of complicated urban areas in the VHR images make it difficult to extract representative features of target objects [9]. On the other hand, many methods depend on designing hand-crafted features [10], whereas the hand-crafted features are usually low/mid-level features and are often unreliable to distinguish objects in complicated circumstances [11].

Deep convolutional neural networks (DCNNs) have recently shown remarkable learning ability in processing and analyzing VHR images [12], such as scene classification [13], urban object detection [14], and semantic segmentation [15], [16]. DCNNs can automatically learn rich contextual features and high-level semantic features from the input images without prior knowledge [17]. Because the downsampling operations in DCNNs enlarge the receptive fields of deeper layers [18], DCNNs often generate low spatial resolution feature maps and do not meet the demands to obtain full-resolution labeling results. To overcome this problem, Long *et al.* [19] proposed fully convolutional networks (FCNs) that can perform pixelwise semantic labeling via an end-to-end learning fashion. FCNs have an encoder–decoder architecture, of which the encoder is used to learn multilevel features and the decoder is used to obtain dense prediction results [20]. Thus, FCNs have become increasingly important for many state-of-the-art semantic segmentation algorithms [21]. One of the significant weaknesses of FCNs is that the output results are often "blobby" because many spatial-related features generated in the shallow layers of DCNNs are discarded [22]. Consequently, it is now an open research topic to develop advanced FCN-models for semantic labeling of VHR images.

There are two key issues to be properly addressed when performing accurate labeling with FCNs. First, lower level features of DCNNs that contain rich spatial information are often utilized to refine the coarse results [23]. Many methods transfer these features to the decoder via skip-connection [24] or up-pooling [25], and then apply the multiscale feature stacking [18]. However, the semantic gaps among different level features are often overlooked [26]. As the receptive field sizes and abstraction degrees of multilevel features are different, higher level features tend to have stronger category discriminability than lower level features [20]. Directly stacking the features of different levels may not achieve the desired fusion outcome or even decrease the accuracy of FCNs [27].

Second, feature selection is often ignored when lower level features are transferred to the decoder, leading to the generation of redundant information [28] and increased computational complexity of FCNs. The transferred features usually contain category-ambiguity and nonboundary-related information that is not helpful for label refinement [20], [29]. To tackle this problem, Islam *et al.* [20] developed a gated feedback refinement network (G-FRNet) that chose a certain number of features from both higher level and lower level encoding stages and fused them for further label refinement. The purpose of their work is to select representative features from each encoding stage with the guidance of higher level features. Increasing studies have used high-level features to guide the selection of lower level features for improved category discriminability [29]–[31]. However, most methods to date only consider the features of the same level to be equally important, which may increase the difficulty of feature selection from such a large number of features.

According to the reasons as mentioned earlier, it is potentially beneficial to improve FCNs by eliminating the semantic gaps among different level features and selecting the most representative features for network delivering. Toward this goal, we focus on applying the visual attention mechanism [32] to the semantic segmentation tasks. The visual attention mechanism aims to learn how to extract salient spatial locations and important feature channels from images by imitating the way of human observation [33], [34]. There are developed attention modules that can be seamlessly integrated with DCNNs [27]. Notably, the squeeze-and-excitation (SE) module proposed by Hu *et al.* [35] allows DCNNs to perform the channelwise attention that can automatically emphasize essential features and suppress less important ones. The SE module motivates us to use the attention mechanism for selecting representative features and declining the semantic gap when fusing multilevel features. At present, integrating the attention mechanisms and FCNs has attracted the interests of many studies [27], [30], [33], [36]. However, few studies have used the attention mechanisms for declining the semantic gaps among different level features and taking advantage of the reweighted features for further label refinement.

In this study, we propose a novel attention-guided label refinement network (ALRNet) to advance the state-of-the-art on semantic segmentation of VHR aerial images. The proposed network follows the paradigm of encoder–decoder architecture, which progressively upsamples and refines the coarse labeling maps of different scales. A new attention-guided feature fusion (AGFF) module is carefully designed to decline the semantic gap between different level features. Specifically, it transforms the higher level features into a weighted vector via the SE module. The obtained weighted vector is used to guide the channelwise recalibrating of lower level features. The AGFF module not only strengthens the category discriminability of each pixel in the lower level features but also reweights the features that are helpful for label refinement. Additionally, rather than transferring all the reweighted features to the decoder, only a small part of these features is selected and leveraged to refine the upsampled labeling maps. We recursively apply the AGFF module across different encoding stages in ALRNet and finally obtain full-resolution classification results. Experiments on three benchmark datasets were conducted to verify the performance of our proposed network.

II. Related Work

A. Semantic Segmentation of VHR Images

Semantic segmentation of the remote-sensing images has been intensively studied in the past few decades [37]. Prior studies can be roughly divided into the pixel-based and object-based classification methods [38], depending on the analysis unit. Machine-learning-based methods and object-based analysis methods [8] are important techniques widely used in previous studies [16]. With the development of the deep learning technology, many researchers have attempted to use deep learning for semantic segmentation. Early studies have used DCNNs for pixelwise classification [39], in which each pixel is classified with the category of its enclosing subpatch. Recently, the FCN-based methods have been applied for the semantic segmentation of VHR images. FCNs can obtain full-resolution classification results via an end-to-end learning framework [19]. Hence, they have been widely used in land cover classification and urban object extraction [17], [40]. It is now common to validate the newly developed FCNs methods based on the international society for photogrammetry and remote-sensing (ISPRS) semantic labeling benchmarks [41]. These methods can be roughly classified into the image-based methods [28] and the data fusion-based methods [42], [43]. The data fusion-based methods use the elevation data (e.g., nDSM) as supplementary elevation information to the 2-D images [44]. By comparison, the image-based methods that only use the 2-D images for semantic segmentation face an enormous challenge in extracting robust contextual and semantic features from images. The objective of this study is to develop a new image-based FCN model to improve the semantic segmentation of the VHR images.

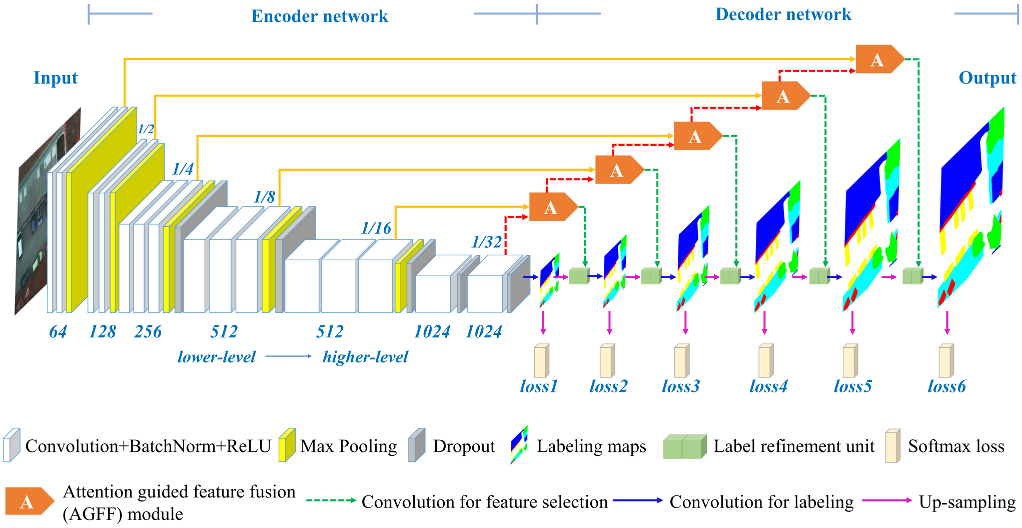

Fig. 1. Overview of the proposed ALRNet. The encoder is based on the modified VGG-16, and the decoder progressively enlarges and refines the coarse labeling maps to obtain the final full-resolution output. The AGFF module fuses higher level and lower level features with the guidance of the channelwise SE module. The label refinement unit produces higher level labeling maps by combining the upsampled labeling maps with the selected features from the AGFF module. A multiscale supervision schema is used to penalize the poor classifications of the labeling maps on each scale.

B. Encoder–Decoder Network Architecture

Most of the state-of-the-art FCN-based methods follow the paradigm of an encoder–decoder architecture, e.g., SegNet [25] and U-Net [45]. The output of FCNs often suffers the loss of spatial details because of the downsampling operations performed in the encoder. To tackle this problem, some studies transfer lower level features to the decoder by skip-connections to refine the coarse prediction results [45]. Some other works aggregate multiscale contextual features that generated by dilated convolution operations [21]. Multilevel feature fusion is a crucial step in FCN-based methods [18]. However, the semantic gaps between different levels of encoder features and between encoder and decoder features are often neglected [26], leading to the generation of redundant information and unexpected results. Our proposed model is inspired by LRN [24] and G-FRNet [20], which progressively refines the coarse labeling maps of different scales. G-FRNet adopts a gate feature fusion strategy to enhance the category discriminability of lower level features. Different from their works, we further reduce the aforementioned semantic gaps among features of different levels in the fusion stage by leveraging the channelwise attention mechanism.

