# Frequency-Driven Edge Guidance Network for Semantic Segmentation of Remote Sensing Images

Jinsong Li, Shujun Zhang, Yukang Sun, Qi Han, Yuanyuan Sun, and Yimin Wang

***Abstract*—Semantic segmentation plays a significant role in parsing remote sensing images. However, mainstream segmentation models lack a thorough understanding of the complex structures and scale differences, and struggle to effectively locate and emphasize diverse edges. Aiming at these limitations, we propose a frequency-driven edge guidance network, named FDEG-Net, for semantic segmentation of remote sensing images. First, we design a joint sparse context aggregation module that integrates both dense local context and sparse long-range context to improve the analysis of intricate and multiscale objects. Second, an edge guidance module is developed for strong interclass edge acquisition. It applies a 2-D discrete wavelet transform, coefficient superposition method, and adaptive edge feature enhancement algorithm to reduce low-frequency information and highlight salient boundaries in spatial features. This module has two significant advantages. 1) The edge positions are defined in pixel intensity with high interpretability. 2) The modular design without additional edge labels is plug-and-play. The effectiveness and robustness of this module are validated through edge visualization results. The proposed FDEG-Net is evaluated on the Potsdam, Vaihingen, and GID datasets, demonstrating its excellent performance in accurately capturing the rich semantics of geographic space features.**

***Index Terms*—Context extraction, edge guidance, remote sensing images (RSIs), semantic segmentation.**

## I. Introduction

SEMANTIC segmentation is a crucial task in comprehending remote sensing images (RSIs) and plays an increasingly important role in land resource utilization, urban planning, and environmental protection [1]. With the rapid advancement of remote sensing technology, images cover more diverse contents with higher spatial resolution [2]. Consequently, traditional image processing and computer vision methods are no longer sufficient for effective analysis. In recent years, deep learning (DL) technology, particularly convolutional neural networks (CNNs), has achieved significant success in various computer vision tasks, including semantic segmentation of RSIs [3]. Due to the excellent feature extraction ability of CNNs, CNN-based semantic segmentation methods in natural images are introduced to RSIs. Specifically, methods based on the fully convolutional network [4] and U-Net [5] have been rapidly developed [6], [7].

Manuscript received 29 December 2023; revised 11 March 2024; accepted 15 April 2024. Date of publication 25 April 2024; date of current version 14 May 2024. This work was supported by the Natural Science Foundation of Shandong Province, China under Grant ZR2021QC120. *(Corresponding author: Shujun Zhang.)*

The authors are with the School of Data Science, Qingdao University of Science and Technology, Qingdao 266061, China (e-mail: plaitkol@mails.qust.edu.cn; zhangsj@qust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3393531

However, convolutional receptive fields are limited in size [8], making it difficult for CNNs to consider enough pixels to establish global context. To overcome this limitation, some studies [9], [10] enlarge the receptive field of CNNs through pooling and dilated convolution. Nevertheless, these methods suffer from sparsity (compressed information and atrous interval), where the object's local discriminant information becomes fuzzy, leading to semantic confusion. One potential solution to this problem is to introduce self-attention [11]. Self-attention uses matrix computation to capture the global correlation between arbitrary pixels. However, self-attention requires significantly more computation than convolution. Further, thanks to the multihead self-attention, some Transformer-based schemes have made promising progress [12], [13]. While these methods have powerful capabilities in capturing global context, they also exhibit limitations in computational resources and memory efficiency [14].

Recently, researchers have explored the large dense convolutions as an alternative to self-attention for context modeling [15]. This method has demonstrated the potential to rival self-attention while maintaining acceptable computational costs. Therefore, inspired by the fuzzy vision of humans, we consider the design of the receptive field from the perspective of the convolutional kernel, aiming to strike a balance between sparsity and density. We develop a joint sparse context aggregation (JSCA) approach, which adopts a large sparse window and dense local context to approximate the human visual field (focused center and fuzzy surroundings). This approach allows small dense kernels to focus on the current object, capturing its discriminant features, whereas large sparse kernels expand the context to establish global understanding simultaneously.

In addition to expanding receptive the fields of CNNs, detailed spatial information is essential for semantic segmentation. While the semantic representation is acquired through image encoding, this comes at the cost of losing spatial details, which hinders dense pixel classification [16]. To address this issue, an effective strategy is to enhance and extract spatial information from low-level features, which can then be combined with high-level feature upsampling during the decoding process [17]. For spatial information, different objects typically display significant variations in their feature expressions, with the key aspect being the identification of edges between them. Therefore, accurately identifying interclass edges is crucial for extracting explicit spatial information, particularly for RSIs [18].

Recent studies on RSI segmentation have mainly focused on boundary supervision [19], [20], [21], [22], [23], [24], [25]

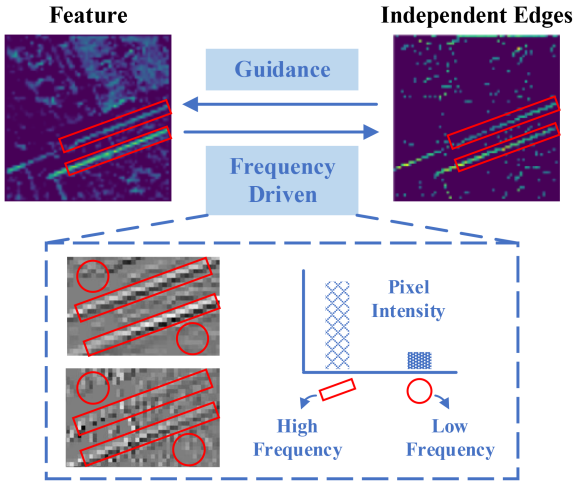


Fig. 1. By frequency-domain transformation, our proposed method generates independent edge features based on pixel intensity without additional data support.

and additional data support [26]. Typically, the ground truth for edges (edge GT) is generated from the segmentation labels to serve as a reference for the network's boundary supervision branch. While these methods partially address the issue of fuzzy boundary segmentation, the introduction of branch networks and additional data results in model bloat, increased parameters, reduced computational efficiency, and limited generality. Traditional boundary detection algorithms, such as differential operators, perform well in natural images, but have received limited attention in RSI segmentation [27], [28]. This is primarily due to their weak noise resistance.

The wavelet transform has been extensively researched in the field of frequency analysis. It has the ability to generate feature maps based on predefined rules and effectively distinguish high-frequency information from low-frequency information [29]. This means that the wavelet transform takes into account image structures, particularly the variations in pixel intensity, during feature extraction. On the other hand, CNNs abstract objects into features, decreasing their complexity and indirectly reducing the impact of noise on high-frequency information. Based on the above observation of wavelet transform and CNNs, our objective is to accurately extract edges from the feature-level space by introducing wavelet transform to improve RSI semantic segmentation. To this end, we design an edge guidance module (EGM) based on 2-D wavelet transform, detecting edges in feature-level space. This module includes a coefficient superposition method and an edge feature enhancement algorithm that adaptively emphasizes the high-frequency features reflected by boundary information from both horizontal and vertical directions. It is worth noting that we do not use any edge GT or additional data, only identify the boundaries in the features from a frequency perspective, as shown in Fig. 1. This is an essential difference from previous methods.