C. Attention Mechanism

Attention mechanisms [32] have shown efficiency in many computer vision applications, such as image classification [34], [35], object detection [35], [46], and semantic segmentation [33], [36]. For the tasks of semantic segmentation, Oktay *et al.* [30] integrated an attention gate model (AG) into the standard U-Net architecture to increase model sensitivity and prediction capability. Fu *et al.* [27] proposed a dual attention network for scene segmentation, of which both a positionwise and a channelwise attention module are appended on top of the dilated FCNs. Motivated by the channelwise SE module [35], Roy *et al.* [47] proposed a spatialwise SE (SSE) module and proved that the integration of SE modules within FCNs yields an improvement for image segmentation. Recent studies have used attention mechanisms for improving the semantic segmentation of VHR images [48], [49]. Luo *et al.* [43] proposed a deep FCN model for semantic labeling of VHR images, where the SE module is used for reweighting single-level features. Panboonyuen *et al.* [50] introduced a channelwise attention module to the global convolutional network for selecting the most discriminative features. Pan *et al.* [51] developed a generative adversarial network (GAN) with channelwise and spatialwise attention modules (GAN-SCA) for building extraction. Different from the aforementioned studies, we try to conduct channelwise attention across different feature fusion stages and carefully design an AGFF module for narrowing the semantic gaps between higher level and lower level features.

III. Attention-Guided Label Refinement Network

A. Overview of the Network

The proposed ALRNet consists of an encoder network and a decoder network (see Fig. 1). The encoder is based on the modified VGG-16 [52], which is stacked by multiple convolutional blocks and max-pooling layers. Each block has a 3 × 3 kernel-size convolutional layer, a batch normalization layer, and a rectified linear unit (ReLU) layer. We replace the Softmax and fully connected layers of the original VGG-16 with two additional convolutional blocks, of which each generates the feature maps with 1024 channels. The layers in the encoder can be grouped into different encoding stages, according to the increasing feature channels and decreasing spatial dimensions. To avoid network overfitting, we add a dropout layer on top of the third to the seventh encoding stages, respectively. As shown in Fig. 1, the encoder network accepts three-band input images and generates feature maps with 1/32 size of the original spatial dimensions.

The output feature maps (with 1024 channels) of the encoder network are passed to the decoder and are, then, convolved into

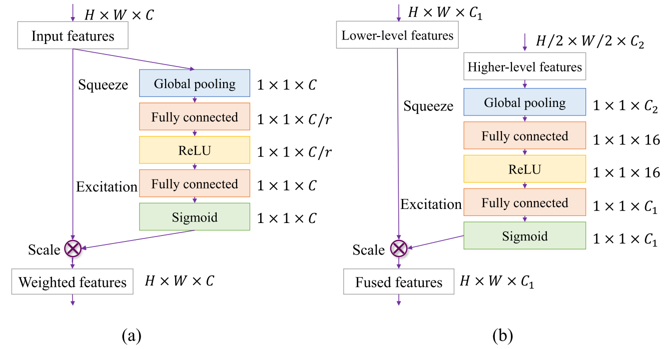

Fig. 2. Schema of (a) the SE module and (b) the proposed AGFF module.

coarse labeling maps with the same number of the classification categories. The decoder progressively enlarges the coarse labeling maps using a standard 2 × upsampling operation and a convolutional operation for five times to obtain full-resolution prediction results (see Fig. 1).

As the lower level features are discarded, the details of the full-resolution prediction maps are still coarse with "blobby" objects [22]. Here, we apply three model components to transfer the representative features from the encoder to refine the coarse prediction maps. First, an AGFF module based on the SE module [35] is proposed to fuse features from different encoding stages. AGFF is recursively applied to each encoding stage to reduce the category ambiguity of lower level features and reweight the features for subsequent selection. Second, a label refinement unit [20] is used repeatedly in the decoder (see Fig. 1). This unit produces higher level labeling maps with larger spatial dimensions by combining the upsampled coarse labeling maps and the convolved features obtained from the AGFF module. The detailed design of the AGFF module and its collaboration with the label refinement unit are demonstrated in Section III-C. Finally, a multiscale supervision scheme [53] is utilized to penalize the poor classifications of the coarse labeling maps on each scale. We first upsample each coarse labeling map to full resolution and then calculate the pixelwise cross-entropy loss between the upsampled labeling maps and the ground truth images. The detailed description of the deep supervision schema is presented in Section III-D. Our proposed ALRNet is an end-to-end trainable network and does not require any sophisticated postprocessing.

B. SE Module

To allow a better understanding of the AGFF module, here we briefly introduce the basic concept of the SE module proposed by Hu *et al.* [35]. The SE module is a channelwise attention module that aims to improve the representational capability of a network. It allows DCNNs to emphasize essential features and suppress less important ones by explicitly modeling the interdependencies of features. Fig. 2(a) illustrates an example of the SE module transforms input features into weighted features. In general, the SE module can be embedded into many DCNNs via three steps of 1) squeeze, 2) excitation, and 3) rescaling as follows.

Let $\mathbf{l} \in \mathbb{R}^{H\times W\times C}$ denote the input features of SE module, H, W, and C denote the height, width, and number of channels of the features, respectively. The squeeze operation is first applied to shrink $\mathbf{l}$ into a channelwise statistic $\mathbf{q} \in \mathbb{R}^{C}$, using the global average pooling operation. The cth element of $\mathbf{q}$ is computed by

$$\mathbf{q}_c = \mathbf{F}_{\text{sq}}\left(\mathbf{l}_c\right) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{l}_c(i,j) \tag{1}$$

where $\mathbf{l}_c$ is cth channel of $\mathbf{l}$. The global pooling operation helps to aggregate and exploit the contextual information of each channel.

The excitation operation is conducted to fully capture the channelwise interdependencies of the features. It transforms the channelwise statistic $\mathbf{q}$ into a weighted vector via a lightweight gate mechanism

$$\mathbf{g} = \mathbf{F}_{\text{ex}}(\mathbf{q},\mathbf{W}) = \sigma\left(\mathbf{W}_2\delta\left(\mathbf{W}_1\mathbf{q}\right)\right) \tag{2}$$

where δ and σ denote the ReLU and sigmoid activation function, respectively; $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ refer to two fully connected layers around the ReLU layer; $\mathbf{r}$ denotes the dimensionality-reduction ratio; $\mathbf{g}$ is the obtained weighted vector with the same number of channels as $\mathbf{l}$. As shown in Fig. 2(a), the number of channels of statistic $\mathbf{q}$ is first reduced to C/r and then scaled up to the original dimension (i.e., C).

Finally, the rescaling operation (i.e., channelwise multiplication) is used to rescale the input features $\mathbf{l}$ with vector $\mathbf{g}$

$$\mathbf{v}_c = \mathbf{F}_{sc}\left(\mathbf{l}_c,\mathbf{g}_c\right) = \mathbf{g}_c\mathbf{l}_c \tag{3}$$

where $\mathbf{v} = [\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_c]$ refers to the weighted features.

Some studies [43], [51] have used the SE module for feature reweighting on single-level features. Different from their works, here we tried to introduce the higher level features into the squeeze and excitation process. The underlying idea is that the weighted vector obtained from the higher level features is more category discriminative than from the lower level ones and, thus, is helpful to guide the reweighting of the lower level features.

C. AGFF Module

The proposed AGFF module is developed based on the SE module to narrow the semantic gaps among different level features and select the most representative information for further label refinement. An example that illustrates the structure of the AGFF module is shown in Fig. 2(b). Rather than performing feature recalibration by self-attention on single-level features, the AGFF module conducts the channelwise attention among features of different levels. It is designed for reducing the semantic gaps between higher level and lower level features and further facilitating feature selection.

Let $\mathbf{l} \in \mathbb{R}^{H\times W\times C_1}$ and $\mathbf{u} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C_2}$ be the lower level and higher level encoding features, respectively. We first transform $\mathbf{u}$ into a weighted vector following the SE steps of the SE module:

$$\mathbf{q}'_c = \mathbf{F}_{\text{sq}}\left(\mathbf{u}_c\right) = \frac{1}{H/2\times W/2}\sum_{i=1}^{H/2}\sum_{j=1}^{W/2}\mathbf{u}_c(i,j) \tag{4}$$

$$\mathbf{g}' = \mathbf{F}_{\text{ex}}\left(\mathbf{q}',\mathbf{W}'\right) = \sigma\left(\mathbf{W}'_2\delta\left(\mathbf{W}'_1\mathbf{q}'\right)\right) \tag{5}$$

where $\mathbf{g}'$ denotes the weighted vector transformed from $\mathbf{u}$; $\mathbf{W}'_1 \in \mathbb{R}^{16\times C_1}$ and $\mathbf{W}'_2 \in \mathbb{R}^{C_1\times 16}$ refer to two fully connected

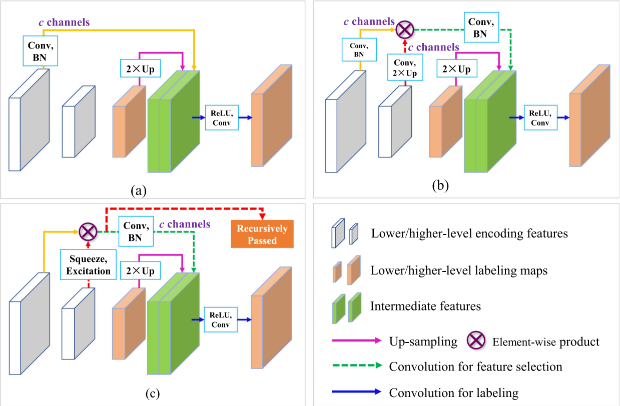

Fig. 3. Schemes are illustrated for different feature fusion modules collaborated with the label refinement unit. Note that the weighted features obtained from the AGFF module are recursively used in the next fusion stage and acted as the higher level inputs. In contrast, the GFF module is separately used in each fusion stage without recursion, and only two adjacent level encoding features are involved. (a) Direct FF. (b) GFF. (c) AGFF.