In this study, we propose a novel network called frequency-driven edge guidance network (FDEG-Net) that tightly couples JSCA and EGM. Context information interpreted by JSCA is scheduled across scales to the decoder. The EGM provides detailed spatial information to both the encoder and decoder in a circular manner. We conduct comprehensive ablation studies and compare FDEG-Net with both CNN-based computer vision methods and RSI-specific approaches on three different RSI datasets to demonstrate the effectiveness of FDEG-Net.

The main contributions of this study are as follows.

1) Inspired by human vision's fuzzy characteristics, we develop a JSCA module that combines local context and large sparse context. This approach allows CNNs to capture object's discriminant semantic and establish global understanding in images.
2) A novel frequency-driven edge feature extraction method is put forward. It calculates edge positions based on the high- and low-frequency representation of the wavelet coefficients in all directions. Independent edge information loops into the encoder and decoder, thus refining spatial features. This method does not utilize additional edge labels and can be independently and flexibly embedded within CNNs.
3) Proposing a FDEG-Net to improve the semantic segmentation of RSIs by tightly coupling context information and fine-grained features across different scales. Compared with 16 advanced methods, the network achieves excellent performance on three publicly available datasets including Potsdam, Vaihingen, and GID datasets.

The rest of this article is organized as follows. Section II introduces related works on semantic segmentation of RSIs. Section III elaborates the proposed FDEG-Net. In Section IV, we present a comprehensive experimental evaluation. Section V discusses this study. Finally, Section VI concludes this article.

## II. Related Works

### A. Context Interpretation for Segmentation in RSIs

The limitation of the receptive field size in CNNs restricts their ability to capture long-range context. Researchers have explored different methods to address this issue. Yu et al. [30] and Diakogiannis et al. [31] adopted a pyramid pooling module [10] to aggregate features at different scales to capture global context in RSI segmentation. Several studies [32], [33], [34], [35], [36], [37] directly utilized atrous spatial pyramid pooling (ASPP) [9], which performs well in natural images, to integrate multiscale contexts. Furthermore, ASPP-based models have been developed to improve the performance for RSI scenes [38]. For example, EaNet [18] proposed a large kernel pyramid pooling that captures multiscale contexts using hybrid asymmetric convolution with atrous rates. However, these sparsity modeling methods only establish context in limited locations. This can lead to interobject semantic confusion, where small objects can be easily obscured by more salient objects, especially when they are close together.

Another approach to capturing global context is self-attention mechanism [39], which has shown promising results [40]. Thereafter, self-attention is applied to model long-range spatial correlation in high-level features of RSIs [41], [42], [43]. For example, Zhang et al. [44] designed an adaptive ASPP that introduces self-attention modules at each scale to enhance semantic

context understanding. Although these methods prove the effectiveness of self-attention, they bring computational burden. In addition, some researchers have put forth Transformer-based approaches [45], [46], [47] but at the cost of computational resources and memory.

Compared with self-attention modeling, the context representation of dilated convolution is indeed sparse or rough. However, recent research suggests that the outstanding performance of self-attention may be attributed to its ability to capture global context through a global window [15]. This finding motivates the exploration of designing large dense kernel convolutions to capture a wide range of context.

### B. Edge Extraction for Segmentation in RSIs

Edge extraction are prominent areas of research in semantic segmentation, particularly in the context of RSIs. Some studies employ traditional image processing algorithms, such as watershed algorithm [48], to extract boundaries. Li et al. [28] proposed an edge distributed attention, which incorporates the Canny operator into self-attention to emphasize edges. Chen et al. [49] simultaneously fused the Canny results of labels and input images into the network. Notably, Azimi et al. [27] inserted the wavelet decompositions of input images into CNNs to segment lane markings in RSIs. They only simply used three wavelet decompositions of input images and did not investigate how to better highlight high-frequency information and reduce low-frequency noise in decomposition results. In this article, we directly decompose features by wavelet transform and adopt four decompositions to emphasize high-frequency information represented by boundaries and suppress noises. In addition, to aid in segmentation, researchers have also introduced additional data with boundary attributes, such as digital surface model (DSM) [26].

Recent studies have focused on optimizing the loss function and incorporating additional edge supervision. The most common approach is to perform edge supervision on the multilevel features in encoder or decoder [50], [51], [52], [53], [54], [55], [56]. Zheng et al. [18] developed an edge-aware loss function to enhance edge information directly from segmentation prediction, facilitating the separation of confusing objects with sharp contours. Sun et al. [20] proposed an adaptive edge loss that optimizes the edge-body segmentation jointly, specifically aimed at identifying tiny objects. Li et al. [21] constructed a semantic boundary awareness network, which incorporates edge ground truth (GT) and employs a multitask loss to supervise the boundaries in the encoder features. Sui et al. [23] first pretrained a boundary detection network using edge GT. Then, the network constrains the segmentation results through loss cycles and feature bootstrapping. To improve boundary prediction performance, some studies employ combination methods. Pan et al. [19] performed Canny and morphological operations to the input image, generating an edge-region map that is supervised by edge GT and embedded into the decoder to identify edges. Jin et al. [22] adopted edge GT to supervise spatial information, guiding multimodal fusion through boundary features.

It is apparent that the complexity of the aforementioned methods is increasing. In light of this, our focus is to reduce the model's reliance on edge GT and the design of the loss function. By solely utilizing the traditional differential operator, we can effectively extract high-quality boundary features.

## III. Methodology

### A. Overview of the FDEG-Net

This section presents the overall structure of FDEG-Net, as illustrated in Fig. 2. FDEG-Net follows an encoder–decoder structure that is based on a variant of U-Net. The encoder applies a ResNeXt-101 [57] as the backbone of FDEG-Net. The encoding features hierarchically incorporate the JSCA module and EGM. The JSCA module captures long-range context at high-level semantics, whereas the EGM refines high-resolution features to produce high-quality edges. Besides, the decoder adopts a cross-scale fusion strategy to gradually align the details and semantics. This strategy consists of three fusion modules, each leveraging contextual features from two scales to guide the integration of detailed features. The three features are aggregated using Conv-BN-ReLU, and a scale attention unit (SAU) is applied to consider the relative importance of features. For more information on the fusion module, please refer to the two dotted boxes on the right side of Fig. 2. Finally, bilinear interpolation is applied in the decoder to restore the spatial resolution for dense prediction.

### B. JSCA Module

When individuals focus their vision on a significant target, they tend to observe their surrounding environment in a generalized manner. However, in the case of RSIs, direct visual observation of images from this perspective is not possible. Nevertheless, at the semantic level, RSIs exhibit certain regularities in its contextual information. For example, cars are expected to be present on roads rather than on the rooftops of buildings. Similarly, buildings are typically accompanied by roads or surrounded by green vegetation.