layers. Note that the obtained weighted vector $\mathbf{g}'$ has the same channel dimension as $\mathbf{l}$. The final output of the AGFF module is calculated by rescaling the lower level features $\mathbf{l}$ with $\mathbf{g}'$

$$\mathbf{v}'_c = \mathbf{F}_{sc}\left(\mathbf{l}_c, \mathbf{g}'_c\right) = \mathbf{g}'_c \mathbf{l}_c \tag{6}$$

where $\mathbf{v}' = [\mathbf{v}'_1, \mathbf{v}'_2, \ldots, \mathbf{v}'_{c_1}]$ refers to the weighted fusion features. As Fig. 1 shows, the weighted features are convolved (selected) and passed to the label refinement unit. Here, a standard convolution operation rather than a specific pooling or convolution operation (e.g., depthwise separable convolution) is applied to the weighted features because the standard convolution operation helps to preserve the completeness of weighted features and smooth the labeling maps. The feature selection results would be presented in Section VI-B.

To better demonstrate the design and novelty of the AGFF module, Fig. 3 illustrates an example of different feature fusion modules incorporated with the label refinement unit. The label refinement unit obtains higher level labeling maps with more details by combining the upsampled labeling maps with the selected features from different feature fusion modules. As shown in Fig. 3(a), the direct feature fusion (FF) module [24] transfers a certain number of lower level features to the label refinement unit. These transferred features may contain category ambiguity information that decreases the accuracy of FCNs. Different from the FF module, the gated feature fusion (GFF) module [see Fig. 3(b)], which is used in G-FRNet [20], integrates the higher level and lower level features into a fusion process. GFF module first fuses the selected features from both higher level and lower level features using the elementwise product and then transfers a certain number of features to the label refinement unit. In the GFF module, multiple convolution operations are performed, and the features of the same encoding stage are considered equally important, which could increase the difficulty of feature selection. The main advantage of the proposed AGFF module is that it does not perform feature selection and upsampling operation before connecting to the label refinement unit [see Fig. 3(c)]. The details of lower level features are retained and reweighted when fused with higher level ones. The weighted features are considered as discriminative and informative enough to refine the coarse labeling maps. Note that the weighted features obtained from the AGFF module are recursively used in the next fusion stage and acted as the higher level inputs [see Figs. 3(c) and 1] to strengthen the connection across multilevel features. Conversely, the GFF module in G-FRNet [20] is separately used in each fusion stage without recursion, and only two adjacent level encoding features are involved.

D. Multiscale Supervision

As shown in Fig. 1, six labeling maps with increasing spatial dimensions are produced at the decoder network. For supervised learning, a straight method is to calculate the pixelwise cross-entropy loss between the last labeling maps and the ground truth images. To capture the characteristic and interaction of target objects on different scales, here we apply the multiscale supervision learning approach, which has been used for typical FCN architectures [42], [53].

Let f_d^S denote the dth decoding stage labeling maps at the down-scaling factor s, and the corresponding upsampled full-resolution labeling maps (F_d) can be obtained using the standard upsampling operation as follows:

$$F_d = \begin{cases} \mathrm{UP}_S\left(f_d^S\right), & d \in \{1, 2, \ldots, 5\} \\ f_6, & d = 6 \end{cases} \tag{7}$$

where UP_S denotes the bilinear interpolation operator that upsamples the labeling maps by a factor s; f_6 is the last full-resolution labeling maps. The down-scaling factor s is set according to the decoding stage at which the labeling maps are located. For example, s is set as 16 if the labeling maps is in the second decoding stage.

The cross-entropy loss function is used to account for the differences between the ground truth images and the upsampled full-resolution labeling maps at each decoding stage. Let $x^{(n)}$ denote the nth pixel in the input image and $y^{(n)}$ denote the corresponding nth category label. The loss calculation is described as follows:

$$P_k^{(n,d)} = \frac{\exp\left(m_k^{(n,d)}\right)}{\sum_k \exp\left(m_k^{(n,d)}\right)} \quad \forall n \in N, \forall d \in \{1, \ldots, 6\} \tag{8}$$

$$\mathrm{loss}_d = -\frac{1}{N}\left[\sum_{n=1}^{N} I\{y^{(n)} = k\} \log p_k^{(n,d)}\right] \tag{9}$$

where k is the category label; N is the total number of pixels in an input image; $m_k^{(n,d)}$ and $P_k^{(n,d)}$ denote the response values of the dth decoding full-resolution labeling maps and the category probability of $x^{(n)}$, respectively; loss_d represents the cross-entropy loss obtained from the dth decoding stage; $I\{y^{(n)} = k\}$ denotes an indicator function.

The proposed ALRNet is optimized using back propagation and the final output loss of ALRNet is calculated by summing up all the loss from each decoding stage

$$\mathrm{loss}_{\text{total}} = \sum_{d=1}^{6} \mathrm{loss}_d\,. \tag{10}$$

TABLE I
INFORMATION ON ALL TRAINING IMAGES, VALIDATION IMAGES, AND TEST IMAGES FOR THREE DATASETS

Dataset	Resolution	Training images	Validation images	Test images
Potsdam dataset	0.05 m	18 (full: 6,000 × 6,000) 7,200 (patch: 480 × 480)	6 (full: 6,000 × 6,000)	14 (full: 6,000 × 6,000)
Vaihingen dataset	0.09 m	16 (full: ∼2494 × 2064) 746 (patch: 480 × 480)	Not used	17 (full: ∼2,494 × 2,064)
WHU dataset	0.3 m	4,736 (patch: 512 × 512)	1,036 (patch: 512 × 512)	2416 (patch: 512 × 512)

IV. Experiment Design

A. Dataset Descriptions

To validate the robustness of our developed model, we conducted experiments on three publicly available VHR aerial image datasets with distinctive characteristics.

ISPRS Potsdam labeling dataset: This is a commonly used benchmark dataset released by the 2-D semantic labeling contest[1] organized by the ISPRS Working Group III/4. It contains 38 VHR orthoimages, lidar-derived elevation products (e.g., DSM and nDSM), and the corresponding annotated images. Each orthoimage has an image size of 6000 × 6000 pixels with a ground sampling distance (GSD) of 5 cm, and consists of 4 spectral bands, i.e., 1) near-infrared (NIR), 2) red (R), 3) green (G), and 4) blue (B). The annotated images are densely classified into six classes: 1) impervious surfaces, 2) buildings, 3) low vegetation, 4) trees, 5) cars, and 6) clutter. All the annotated images have been made available to the public. In total, 24 images were used for training and validation, and the remaining 14 images are used for online comparison. The validation images includes six tiles (ID: 3_12, 4_12, 5_12, 6_12, 7_8, 7_12). We used the three-band images NIR-R-G as inputs to the network, without using the elevation data.

ISPRS Vaihingen labeling dataset: Another benchmark dataset provided by ISPRS for semantic labeling contest. The Vaihingen dataset consists of 33 spectral orthoimages and corresponding annotated images. Each image has an averaging size of 2494 × 2064, with a spatial resolution of 9 cm, and is composed of only three bands, i.e., NIR-R-G. The annotated images have the same classification schema as the Potsdam dataset. In our experiments, we used all the 16 labeled images for model training and the remaining 17 images for testing. Only NIR-R-G images were used as inputs to the network.

WHU building dataset: The Wuhan University (WHU) building dataset covers an area of about 45 km^2, with more than 187 000 buildings with diverse shapes and appearances in New Zealand[2] [54]. It is set up for evaluating the performance of different methods on building extraction. A total of 8188 RGB images of 512 × 512 pixels are provided at a GSD of 0.3 m. The building footprints have been manually edited and rasterized into images. The dataset was divided into a training set (4736 images), a validation set (1036 images), and a test dataset (2416 images), respectively.[3]

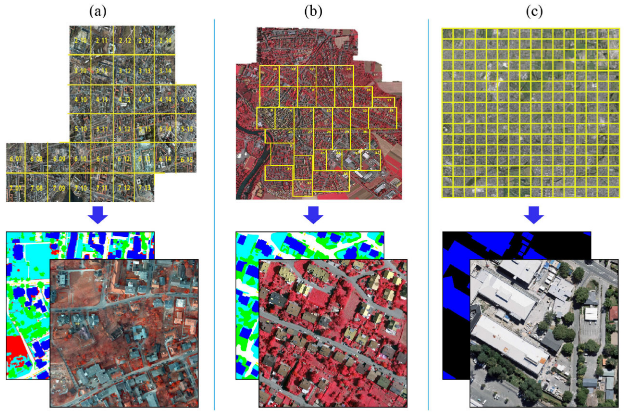

Fig. 4. Image samples and corresponding ground reference images for (a) the Potsdam labeling dataset, (b) the Vaihingen labeling dataset, and (c) the WHU aerial building dataset.

Table I lists the detailed information about these datasets, and Fig. 4 displays some spectral image samples and the corresponding ground truth images. There are many intricate human-made objects and complex scenarios in the VHR images, which pose challenges for obtaining both consistent and accurate semantic segmentation results.

B. Implementation Details

The large images were cropped into subpatches as network inputs due to the limitation of the GPU memory sizes. For both the Potsdam and Vaihingen datasets, each training image was cropped into a sequence of patches (size of 480 × 480 pixels) with an overlap of 200 pixels, whereas each validation and test image was cropped into patches (size of 480 × 480 pixels) with an overlap of 100 pixels. We did not crop the images for the WHU building dataset. Details on the cropped patches are listed in Table I. In the training stage, data augmentation techniques were applied to increase the diversity of training samples. Each training patch was randomly rotated every 90° in the clockwise direction, flipped in the horizontal direction, and its brightness and contrast were also adjusted (the adjustment ratios are set as 0.7 or 1.3). In the inference stage, the full-resolution probability map was created by merging all the patch results, and the class probability of each overlapping pixel was obtained by averaging the prediction values of that pixel.

The proposed ALRNet was implemented using Caffe [55] on an NVIDIA GTX Titan X GPU. The Adam optimizer [56] was used to improve the model convergence because it could iteratively update the learning rate based on the training data. The parameters β_1 and β_2 were set as 0.9 and 0.999, respectively. The initial learning rate of 0.00005 was determined through

[1] Online. [Available]: https://www2.isprs.org/commissions/comm2/wg4/benchmark/semantic-labeling/

[2] Online. [Available]: http://study.rsgis.whu.edu.cn/pages/download/building_dataset.html

[3] Online. [Available]: https://www2.isprs.org/commissions/comm2/wg4/results/vaihingen-2d-semantic-labeling-contest/

trials and errors (from 0.001 to 0.00001) based on the speed of convergence and accuracy on the Potsdam validation dataset. The maximum iteration number was 200 000 for the Potsdam dataset, 30 000 for the Vaihingen dataset, and 300 000 for the WHU building dataset, respectively. The minibatch size was default as 4. Additionally, we initialized the parameters for the encoder part of ALRNet using the pretrained weights of VGG-16.

C. Method Comparisons

To understand the performance of ALRNet, five FCN models, including U-Net [45], SegNet [25], Att-U-Net [30], G-FRNet [20], and GRRNet [29], were used for method comparisons. The main reason for selecting these models is that all these models were verified in the field of computer vision or remote-sensing image processing, and all models are open source and easy for implementation. Besides, the feature transmission and the fusion mechanism of these models have motivated the design of ALRNet.

SegNet and U-Net are commonly used as baselines for semantic segmentation comparisons because of their elegant encoder–decoder architectures. To restore the spatial details lost during the downsampling operations, SegNet delivers the indices of max-pooling to the decoder, whereas U-Net fused the corresponding lower level and higher level features by skip connections. Based upon the architecture of U-Net, Att-U-Net introduces a new attention module into each feature fusion stage and enhances the network prediction capabilities. Both G-FRNet and GRRNet were developed based on the label refinement network (LRN) [24] that gradually up-samples and refines the coarse labeling maps in each decoding stage. The aforementioned models applied the same training strategy as ALRNet to facilitate direct method comparisons. Note that the minibatch size of each network was different and was set according to the GPU memory size. SegNet, G-FRNet, and GRRNet initialized their weights using the corresponding pretrained models, e.g., VGGNet and ResNet-50 [57], and U-Net and Att-U-Net initialized the weights from scratch.

D. Evaluation Metrics

To facilitate a comparison with state-of-the-art methods, different metrics are considered regarding different datasets. For the Potsdam and Vaihingen test datasets, the overall accuracy (OA) and F1 score are used, whereas for the WHU test dataset, the precision, recall, F1, and IoU metrics are employed. The above-employed metrics are consistent with those used in the relevant literature [17], [41], [54]. Typically, for the Potsdam validation dataset, we also use IoU for evaluation, since IoU is often used as one of the essential indicators of semantic segmentation. OA is the proportion of the correctly classified pixel numbers to the total pixel numbers in a single image or a whole image dataset. F1 could be regarded as the harmonic mean of precision and recall. IoU, also called the Jaccard similarity coefficient, is the ratio of correctly classified pixel numbers to the total amounts of reference annotated pixels and the detected pixels. All of the metrics mentioned before, except for OA, are explicitly calculated for each category and are defined, respectively, as follows:

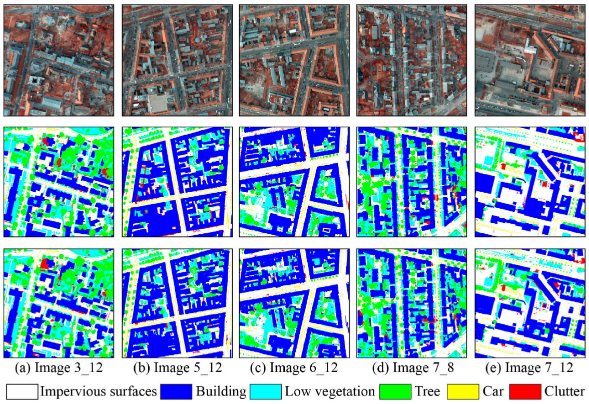
(a) Image 3_12 (b) Image 5_12 (c) Image 6_12 (d) Image 7_8 (e) Image 7_12

Fig. 5. False-color-composite (NIR-R-G) images for the Potsdam validation dataset (top row), the ground references (middle row), and classification results produced by ALRNet (bottom row) are shown for different image tiles, respectively. The ID for image tile is shown for each column. (a) Image 3_12. (b) Image 5_12. (c) Image 6_12. (d) Image 7_8. (e) Image 7_12.

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{11}$$

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

$$F_1 = \frac{2\times \text{ Precision} \times \text{ Recall}}{\text{Precision}+\text{Recall}} \tag{13}$$

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively. Note that all the metrics for the Potsdam and Vaihingen datasets were computed on eroded boundary ground truths, so as to reduce the impact of uncertain boundary annotations.

V. Results

A. Potsdam Dataset Results

Fig. 5 presents the classification results of ALRNet on the Potsdam validation dataset. Comparing with the reference annotated images, ALRNet obtained satisfactory results in different scenarios. As shown in Table II, ALRNet achieved an OA of 92.6% and a mean IoU of 76.4% and outperformed other methods. ALRNet achieved the best classification accuracies on extracting impervious surfaces, buildings, and trees. Fig. 6 displays some close-up views of the Potsdam validation images and the results produced by the comparison methods. Both U-Net and Att-U-Net have difficulties in detecting the complete building objects (e.g., the second and third rows in Fig. 6), and their results are less accurate than others (see Table II). SegNet has misclassification between the building and the background objects in some scenarios (e.g., the second row in Fig. 6). As shown in the first row in Fig. 6, most methods did not extract the entire building, whereas ALRNet did well in this case.

To further validate the semantic segmentation performance of ALRNet, we applied the best training model to the Potsdam test dataset. As shown in Table III, ALRNet achieved an OA of

TABLE II
THE STATISTICAL RESULTS OBTAINED USING ALRNET AND COMPARISON METHODS ON THE POTSDAM VALIDATION DATASET

Model	Imp surf		Building		Low veg		Tree		Car		Clutter		Overall	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	OA	mIoU
Att-U-Net [30]	90.4	82.6	94.4	89.4	81.9	69.4	86.8	76.7	95.2	90.9	34.5	20.9	89.2	71.6
U-Net [45]	91.1	83.6	94.2	89.0	81.5	68.8	86.1	75.6	95.5	91.4	35.8	21.8	89.4	71.7
SegNet [25]	93.3	87.5	96.1	92.4	84.8	73.7	88.2	78.8	**96.1**	92.4	39.1	24.3	91.9	74.9
GRRNet [29]	93.4	87.6	96.7	93.6	84.5	73.1	88.0	78.6	**96.1**	**92.5**	38.0	23.5	92.1	74.8
G-FRNet [20]	93.8	88.4	96.8	93.7	**85.5**	**74.7**	88.2	78.9	95.5	91.5	43.2	27.5	92.5	75.8
ALRNet	**93.9**	**88.6**	**97.3**	**94.7**	85.3	74.4	**88.3**	**79.1**	96.0	92.3	**45.1**	**29.1**	**92.6**	**76.4**

The bold values represent the best result and the underlined values represent the second-best result achieved by models. “mIoU” denotes the mean IoU of all categories.

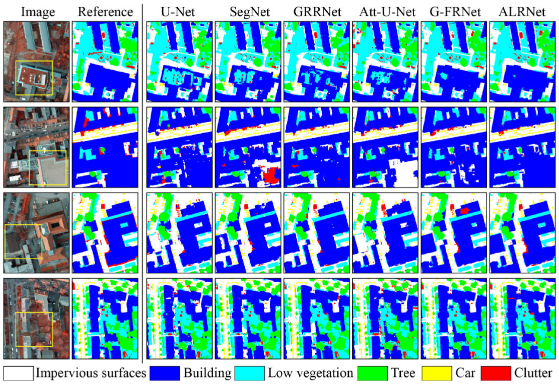

Fig. 6. Close-up views of the Potsdam validation dataset and the classification results using ALRNet and comparison methods. The distinct differences are marked with yellow rectangles.