CNNs exhibit a characteristic wherein large receptive fields allow the network to consider a broader context. On the other hand, small receptive fields are good at capturing local details. By employing dilated convolution with different atrous rates, CNNs effectively obtain a multiscale context. Therefore, utilizing local context with multiscale receptive fields enables the network to perceive the inherent contextual semantics present in RSIs.

Dilated convolution is utilized to simulate a large-window sparse context, which is then combined with local details to generate a joint sparse context. This joint sparse context is leveraged to foster the learning of long-range relationships. Consequently, a JSCA module is designed to extensively explore the contextual information present in the images. The JSCA module consists of five branches as shown in Fig. 3, and its operation process is as follows.

Let $X \in \mathbb{R}^{C\times H\times W}$ denotes the input feature map, with $C$, $H$, and $W$ representing its channel, height, and width, respectively. The subscript $i$ is the number of branches, where $i = \{1, 2, 3\}$.

First, the three main branches consist of $3 \times 3$ depthwise separable convolutions (DSConv) to compute contextual features

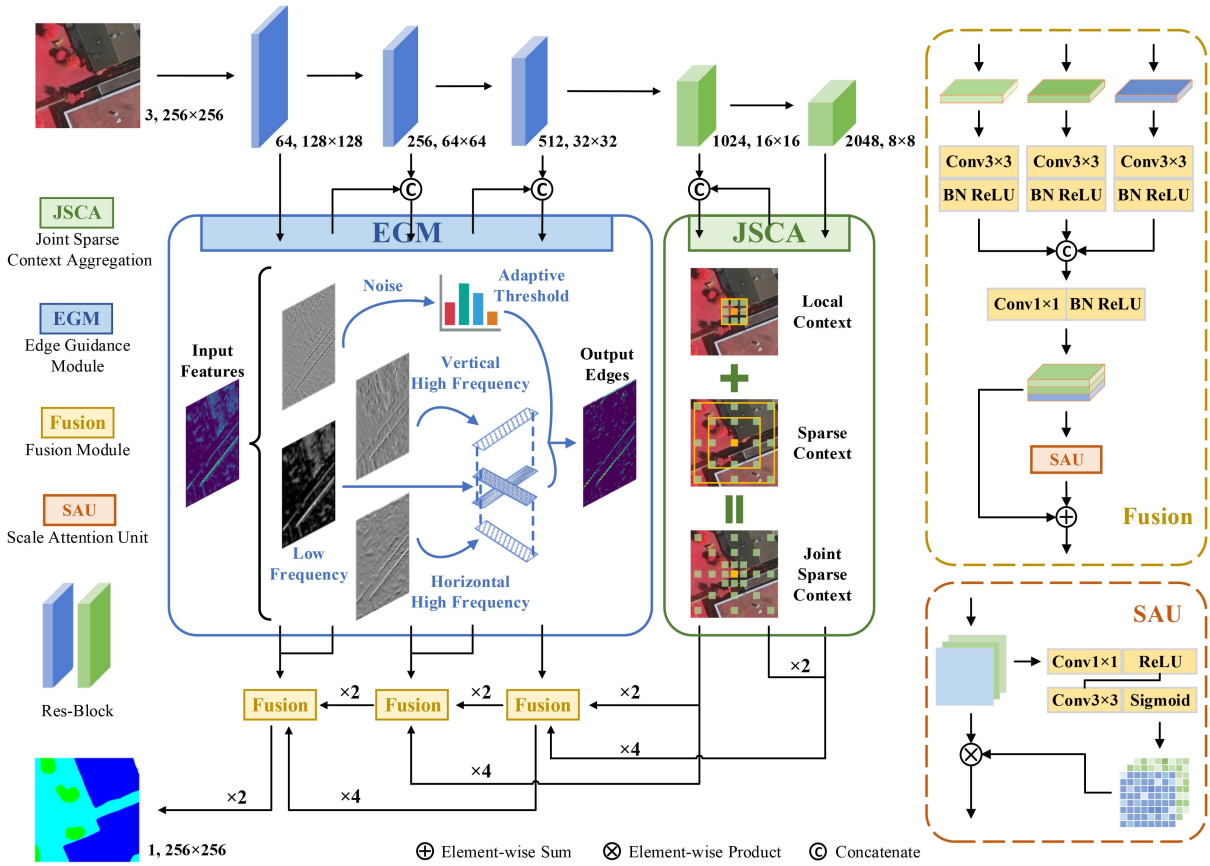


Fig. 2. Overview of the proposed FDEG-Net. The JSCA module establishes long-range associations in high-level features, whereas the EGM extracts and reinforces edge features in low-level features.

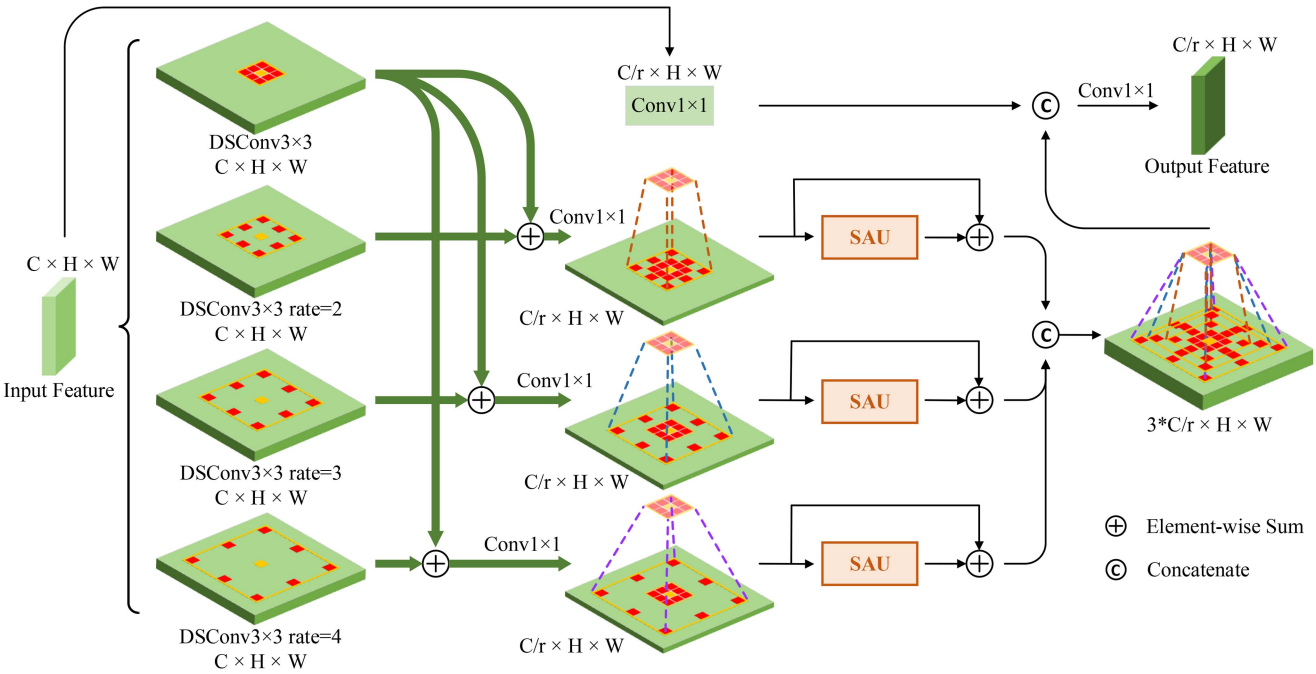


Fig. 3. Detailed design of the JSCA module. Local information is attached to each rate-scale feature to represent the joint sparse context. The SAU denotes scale attention unit the same as in Fig. 2.

$F_{\mathrm{SC}}$ of different scale. These context features can be described as follows:

$$F_{\mathrm{SC}_i} = \mathrm{DSConv}_{3\times 3}(X|a_i) \tag{1}$$

where $a$ is the atrous rate. The three atrous rates are set to 2, 3, and 4. This operation can preserve original semantics on a single channel while reducing the computational burden. Therefore, the shape of the $F_{\mathrm{SC}}$ is $C \times H \times W$. Note that the atrous rate in the other JSCA module is set to 3, 5, and 7, and the smaller rate combination is used to parse the top-level context.

The atrous rates for the two groups are set as follows. The objective of the proposed JSCA module is to capture long-range context, covering the entire feature map. On one hand, a larger rate leads to ineffective feature learning as it reduces the valid filter weights [9]. To preserve more valid filter weights, the receptive field of the maximum atrous rate should be close to the feature size. Consequently, the maximum atrous rates for the first and second groups are set to 4 and 7, respectively. Since the maximum atrous rate of the first group is 4, the other two atrous rates can only be reduced to 2 and 3. On the other hand, an improper group of atrous rates will result in a gridding problem [58]. Therefore, we set the atrous rates at two intervals to prevent gridding problem and capturing duplicate semantic information.

Subsequently, local features $F_{\mathrm{L}}$ are extracted using a standard $3 \times 3$ DSConv. They are then added to the $F_{\mathrm{SC}_i}$ pixel by pixel, resulting in the joint sparse context $F_{\mathrm{JSC}_i}$. Afterward, the $F_{\mathrm{JSC}}$ from each scale undergoes fusion using $1 \times 1$ convolution to learn the associations between local features and long-range context. The $F_{\mathrm{JSC}_i}$ is expressed as follows:

$$\begin{aligned} F_{\mathrm{JSC}_i} &= \mathrm{Conv}^{r}{}_{1\times 1}(F_{\mathrm{L}} + F_{\mathrm{SC}_i}) \\ &= \mathrm{Conv}^{r}{}_{1\times 1}\left(\mathrm{DSConv}_{3\times 3}(X) + F_{\mathrm{SC}_i}\right) \end{aligned} \tag{2}$$

where $r$ represents the dimension reduction ratio of channels, and $r = 4$.

Lastly, the SAU is employed to process the three main branches to enhance the expression of the joint sparse context for each scale. Then, they are stacked as multiscale joint sparse context $F_{\text{MS-JSC}}$, and aggregated with the top $1 \times 1$ convolution branch to generate the output features $X_{\text{out}}$. The $F_{\text{MS-JSC}}$ and $X_{\text{out}}$ are defined as follows:

$$F_{\text{MS-JSC}} = \mathrm{Concat}((F_{\mathrm{JSC}_i} + \mathrm{SAU}(F_{\mathrm{JSC}_i})) \tag{3}$$

$$X_{\text{out}} = \mathrm{Conv}^{r}{}_{1\times 1}(\mathrm{Concat}(F_{\text{MS-JSC}_i}, \mathrm{Conv}^{r}{}_{1\times 1}(X))). \tag{4}$$

Within SCEG-Net, the JCSA module is applied to the top two layers of ResNeXt-101. Although these layers have large receptive fields relative to the inputs, the JSCA can still acquire inherent contextual knowledge by utilizing scale-cropped original inputs. In summary, the JCSA module incorporates dense local kernels into large sparse kernels to proficiently capture object discriminant features and abundant long-range context.

### C. Edge Guidance Module

Since features of CNNs essentially capture the abstract mapping of corresponding objects, the final details within the objects tend to be attenuated, resulting in heightened high-frequency information at the interclass edges. Therefore, we construct a frequency-driven EGM without any parameters, as depicted in Fig. 4. This module comprises of four steps: First, the 2-D discrete wavelet transform is employed to extract frequency information in feature maps. Second, the coefficients representing the frequency are superimposed to determine the edge region, obtaining edge texture maps. Third, an adaptive threshold algorithm is applied to remove noise and enhance the edge representation in edge texture maps. Fourth, the edge texture maps are then fed into two paths: 1) They are concatenated with details at the next level. 2) They are aggregated with higher-level semantics. This allows the edge information to be reused at different resolutions, thereby enhancing the model's ability to detect edge regions. We will now provide illustrations for these four steps.

Let the input feature map be $F \in \mathbb{R}^{C\times H\times W}$. The $C$, $H$, and $W$ denote the channel, height, and width of the $F$, respectively. For each channel, assuming that

$$F_c = f_c + \varepsilon_c \tag{5}$$

where $f$ represents the clear body feature, $\varepsilon$ stands for the unrecognizable smaller feature (as noise), and $c$ is the channel number. The $f$ and $\varepsilon$ are independent of each other.

$F$'s wavelet coefficient matrix Coe is generated via a 2-D discrete wavelet transform. The transformation is as follows:

$$\mathbf{Coe}_c = \boldsymbol{\omega} F_c \boldsymbol{\omega}^T \tag{6}$$

where $\boldsymbol{\omega}$ is a standard Haar wavelet transform matrix. The Haar wavelet transform can decompose feature map components of different directions with fewer parameters and inference delay than other wavelet basis functions [59].

The $\boldsymbol{\omega}$ contains the Haar basis function of the following form:

$$\begin{cases} h_0(t) = h_{00}(t) = \frac{1}{\sqrt{H}}, & t \in [0,1] \\ h_k(t) = h_{qp}(t) = \frac{1}{\sqrt{H}} \begin{cases} 2^{\frac{q}{2}}, & \frac{p-1}{2^q} \le t < \frac{p-0.5}{2^q} \\ -2^{\frac{q}{2}}, & \frac{p-0.5}{2^q} \le t < \frac{p}{2^q} \\ 0, & \text{otherwise}, t \in [0,1] \end{cases} \end{cases} \tag{7}$$

where $t \in [0, 1]$. Since $H = W$ in the input feature $F$, the matrix size in the following formula is represented by $H$. $k = \{0, 1, 2, \ldots, H-1\}$, $H = 2^l$, and $l \in N^*$. The $p$ and $q$ represent integer factorization of $k$, and $k = 2q + p - 1$, where $0 \le p \le l-1$, $p = 0$ or 1 for $q = 0$, and $1 \le p \le 2q$ for $q \ne 0$. Each row of $\boldsymbol{\omega}$ contains the elements $h_{k,r}(t)$, which make up the 1-D vector $\vec{\boldsymbol{h}}_k$, where $r = \{0, 1, 2, \ldots, H-1\}$ is the position number, and $t = \{0, 1/H, 2/H, \ldots, (H-1)/H\}$. Therefore, $\boldsymbol{\omega}$ is the following matrix:

$$\boldsymbol{\omega} = \begin{pmatrix} \vec{\boldsymbol{h}}_0 \\ \vec{\boldsymbol{h}}_1 \\ \vdots \\ \vec{\boldsymbol{h}}_{H-1} \end{pmatrix} = \begin{pmatrix} h_{0,0}(t) & \cdots & h_{0,H-1}(t) \\ h_{1,0}(t) & \cdots & h_{1,H-1}(t) \\ \vdots & \ddots & \vdots \\ h_{H-1,0}(t) & \cdots & h_{H-1,H-1}(t) \end{pmatrix}. \tag{8}$$