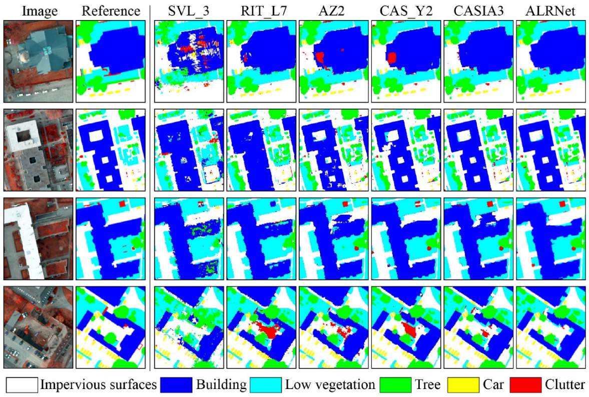

Fig. 7. Close-up views of the Potsdam test dataset and the classification results using ALRNet and five state-of-the-art methods.

TABLE III
QUANTITATIVE COMPARISON (%) WITH STATE-OF-THE-ART METHODS ON THE POTSDAM TEST DATASET

Method	Elevation	Imp sur	Building	Low veg	Tree	Car	OA
SVL_3 [44]	w/	84.0	89.8	72.0	59.0	69.8	77.2
UFMG_4 [58]	w/	90.8	95.6	84.4	84.3	92.4	87.9
RIT_L7	w/	91.2	94.6	85.1	85.1	92.8	88.4
AZ2	w/o	92.3	96.0	86.4	87.6	95.1	89.9
DST_5 [60]	w/	92.5	96.4	86.7	88.0	94.7	90.3
RIT4 [59]	w/	92.6	97.0	86.9	87.4	95.2	90.3
CAS_Y2	w/o	92.6	96.2	87.3	87.7	95.7	90.4
BUCTY5	w/	93.1	**97.3**	86.8	87.1	94.1	90.6
HUSTW2	w/	93.2	96.5	87.3	88.3	93.9	90.9
CASIA3 [41]	w/o	93.4	96.8	**87.6**	88.3	**96.1**	**91.0**
ALRNet	w/o	**93.5**	96.1	87.3	**89.0**	95.5	90.9

“w/o” denotes the method without using elevation data and “w/” denotes the method using elevation data.

TABLE IV
QUANTITATIVE COMPARISON (%) WITH STATE-OF-THE-ART METHODS ON THE VAIHINGEN TEST DATASET[4]

Method	Elevation	Imp surf	Building	Low veg	Tree	Car	OA
SVL_4 [44]	w/	86.1	90.9	77.6	84.9	59.9	84.7
DST_2 [60]	w/	90.5	93.7	83.4	89.2	72.6	89.1
ONE_7 [61]	w/	91.0	94.5	**84.4**	**89.9**	77.8	89.8
INR [62]	w/	91.1	94.7	83.4	89.3	71.2	89.5
RIT_7 [59]	w/	91.7	95.2	83.5	89.2	82.8	89.9
DLR_9 [2]	w/	**92.4**	95.2	83.9	**89.9**	81.2	90.3
WUH_W4	w/o	92.3	94.9	83.2	89.0	85.3	90.1
GSN3 [28]	w/o	92.2	95.1	83.7	**89.9**	82.4	90.3
ALRNet	w/o	**92.4**	**95.4**	83.9	89.6	**85.6**	**90.5**

“w/o” denotes the method without using elevation data and “w/” denotes the method using elevation data.

90.9%, which is competitive with some state-of-the-art methods (e.g., HUSTW2, CASIA3 [41]). The F1 scores for impervious surfaces and trees are 93.5% and 89.0%, respectively, both of which are higher than the results derived from CASIA3. Compared with other methods using elevation data, e.g., RIT4 [59] and HUSTW2, ALRNet has slightly lower accuracy on building extraction but has a more robust performance on extracting impervious surfaces and trees. The visual comparison results in Fig. 7 further show that ALRNet identified accurate and precise boundary objects in the Potsdam test images. Accordingly, ALRNet achieved reliable classification accuracy on the Potsdam dataset using only spectral images, and the accuracy of some individual categories is better than some methods that currently achieve top performance in the ISPRS contests.[4]

B. Vaihingen Dataset Results

As shown in Fig. 8, the prediction results of ALRNet on the Vaihingen test dataset are close to the ground truth images and the outlines of urban objects are distinguishable. Table IV compares the quantitative results of ALRNet with state-of-the-art methods. The OA of ALRNet is 90.5%, which is higher than other methods that did not employ the auxiliary elevation data, e.g., GSN3 [28] and WUH_W4. ALRNet obtained the best accuracies on detecting impervious surfaces, buildings, and cars. Fig. 9 exhibits the close-up results of ALRNet and other

[4] Online. [Available]: https://www2.isprs.org/commissions/comm2/wg4/results/potsdam-2d-semantic-labeling/

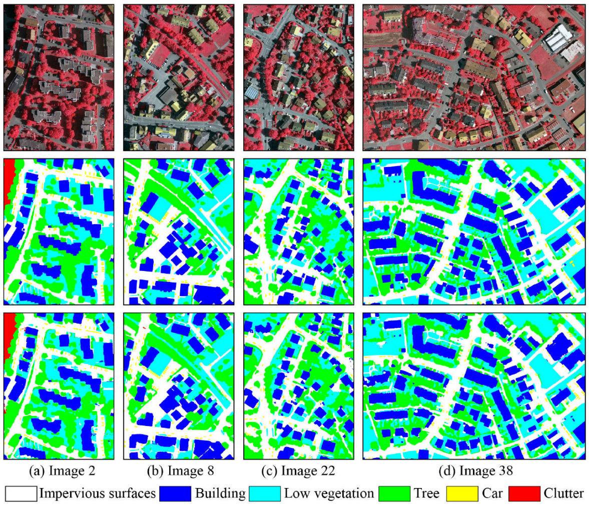

Fig. 8. False-color-composite images (NIR-R-G) for the Vaihingen test dataset (top row), the ground references (middle row), and classification results produced by ALRNet (bottom row) are shown for different image tiles. The ID for image tile is shown for each column. (a) Image 2. (b) Image 8. (c) Image 22. (d) Image 38.

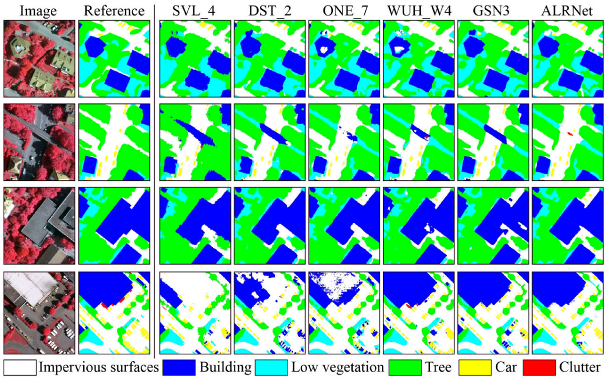

Fig. 9. Close-up views of the Vaihingen test dataset and the classification results using ALRNet and five state-of-the-art methods.

methods. There were apparent missing building objects in the results of SVL_4 [44], DST_2 [60], and ONE_7 [61] (e.g., the fourth row in Fig. 9). ALRNet obtained robust classification results, and notably, it correctly identified the overhead road (e.g., the second row in Fig. 9).

C. WHU Dataset Results

While tests of ALRNet on the ISPRS benchmarks have demonstrated that its ability to deal with the multiclass classification problem in VHR images, we further validated ALRNet on the WHU building dataset for recognizing the building objects from images of different spatial resolution and band composition. As shown in Fig. 10, ALRNet performed well in diverse scenarios with excellent segmentation of object contours. According to the quantitative results summarized in Table V, ALRNet produced the best accuracies with an F1 score of 94.5% and an IoU of 89.5% among all methods. SRI-Net [17] obtained higher precision but lower recall than ALRNet. GRRNet has the lowest recall, indicating that this method needs to be improved in terms of completeness. In our experiments, U-Net and Att-U-Net obtained low precision, which affected their OA. As shown in Fig. 11, ALRNet obtained more accurate and complete extraction results than other tested methods.

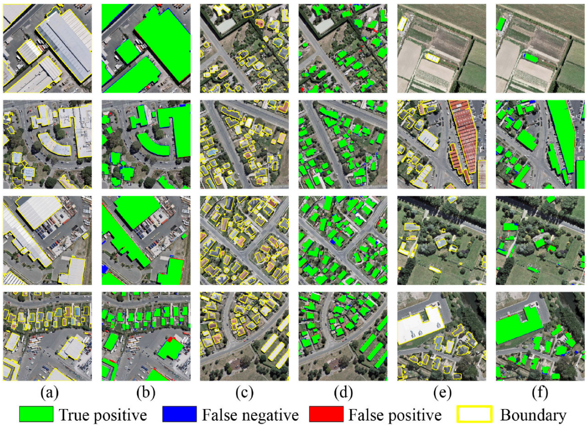

Fig. 10. Example images for the WHU building test dataset and the classification results produced by ALRNet. The building boundary (marked in yellow polygon) is obtained by vectorizing the segmentation results.

TABLE V
QUANTITATIVE RESULTS (%) OBTAINED USING ALRNET AND OTHER METHODS ON THE WHU BUILDING TEST DATASET

Method	Precision	Recall	F1	IoU
U-Net [45]	90.6	95.7	93.1	87.0
SegNet [25]	94.0	94.0	94.0	88.7
GRRNet [29]	94.8	90.4	92.6	86.2
Att-U-Net [30]	92.9	94.8	93.8	88.4
G-FRNet [20]	94.2	93.8	94.0	88.7
SiU-Net [54]	93.8	93.9	—	88.4
SRI-Net [17]	**95.2**	93.3	94.2	89.1
ALRNet	94.9	94.1	**94.5**	**89.5**

The bold values represent the best result and the underlined values represent the second-best result achieved by models.

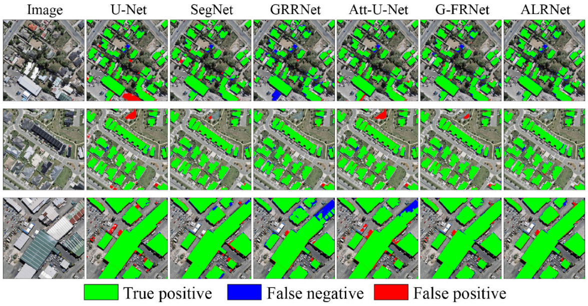

Fig. 11. Visual comparison with comparison methods on the WHU building test dataset.

VI. DISCUSSIONS

A. Ablation Experiments for Model Components Analysis

To examine the contribution of each component on the performance of ALRNet, some ablation experiments were conducted. Tables VI and VII list some ALRNet variants' results on the Potsdam and the WHU building validation dataset, respectively.

TABLE VI
VALIDATION RESULTS (%) OBTAINED USING ALRNET AND ITS VARIANTS ON THE POTSDAM DATASET

Model	Imp surf		Building		Low veg		Tree		Car		Clutter		Overall	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	OA	mIoU
ALRNet_noMSS	93.6	88.1	96.8	93.8	84.7	73.5	88.1	78.7	95.7	91.8	43.1	27.5	92.1	75.5
ALRNet_noMSS_noRec	93.4	87.7	96.8	93.8	84.2	72.6	87.7	78.1	95.6	91.6	40.4	25.3	91.9	74.9
ALRNet_ResNet	93.3	87.5	96.6	93.4	84.7	73.4	88.2	78.9	**96.2**	**92.7**	43.0	27.4	92.0	75.5
ALRNet_noFF	93.2	87.3	96.4	93.0	84.1	72.6	88.2	78.8	92.6	86.3	41.2	25.9	91.7	74.0
ALRNet_FF	93.6	87.9	96.9	94.0	85.0	73.8	88.2	78.8	96.0	92.4	**45.7**	**29.7**	92.1	76.1
ALRNet_GFF	93.6	87.9	96.6	93.4	84.8	73.5	88.0	78.6	95.9	92.0	38.5	23.8	92.1	74.9
ALRNet_noRec	93.7	88.1	96.7	93.7	85.2	74.2	**88.3**	**79.1**	**96.2**	**92.7**	43.5	27.8	92.4	76.0
ALRNet	**93.9**	**88.6**	**97.3**	**94.7**	**85.3**	**74.4**	**88.3**	**79.1**	96.0	92.3	45.1	29.1	**92.6**	**76.4**

"_noMSS" denotes the network without using multi-scale supervision. "_noRec" denotes the network that applies the AGFF module without recursion. "_noFF" denotes the network without transferring encoding features. "_ResNet" denotes the network that replaces the VGG-16 with ResNet-50.

TABLE VII
VALIDATION RESULTS (%) OBTAINED USING ALRNET AND ITS VARIANTS ON THE WHU BUILDING DATASET

Method	Precision	Recall	F1	IoU
ALRNet_noMSS	96.2	94.3	95.2	90.9
ALRNet_noMSS_noRec	94.5	**95.6**	95.0	90.5
ALRNet_ResNet	92.1	93.4	92.7	86.5
ALRNet_noFF	89.7	87.3	88.5	79.4
ALRNet_FF	94.4	94.8	94.6	89.8
ALRNet_GFF	**96.3**	93.5	94.9	90.3
ALRNet_noRec	95.3	94.7	95.0	90.5
ALRNet	95.9	94.8	**95.3**	**91.1**

Here, we use ALRNet with AGFF module as the baseline (ALRNet). The accuracy of ALRNet_noMSS was lower than ALRNet, indicating that the multiscale supervision schema could effectively improve the performance of ALRNet. According to the results of ALRNet_ResNet, an encoder with deeper depth did not increase the OA of ALRNet. The model with no encoding feature transferred (ALRNet_noFF) got the lowest accuracy among all methods, indicating that the proper use of encoding features is key to improving our model. By comparing the performance of different feature fusion modules (i.e., FF and GFF module), the AGFF module has better performance than others, showing that the channelwise attention is feasible for feature fusion and feature selection in ALRNet. ALRNet with recursion has higher accuracy than that without recursion (ALRNet_noRec). By comparing the accuracies of "ALRNet_noMSS_noRec" and "ALRNet_noMSS," we found that the AGFF module without recursion was still effective and could improve the performance of ALRNet without multiscale supervision.

B. Model Visualization

To better understand the attention mechanism, we carried out feature map visualization to analyze how the AGFF module affects the performance of ALRNet. Fig. 12 shows the stagewise classification results of ALRNet using different feature fusion modules for an input image. All the methods obtained coarse prediction results in the first decoding stage. The AGFF module progressively rectified the misclassification pixels since the second decoding stage, whereas the FF and GFF modules had difficulties in improving the subsequent outputs. ALRNet_noMSS did not obtain precise classification results until the last decoding stage because it did not apply the deep supervision schema, but once again shows the effectiveness of the proposed AGFF module.

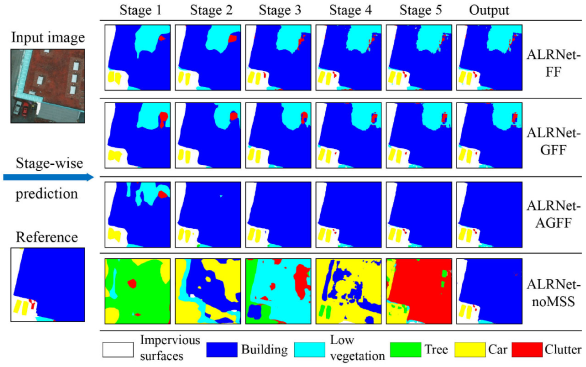

Fig. 12. Visual comparisons of stagewise classification results among the proposed ALRNet (ALRNet_AGFF) and its variants on the close-up view of the Potsdam validation image.

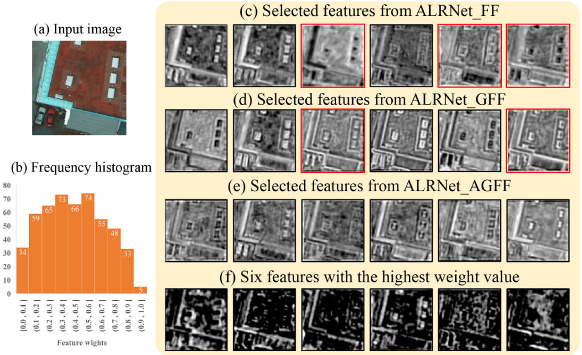

Fig. 13. Weight values and selected feature maps in the second feature fusion stage. (a) Close-up view of the Potsdam validation image. (b) Weight value frequency histogram. (c)–(e) Selected features from different feature fusion modules. (f) Six features with the highest weight value.

The proposed AGFF module has an effect on ALRNet by influencing feature weighting and selection. Figs. 13 and 14 display the obtained feature maps in the second and the fifth feature fusion stages, respectively. The feature weight values are mostly distributed between 0.3 and 0.6 and decrease toward both ends [see Figs. 13(b) and 14(b)]. That means a few numbers of features have high weight values and are emphasized by the attention-guided modules. Figs. 13(f) and 14(f) show top six features with the highest weight values, respectively. These features represent different locations of the image, and some are particularly sensitive to edges and textures. By the channelwise feature reweighting, the AGFF module could reduce the

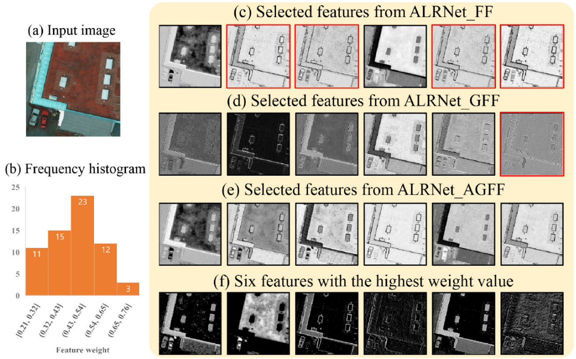

Fig. 14. Weight values and selected feature maps in the fifth feature fusion stage. (a) Close-up view of the Potsdam validation image. (b) Weight value frequency histogram. (c)–(e) Selected features from different feature fusion modules. (f) Six features with the highest weight value.