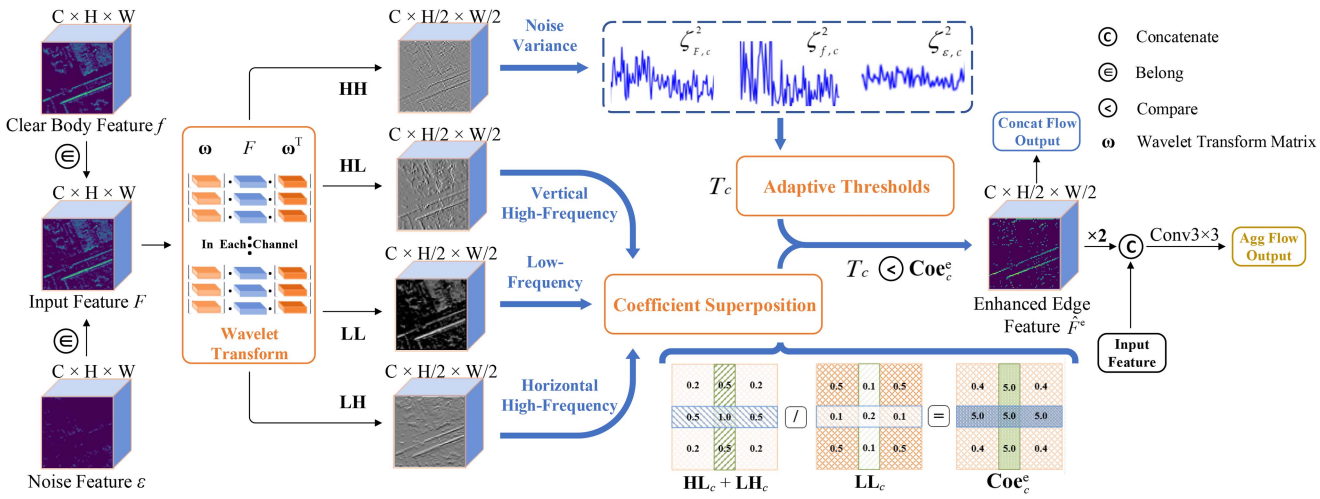


Fig. 4. Detailed design of the EGM. It leverages wavelet transform to obtain feature subbands with edge attributes. Subsequently, these subbands are used to generate adaptive thresholds and superpose coefficients, highlighting edge features.

The decomposition process of transformation can be described as follows: Initially, a 1-D discrete wavelet transform is applied to each row of the feature map, resulting in the extraction of low- and high-frequency components in the horizontal direction. Subsequently, a 1-D discrete wavelet transform is performed on each column of the transformed feature map, producing the low-frequency subband **LL** and the high-frequency subbands **LH**, **HL**, and **HH** in the vertical, horizontal, and diagonal directions. Consequently, the **Coe** can be represented by the following matrix:

$$\mathbf{Coe}_{z,c} = \begin{pmatrix} \mathbf{LL_{z,c}} & \mathbf{LH}_{z,c} \\ \mathbf{HL}_{z,c} & \mathbf{HH}_{z,c} \end{pmatrix} \tag{9}$$

where $z$ denotes the number of the wavelet transform, and $z = 1$. For simplicity, the following formulas will not be accompanied by $z$.

Each subband has a size of $(C, H/2, W/2)$. With the exception of the **LL** subband, the **HL**, **LH**, and **HH** subbands contain edge texture information and high-frequency noise. In the subbands, the magnitude of wavelet coefficients directly represents the strength of edge texture. This implies that if there are edges present in the feature map, the sum of corresponding coefficients in the **LH** and **HL** subbands should be greater than the coefficient value in nonboundary regions. In addition, the **LL** subband exhibits lower coefficient values at high-frequency locations. Therefore, based on this frequency domain characteristic, it is possible to determine whether edges exist in the corresponding feature map of the coefficient matrix. The edge texture features $\mathbf{Coe}^{\mathrm{e}}$ are determined as follows:

$$\mathbf{Coe}^{\mathrm{e}}_{c}(i,j) = \frac{\mathbf{LH}_c(i,j) + \mathbf{HL}_c(i,j)}{\mathbf{LL}_c(i,j)} \tag{10}$$

where $i$ and $j$ are the position number.

To mitigate the impact of noise, only the **HL** and **LH** subbands, which contain edge texture in the horizontal and vertical directions, are retained. There are two main reasons for this: 1) The **HL** and **LH** subbands also show significant wavelet coefficients at diagonal edge locations. This is because the Haar wavelet basis function has a small filter length (=2), which results in a small transformation window. In cases where an edge has a distinct skew or curve, it can be seen as a combination of multiple small horizontal and vertical lines. Therefore, the **HL** and **LH** subbands are effective in capturing edge features in directions other than just horizontal and vertical. 2) The **HH** subband contains more noise. Fragmented features are often irregular and are rarely presented in regular horizontal and vertical directions. Since the **HH** subband represents texture in oblique directions, it is more susceptible to capturing fragmented features and thus, more noise. The visualization effect of (10) is shown at the bottom of Fig. 4. The green and blue areas represent vertical and horizontal edges, whereas the other areas correspond to low-frequency information. Obviously, by performing matrix division, the edge region is effectively highlighted. At this point, a feature map containing edge information of moderate intensity has been successfully extracted.

The high-frequency information is retained and sharpened when the features are translated into the frequency domain space. However, high-frequency noise independent of the edge is also embedded. Therefore, based on the wavelet threshold theory [60], an adaptive edge feature enhancement algorithm is designed to highlight the wavelet coefficients on the edge region. The pseudocode is outlined in Algorithm 1.