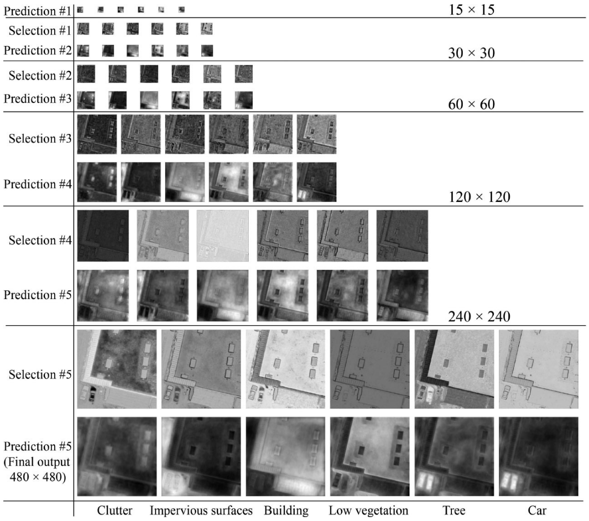

Fig. 15. Feature selection and prediction results of ALRNet in different feature fusion stages.

features passed in the network and obtain more representative characteristics to refine the coarse labeling maps. Figs. 13(c)–(e) and 14(c)–(e) display the selected (convolved) features that are passed to the label refinement unit. Compared to the FF and the GFF modules, the features obtained by the AGFF module are not duplicated but have distinguishable object outlines and extra response to different locations.

To further analyze how the AGFF module influences the refinement process of ALRNet, Fig. 15 shows the feature selection results and prediction results obtained by ALRNet in each stage. We found that the selected features were close to the input image, and their spatial details were more prosperous than the coarse labeling maps at the corresponding stage. The coarse labeling maps contained much information related to the category determination but without more precise object outlines. By merging the selected features from the AGFF module, the coarse labeling maps were gradually rectified into higher resolution and higher confidence labeling maps.

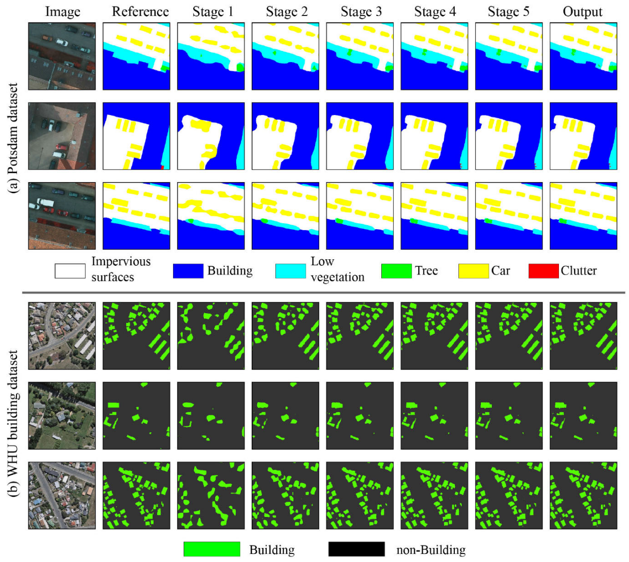

Fig. 16. Classification results of ALRNet in typical scenes containing small objects (e.g., cars and small-size buildings).

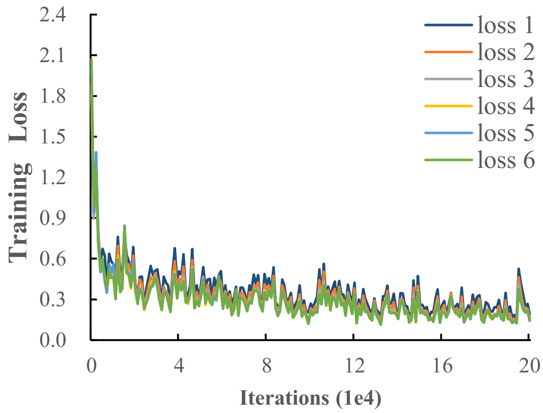

Fig. 17. Multistage training losses of ALRNet on the Potsdam dataset.

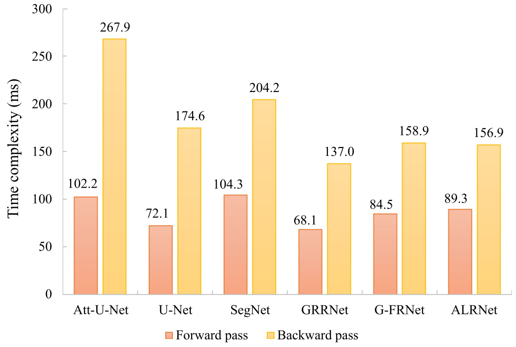

Fig. 18. Comparisons of the efficiency among different methods. The Caffe time command was applied to calculate the time requirements as averaged over 200 iterations with an input image size of 480 × 480 pixels.

C. Effect of Multiscale Supervision

According to Section VI-A, multiscale supervision can effectively improve the classification accuracy of ALRNet in different categories. The bilinear interpolation operators with various scaling factors are employed to capture target objects'

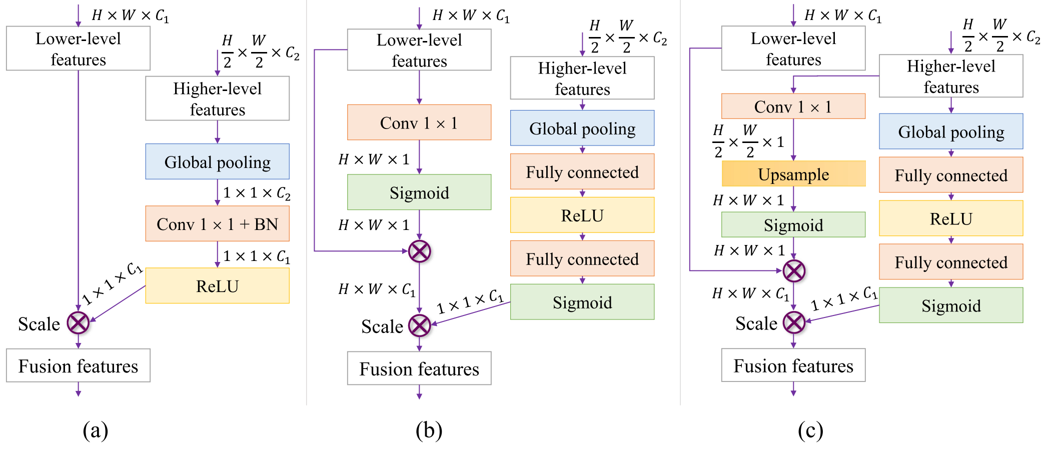

Fig. 19. Schema of three tested models with different attention strategies. (a) Model_1. (a) Model_2. (a) Model_3.

interaction on different scales. To analyze whether such upsampling operations will affect small objects' classification results, some sample images from the Potsdam and the WHU building test datasets were predicted. As Fig. 16 shows, ALRNet could successfully detect small objects such as cars and small-size buildings at different supervision stages (scales). Meanwhile, the small objects in the earlier supervision stages were gradually refined with sharper edges.

D. Model Efficiency Analysis

Fig. 17 displays the training convergence of ALRNet under different supervision stages. Specifically, the training losses in different stages show a rapid downward trend until 8000 iterations, and they reached convergence at around 160 000 iterations. In general, the early stages' losses (e.g., loss 1) are higher than the losses in the later stages (e.g., loss 6). The coarse labeling maps in the early stages lack more accurate spatial details, so it is challenging to keep alike with the ground truth images.

Fig. 18 compares the time complexity of ALRNet with other methods. GRRNet requires less training and inference time than others because of its ResNet-based encoder and limited feature maps transferred. The computational complexity of ALRNet is close to G-FRNet, meaning that the introduction of the SE module does not largely affect the efficiency of model training and testing. ALRNet has higher computational efficiency as compared to U-Net and SegNet because the latter two models do not constraint the number of features transmitted from the encoder. The AG module introduced by Att-U-Net [30] increases the computational time as compared with U-Net because the AG module performs more complicated feature fusion operations.

E. Failed Attempts Using Different Attention Strategies

Before developing the AGFF module, many efforts have been made to find alternative attention strategies feasible for multilevel feature fusion in ALRNet. On the one hand, a global attention upsample module proposed in [26] was adapted to ALRNet. As Fig. 19(a) shows, the global context aggregated from higher level features are passed to a 1 × 1 kernel-size convolutional block and multiplied by the lower level features.

TABLE VIII
QUANTITATIVE COMPARISON (%) WITH TESTED MODELS ON POTSDAM VALIDATION DATASET WITH OA, MEAN IOU AS WELL AS IOU SCORE FOR EACH CATEGORY

Models	Imp_suf	Building	Low veg	Tree	Car	OA	mIoU
Model_1	87.9	94.1	73.5	78.4	92.0	92.3	76.0
Model_2	87.3	94.2	74.0	79.0	**92.7**	92.1	75.8
Model_3	87.6	93.5	73.9	**79.1**	92.3	92.2	76.2
ALRNet	**88.6**	**94.7**	**74.4**	**79.1**	92.3	**92.6**	**76.4**

On the other hand, motivated by Roy *et al.* [47], we combined the channelwise SE module with an SSE for feature fusion. In the SSE module, input features are first compressed into 1-D and then passed to the sigmoid activation function. Two models [see Figs. 19(b) and (c)] are designed by placing the SSE module on top of the lower level and higher level features, respectively. According to the accuracy reported in Table VIII, ALRNet with AGFF module surpasses other tested models, showing that the channelwise SE module is more suitable for ALRNet.