It is worth noting that the whole process from feature input to Algorithm 1 does not introduce learnable parameters, so it is necessary to learn edge information through feature aggregation. Specifically, $\hat{F}^{\mathrm{e}}$ is fed to the two types of pipeline flows to provide edge guidance. In the concatenating flow, it is utilized to help determine the position of the object in neighboring layer features $F_{\mathrm{al}}$. In the aggregation flow, its guidance aligns the details F and two high-level semantics ($F_{\mathrm{s1}}$ and $F_{\mathrm{s2}}$). The two types of guidance are described as follows:

$$\begin{cases} \text{Concat flow: Concat}(\hat{F}^{\mathrm{e}}, F_{\mathrm{al}}) \\ \text{Agg flow: Concat}(\text{Concat}(\hat{F}^{\mathrm{e}}, F), F_{\mathrm{s1}}, F_{\mathrm{s2}}) \end{cases}. \tag{11}$$

**Algorithm 1:** Adaptive Edge Feature Enhancement Algorithm.

**Input:** Subband **HH** $\in \mathbb{R}^{C\times(H/2)\times(W/2)}$ and edge texture features $\mathbf{Coe}^{\mathrm{e}}$
**Output:** Enhanced edge features $\hat{F}^{\mathrm{e}}$
# Line 1-15: Adaptive modeling on **HH**. Let $\zeta_{F,c}$, $\zeta_{f,c}$, $\zeta_{\varepsilon,c}$ denote the standard deviation of $F_c$, $f_c$, $\varepsilon_c$.
1: According to (5), so $\zeta^2_{F,c} = \zeta^2_{f,c} + \zeta^2_{\varepsilon,c}$.
2: **for** each channel $HH_c \in$ **HH do**
3: **for** all pixel $h_{i,j} \in HH_c$ **do**
4: sort($h_{i,j}$), from the smallest to largest
5: $h_{\mathrm{med}}$ = median(sort($h_{i,j}$)), and $h_{\max}$ = max(sort($h_{i,j}$))
6: Compute noise standard deviation: $\zeta_{\varepsilon,c} = h_{\mathrm{med}}/0.6745$
7: Compute variance of $F$: $\zeta^2_{F,c} = \frac{\sum_{i=1,j=1}^{H/2} h^2_{i,j}}{(H/2)\times(W/2)}$
8: **end for**
9: Compute variance of $f$: $\zeta^2_{f,c} = \sqrt{\max(\zeta^2_{F,c} - \zeta^2_{\varepsilon,c}, 0)}$
10: **if** $\zeta^2_{\varepsilon,c} \geq \zeta^2_{F,c}$ **then**
11: $\zeta^2_{f,c} = 0$, but compute adaptive threshold $T_c = h_{\max}$
12: **else**
13: Compute adaptive threshold $T_c = \zeta^2_{\varepsilon,c}/(2 \times \zeta^2_{f,c})$
14: **end if**
15:**end for** # Adaptive threshold $T$ are obtained.
# Line 16-24: Strong edge modeling on $\mathbf{Coe}^{\mathrm{e}}$.
16:**for** each channel $\mathbf{Coe}_c \in \mathbf{Coe}^{\mathrm{e}}$ **do**
17: **for** all pixel $x_{i,j} \in \mathbf{Coe}_c$ **do**
18: **if** $x_{i,j} > T_c$ **then**
19: $\hat{F}^{\mathrm{e}}(i,j) = x_{i,j}$
20: **else**
21: $\hat{F}^{\mathrm{e}}(i,j) = 0$
22: **end if**
23: **end for**
24:**end for** # Enhanced edge features $\hat{F}^{\mathrm{e}}$ are obtained.

## IV. Experiment

### A. Datasets and Training Details

The experiments are conducted on three datasets: Potsdam and Vaihingen of ISPRS Remote Sensing Image Segmentation Challenge and the Gaofen Image Dataset (GID) [61].

*Potsdam dataset:* The Potsdam dataset consists of 38 high-resolution RSIs that cover the northeastern region of Germany. These images showcase a diverse urban landscape, including buildings, streets, and vegetation. Each image has a size of 6000 × 6000 pixels and has a spatial resolution of 0.05 m. The dataset provides six bands including red, green, blue, near infrared, DSM, and normalized DSM. Note that the experiments only utilize the first three bands. In addition, the labels are divided into six categories: impervious surfaces, buildings, low vegetation, trees, cars, and background. For testing, the chosen images include IDs 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15, and 7_13. ID 2_10 is selected as the validation set. The training phase utilizes the remaining 22 images, excluding ID 7_10 due to a labeling error.

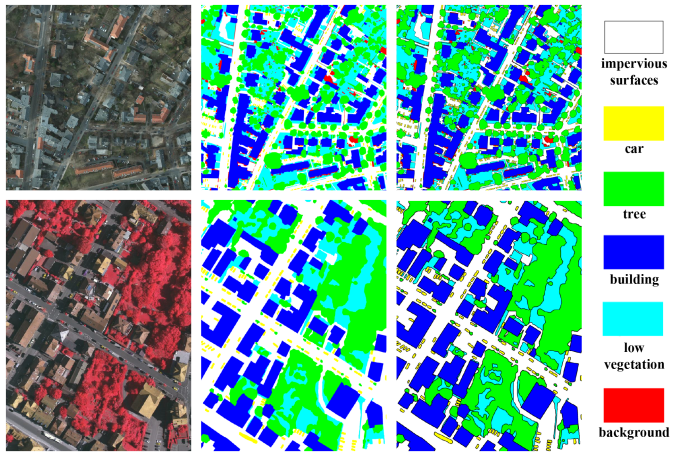


Fig. 5. Potsdam and Vaihingen datasets. The second and third columns are GTs without and with erosion boundaries.

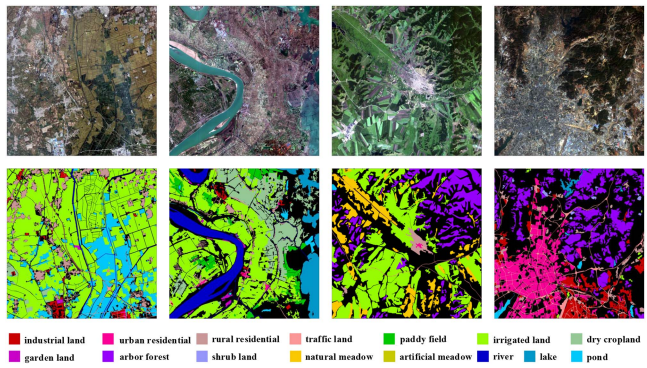


Fig. 6. GID dataset. The first row is images. The second row is GTs.

*Vaihingen dataset:* The Vaihingen dataset contains 33 images captured in central Germany, showcasing typical village characteristics such as compact buildings and lush vegetation. The average size of the image is 2494 × 2064 pixels, with a spatial resolution of 0.09 m. The dataset provides four bands: red, green, near infrared, and DSM. The first three bands are utilized in the experiments. In addition, the dataset follows the same category classification as the Potsdam dataset. The test set consists of images with IDs 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, and 38. Image ID 30 is selected as the validation set. The training set is constructed using the remaining 15 images.

Examples of the above two datasets are shown in Fig. 5. The GT is classified into two categories: nonerosive boundary and erosive boundary. For the latter, black pixels with a circular disc of a 3-pixel radius are added at the object edge to mitigate the influence of uncertain border definitions during evaluation. It is important to note that, except for a proving test, all experiments in this article utilize ground reference data without erosion boundary. Consequently, the experimental results are lower when compared with those obtained using label data with erosion boundary.

*GID:* As illustrated in Fig. 6, the GID comprises 10 RGB images of China, capturing 15 land use categories. These images were collected by the Gaofen-2 satellite. Each image

TABLE I
RESULTS OF DIFFERENT MODELS ON THE POTSDAM TEST SET

| Model | Building | Imp. Surfaces | Low Veg. | Tree | Car | MIoU | AF | OA | K |
|---|---|---|---|---|---|---|---|---|---|
| U-Net | 84.068 / 89.512 | 78.041 / 83.914 | 68.599 / 79.766 | 67.680 / 80.131 | 75.080 / 85.938 | 74.694 | 83.852 | 83.544 | 79.694 |
| PSPNet | 87.277 / 92.335 | 79.858 / 86.863 | 69.961 / 81.014 | 71.309 / 82.930 | 75.875 / 83.790 | 76.856 | 85.386 | 85.384 | 81.752 |
| DeepLabV3+ | 88.114 / 92.812 | 81.382 / 87.851 | 70.643 / 81.740 | 72.049 / 82.742 | 78.587 / 87.352 | 78.155 | 86.499 | 86.334 | 82.554 |
| DANet | 88.729 / 93.426 | 82.558 / 89.774 | 70.728 / 82.785 | 72.662 / 83.683 | 80.782 / 88.743 | 79.092 | 87.682 | 87.102 | 83.118 |
| OCRNet | 88.832 / 92.689 | **84.292** / 89.352 | 71.187 / 83.362 | 71.562 / 82.467 | 80.468 / 88.413 | 79.268 | 87.257 | 86.792 | 82.883 |
| SCAttNet | 88.754 / 92.563 | 81.765 / 88.362 | 71.472 / 82.684 | 70.758 / 82.476 | 79.312 / 87.496 | 78.412 | 86.716 | 86.441 | 82.624 |
| MANet | 89.825 / 94.468 | 83.174 / 90.335 | 73.285 / 84.345 | 73.378 / 84.224 | 82.772 / 90.270 | 80.487 | 88.728 | 87.632 | 83.671 |
| EaNet | 90.192 / 94.843 | 83.076 / 90.756 | 72.599 / 84.124 | 72.495 / 84.055 | 84.355 / 91.514 | 80.543 | 89.058 | 87.765 | 83.873 |
| LANet | 89.455 / 94.434 | 81.839 / 90.013 | 71.293 / 83.241 | 71.078 / 83.095 | 81.118 / 89.575 | 78.957 | 88.072 | 87.035 | 82.892 |
| BANet | 89.452 / 94.433 | 82.221 / 90.243 | 72.674 / 84.175 | 72.494 / 84.054 | 82.055 / 90.143 | 79.779 | 88.610 | 87.118 | 83.096 |
| DCST | 88.603 / 93.476 | 80.311 / 89.577 | 71.920 / 83.460 | 71.016 / 83.012 | 81.331 / 89.901 | 78.636 | 87.885 | 86.497 | 80.837 |
| UNetFormer | 90.514 / 94.931 | 83.247 / 90.750 | 73.266 / 84.533 | 73.619 / 84.631 | 84.516 / 91.271 | 81.032 | 89.223 | 88.043 | 84.169 |
| MAResU-Net | 90.120 / 94.803 | 82.775 / 90.576 | 72.698 / 84.191 | 73.108 / 84.465 | 82.883 / 90.640 | 80.317 | 88.935 | 87.826 | 83.979 |
| MACU-Net | 85.696 / 92.297 | 79.284 / 88.445 | 70.193 / 82.486 | 67.461 / 80.569 | 78.099 / 87.703 | 76.147 | 86.300 | 85.006 | 80.184 |
| $A^2$-FPN | 89.572 / 94.499 | 82.374 / 90.335 | 73.006 / 84.397 | 72.473 / 84.040 | 81.827 / 90.005 | 79.850 | 88.655 | 87.636 | 83.697 |
| BESNet | 90.472 / 94.550 | 83.252 / 90.651 | 73.389 / 84.186 | 73.601 / 84.393 | 83.301 / 91.095 | 80.803 | 88.975 | 87.757 | 83.871 |
| HBCNet | 88.867 / 94.028 | 82.905 / 90.344 | 73.181 / 84.045 | 72.882 / 84.314 | 82.860 / 90.351 | 80.139 | 88.616 | 87.474 | 83.495 |
| GCDNet | 90.485 / 94.590 | 82.929 / 90.554 | 72.692 / 83.956 | 72.440 / 83.865 | 81.930 / 90.118 | 80.095 | 88.617 | 87.561 | 83.543 |
| CMTFNet | 90.444 / 94.921 | 83.633 / 90.746 | 73.687 / 84.676 | 73.657 / 84.626 | 84.194 / 91.402 | 81.123 | 89.274 | 87.985 | 84.168 |
| FDEG-Net(Ours) | **91.007 / 95.123** | 83.643 / **91.085** | **73.728 / 84.765** | **74.034 / 85.201** | **84.778 / 91.753** | **81.438** | **89.585** | **88.362** | **84.713** |

(The black bold indicates the highest, and the underline indicates the second highest). Accuracy of each category is presented in the IoU / F1 Form.

spans a geographic area of 506 $km^2$, with a resolution of 7200 $\times$ 6800 pixels. To facilitate training, we partition each image into nonoverlapping patches of size 256 $\times$ 256. Moreover, we randomly allocate 60% of the patches for training, 20% for validation, and 20% for testing.

Data augmentation applied to the original training images encompassed cropping them into nonoverlapping 256 $\times$ 256 blocks, performing random rotations (90°, 180°, 270°), applying random scale cropping (scaling factor: 0.2–1.0), horizontal flipping, and vertical flipping. In addition, the models are evaluated directly using the 256 $\times$ 256 test images.

For training, the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) is employed, using an initial learning rate of 0.0001. The training process consists of 300 epochs, with the learning rate halved every 30 epochs. In addition, a batch size of 16 is selected and the model is trained using the cross-entropy loss function. PyTorch is utilized as the DL framework for all experiments conducted on a single NVIDIA TITAN XP GPU with 12 GB of RAM.

### B. Evaluation Metrics

To quantitatively assess the performance of the proposed model, the evaluation of the results is based on the four metrics: overall accuracy (OA), intersection over union (IoU), Kappa coefficient (K), and F1-score (F1). AF means the average F1. The calculation methods are as follows:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}} \tag{12}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{13}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{14}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{15}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{16}$$

$$\mathrm{K} = \frac{\mathrm{OA} - p_e}{1 - p_e} \tag{17}$$

where TP, TN, FP, and FN represent the number of true positive, true negative, false positive, and false negative predicted pixels. The $p_e$ is the hypothetical probability of chance agreement and

$$p_e = \frac{(\mathrm{TP}+\mathrm{FP}) \times (\mathrm{TP}+\mathrm{FN}) \times (\mathrm{FP}+\mathrm{TN}) \times (\mathrm{FN}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN})^2}. \tag{18}$$

### C. Experimental Results

To demonstrate the effectiveness of the proposed network, a comprehensive analysis is performed comparing it with various advanced segmentation models, namely, U-Net [5], PSPNet [10], DeepLabV3+ [17], DANet [40], OCRNet [62], EaNet [18], SCAttNet [63], MANet [41], LANet [8], BANet [12], DCST [13], UNetFormer [45], MAResU-Net [11], MACU-Net [64], $A^2$-FPN [42], BESNet [52], HBCNet [51], GCDNet [54], and CMTFNet [46]. For a fair comparison, these models with the ResNet family backbone adopt ResNeXt-101. In addition, unless otherwise specified, all quantitative experimental results are presented as percentages (%).