VII. CONCLUSION

In this article, we presented an ALRNet for VHR image segmentation. ALRNet progressively refines the coarse labeling maps of different scales by using the channelwise attention mechanism. Specifically, a novel AGFF module is proposed based on the SE module to reduce the semantic gaps among features of different levels. The AGFF module not only improves the category discriminability of each pixel in the lower level features but also assigns different weights to the fusion features that are helpful for a subsequent feature selection. The ablation experiments demonstrate that our AGFF module can effectively refine the stagewise classification results of ALRNet. Moreover, model visualization indicates that the proposed module effectively assigns different weights to the fusion features, which helps select representative and nonredundant features. The proposed method not only achieves the outstanding performances on the ISPRS 2-D Semantic Labeling Contest for Potsdam and Vaihingen but also surpasses the state-of-the-art methods on the WHU aerial building dataset. The designed attention-guided module can be integrated with other FCNs. Source code[5] and pretrained models are made publicly available for further studies.

Acknowledgment

The authors would like to thank the ISPRS for making the Vaihingen and Potsdam datasets available and organizing the semantic labeling contest. They also would like to thank Shunping Ji and his teammates for creating the high-quality WHU aerial imagery dataset for building extraction. The authors wish to thank anonymous reviewers for their constructive comments.

References

[1] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS J. Photogrammetry Remote Sens.*, vol. 115, pp. 22–36, May 2016.

[2] D. Marmanis, K. Schindler, J. 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, Jan. 2018.

[3] C. Zhang *et al.*, "Joint deep learning for land cover and land use classification," *Remote Sens. Environ.*, vol. 221, pp. 173–187, Feb. 2019.

[4] H. L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri, "Building extraction at scale using convolutional neural network: Mapping of the United States," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2600–2614, Aug. 2018.

[5] Y. Sun, X. Zhang, Q. Xin, and J. Huang, "Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 143, pp. 3–14, Sep. 2018.

[6] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 66, no. 3, pp. 247–259, May 2011.

[7] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogrammetry Remote Sens.*, vol. 114, pp. 24–31, Apr. 2016.

[8] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, and Y. Liu, "A review of supervised object-based land-cover image classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 130, pp. 277–293, Aug. 2017.

[9] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. D. Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 130, pp. 139–149, Aug. 2017.

[10] 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.

[11] W. Zhao, S. Du, Q. Wang, and W. J. Emery, "Contextually guided very-high-resolution imagery classification with semantic segments," *ISPRS J. Photogrammetry Remote Sens.*, vol. 132, pp. 48–60, Oct. 2017.

[12] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Trans. Geosci. Remote Sens.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[13] B. Huang, B. Zhao, and Y. Song, "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery," *Remote Sens. Environ.*, vol. 214, pp. 73–86, Sep. 2018.

[14] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[15] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, "Learning aerial image segmentation from online maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6054–6068, Nov. 2017.

[16] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 60–77, 2018.

[17] P. Liu *et al.*, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, Apr. 2019, Art. no. 830.

[18] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution aerial image labeling with convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7092–7103, Dec. 2017.

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

[20] M. A. Islam, M. Rochan, N. D. B. Bruce, and Y. Wang, "Gated feedback refinement network for dense image labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jul. 2017, pp. 4877–4885.

[21] 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.

[22] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2019, pp. 1480–1484.

[23] 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 Recog.*, Jul. 2017, pp. 5168–5177.

[24] M. A. Islam, S. Naha, M. Rochan, N. Bruce, and Y. Wang, "Label refinement network for coarse-to-fine semantic segmentation," 2017, *arXiv:1703.00551*. [Online]. Available: https://arxiv.org/abs/1703.00551

[25] 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.

[26] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*. [Online]. Available: https://arxiv.org/abs/1805.10180

[27] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. Conf. Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2019, pp. 3141–3149.

[28] H. Wang, Y. Wang, Q. Zhang, S. Xiang, and C. Pan, "Gated convolutional neural network for semantic segmentation in high-resolution images," *Remote Sens.*, vol. 9, no. 5, May 2017, Art. no. 446.

[29] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, pp. 91–105, May 2019.

[30] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*. [Online]. Available: https://arxiv.org/abs/1804.03999

[31] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jul. 2017, pp. 936–944.

[32] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, Mar. 2001.

[33] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.* 2016, Jun. 2016, pp. 3640–3649.

[34] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jul. 2017, pp. 6450–6458.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2018, pp. 7132–7141.

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

[37] L. Bruzzone and B. Demir, "A review of modern approaches to classification of remote sensing data," in *Land Use and Land Cover Mapping in Europe. Remote Sensing and Digital Image Processing*. Dordrecht, The Netherlands: Springer, 2014, pp. 127–143.

[38] S. W. Myint, P. Gober, A. Brazel, S. Grossman-Clarke, and Q. Weng, "Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery," *Remote Sens. Environ.*, vol. 115, no. 5, pp. 1145–1161, May 2011.

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

[40] G. Cheng, Y. Wang, S. Xu, H. Wang, S. Xiang, and C. Pan, "Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3322–3337, Jun. 2017.

[41] 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, Nov. 2018.

[5] Online. [Available]: https://github.com/CHUANQIFENG/ALRNet

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

[43] H. Luo, C. Chen, L. Fang, X. Zhu, and L. Lu, "High-resolution aerial images semantic segmentation using deep fully convolutional network with channel attention mechanism," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3492–3507, Sep. 2019.

[44] M. Gerke, *Use of the Stair Vision Library Within the ISPRS 2D Semantic Labeling Benchmark (Vaihingen)*. Enschede, The Netherlands: Univ. Twente, Jan. 2014.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, Cham, Switzerland: Springer, 2015, pp. 234–241.

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

[47] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel squeeze & excitation in fully convolutional networks," in *Proc. Med. Image Comput. Comput. Assist. Intervention*, 2018, pp. 421–429.

[48] H. Yang, P. Wu, X. Yao, Y. Wu, B. Wang, and Y. Xu, "Building extraction in very high resolution imagery by dense-attention networks," *Remote Sens.*, vol. 10, no. 11Nov. 2018, Art. no. 1768.

[49] R. Xu, Y. Tao, Z. Lu, and Y. Zhong, "Attention-mechanism-containing neural networks for high-resolution remote sensing image classification," *Remote Sens.*, vol. 10, no. 10, Oct. 2018, Art. no. 1602.

[50] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, "Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning," *Remote Sens.*, vol. 11, no. 1, Jan. 2019, Art. no. 83.

[51] X. Pan *et al.*, "Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms," *Remote Sens.*, vol. 11, no. 8Apr. 2019, Art. no. 917.

[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[53] S. Xie and Z. Tu, "Holistically-nested edge detection," *Int. J. Comput. Vis.*, vol. 125, no. 1–3, pp. 3–18, Mar. 2017.

[54] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[55] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–41.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016.

[58] 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.

[59] S. Piramanayagam, E. Saber, W. Schwartzkopf, and F. Koehler, "Supervised classification of multisensor remotely sensed images using a deep learning framework," *Remote Sens.*, vol. 10, no. 9, Sep. 2018, Art. no. 1429.

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

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

[62] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution semantic labeling with convolutional neural networks," 2017, *arXiv:1611.01962*. [Online]. Available: https://arxiv.org/abs/1611.01962

Jianfeng Huang received the B.S. and Ph.D. degrees in geographical information science from Sun Yat-sen University, Guangzhou, China, in 2013 and 2020, respectively.

He is currently a Postdoctoral Researcher with the School of Atmospheric Sciences, Sun Yat-sen University. His research interests include deep learning in remote-sensing images and remote-sensing applications.

Xinchang Zhang received the B.S. degree in cartography from the Wuhan Institute of Surveying and Mapping, Wuhan, China, in 1982, the M.S. degree in cartography from the Wuhan Technical University of Surveying and Mapping, Wuhan, in 1994, and the Ph.D. degree in resources and environmental sciences from Wuhan University, Wuhan, in 2004.

He is currently a Professor with the School of Geography and Remote Sensing, Guangzhou University, Guangzhou, China, and a Chair Professor with Henan University, Kaifeng, China. His research interests include spatial database updating, spatial data integration, and smart city.

Ying Sun received the B.S. degree in surveying and mapping from Chang'an University, Xi'an, China, in 2005, and the M.S. degree in surveying and mapping from Wuhan University, Wuhan, China, in 2007, and the Ph.D. degree in geographical information science from Sun Yat-sen University, Guangzhou, China, in 2014.

She was a Visiting Scholar with the Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Hong Kong, in 2018. She is currently an Associate Research Fellow with the School of Geography and Planning, Sun Yat-sen University. Her research interests include high-resolution remote sensing, deep learning in remote sensing images, and ecology remote sensing.

Qinchuan Xin (Member, IEEE) received the B.S. degree in physics from Peking University, Beijing, China, in 2005, and the Ph.D. degree in geography from Boston University, Boston, MA, USA, in 2012.

From 2012 to 2015, he was a Postdoctoral Researcher with Tsinghua University, Beijing. He is currently an Associate Professor with the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China. His research interests include ecological remote sensing and terrestrial ecological model.

Dr. Xin is an Associate Editor for the *International Journal of Remote Sensing*.