*Quantitative analysis:* The precision results on the three datasets are presented in Tables I–III. In these tables, the black bold font indicates the highest accuracy, whereas the underlined figures indicate the second highest accuracy. Traditional baselines for natural images perform poorly when applied to RSIs with wide fields of view and complex scenes. Sparse context modeling methods such as PSPNet and DeepLabV3+

TABLE II
RESULTS OF DIFFERENT MODELS ON THE VAIHINGEN TEST SET

| Model | Building | Imp. Surfaces | Low Veg. | Tree | Car | MIoU | AF | OA | K |
|---|---|---|---|---|---|---|---|---|---|
| U-Net | 81.723 / 88.333 | 75.720 / 86.182 | 63.213 / 77.461 | 74.674 / 85.501 | 48.989 / 67.100 | 68.864 | 80.915 | 83.724 | 79.487 |
| PSPNet | 83.064 / 87.696 | 75.885 / 86.742 | 63.574 / 77.319 | 74.524 / 84.891 | 55.481 / 74.154 | 70.506 | 82.160 | 83.990 | 79.820 |
| DeepLabV3+ | 83.362 / 88.159 | 77.341 / 87.562 | 63.657 / 78.013 | 75.512 / 85.331 | 57.631 / 75.483 | 71.501 | 82.910 | 85.325 | 81.121 |
| DANet | 84.728 / 90.397 | 78.661 / 88.253 | 64.567 / 78.366 | 75.734 / 85.871 | 62.496 / 76.838 | 73.237 | 83.945 | 85.981 | 81.556 |
| OCRNet | 85.360 / 91.688 | 78.796 / 88.596 | 64.358 / 78.985 | 75.629 / 86.108 | 63.642 / 76.451 | 73.557 | 84.366 | 86.025 | 81.681 |
| SCAttNet | 84.879 / 90.362 | 78.596 / 87.552 | 63.413 / 77.965 | 75.332 / 85.058 | 63.159 / 75.478 | 73.076 | 83.283 | 85.287 | 81.114 |
| MANet | 86.509 / 92.767 | 79.880 / 88.815 | 65.497 / 79.152 | 76.486 / 86.676 | 65.356 / 79.049 | 74.746 | 85.292 | 86.653 | 82.373 |
| EaNet | 86.672 / 92.860 | 79.820 / 88.777 | 65.624 / 79.244 | 76.511 / 86.693 | 64.030 / 78.071 | 74.531 | 85.129 | 86.886 | 82.652 |
| LANet | 84.666 / 91.696 | 78.613 / 88.026 | 64.064 / 78.097 | 75.145 / 85.809 | 53.956 / 70.093 | 71.289 | 82.744 | 85.438 | 81.240 |
| BANet | 85.715 / 92.308 | 79.249 / 88.423 | 65.375 / 79.063 | 76.189 / 86.486 | 61.387 / 76.074 | 73.583 | 84.471 | 86.392 | 81.989 |
| DCST | 85.343 / 91.224 | 78.873 / 88.156 | 63.863 / 78.530 | 75.269 / 85.343 | 62.731 / 75.861 | 73.216 | 83.829 | 86.255 | 81.583 |
| UNetFormer | 86.864 / 92.970 | 80.245 / 88.855 | 65.936 / 79.472 | 76.267 / 86.461 | 66.676 / 79.503 | 75.198 | 85.452 | 86.891 | 82.657 |
| MAResU-Net | 86.497 / 92.760 | 80.172 / 88.995 | 64.920 / 78.729 | 76.177 / 86.478 | 63.481 / 77.662 | 74.249 | 84.925 | 86.482 | 82.164 |
| MACU-Net | 83.267 / 90.870 | 78.109 / 87.709 | 63.694 / 77.821 | 74.936 / 85.672 | 57.240 / 74.806 | 71.449 | 83.376 | 85.064 | 80.857 |
| $A^2$-FPN | 86.679 / 92.864 | 79.635 / 88.663 | 65.481 / 79.140 | 76.049 / 86.395 | 64.294 / 78.267 | 74.428 | 85.066 | 86.580 | 82.271 |
| BESNet | 86.087 / 92.523 | 79.115 / 88.340 | 64.851 / 78.678 | 76.180 / 85.864 | 63.503 / 77.626 | 73.947 | 84.606 | 86.474 | 82.126 |
| HBCNet | 85.104 / 91.953 | 78.626 / 88.034 | 64.165 / 78.172 | 75.797 / 86.232 | 63.687 / 77.699 | 73.476 | 84.418 | 86.096 | 81.600 |
| GCDNet | 85.971 / 92.309 | 79.271 / 88.375 | 64.903 / 78.608 | 75.842 / 85.504 | 63.420 / 77.323 | 73.881 | 84.424 | 86.257 | 81.511 |
| CMTFNet | 86.793 / 92.664 | 79.862 / 88.600 | 65.666 / 79.381 | 76.606 / **86.793** | 65.200 / 78.623 | 74.825 | 85.212 | 86.704 | 82.416 |
| FDEG-Net(Ours) | **87.103 / 93.185** | **80.557 / 89.217** | **66.071 / 79.583** | **76.677** / 86.740 | **67.002 / 80.061** | **75.482** | **85.757** | **87.073** | **82.868** |

(The black bold indicates the highest, and the underline indicates the second highest). Accuracy of each category is presented in the IoU / F1 Form.

TABLE III
RESULTS OF DIFFERENT MODELS ON THE GID TEST SET

| Model | MIoU | AF | OA | K |
|---|---|---|---|---|
| U-Net | 57.842 | 74.092 | 78.717 | 74.498 |
| PSPNet | 63.077 | 76.215 | 81.505 | 76.773 |
| DeepLabV3+ | 63.697 | 76.740 | 81.632 | 76.854 |
| DANet | 63.358 | 76.497 | 81.619 | 76.957 |
| OCRNet | 63.228 | 76.425 | 81.619 | 76.836 |
| SCAttNet | 63.174 | 75.942 | 81.568 | 76.498 |
| MANet | 62.801 | 76.932 | 81.865 | 77.238 |
| EaNet | 63.667 | 76.987 | 82.643 | 78.181 |
| LANet | 63.177 | 76.275 | 82.208 | 77.637 |
| BANet | 62.960 | 77.214 | 82.275 | 77.654 |
| DCST | 63.790 | 76.547 | 81.610 | 77.345 |
| UNetFormer | 64.396 | 76.808 | 82.537 | 78.074 |
| MAResU-Net | 64.126 | 77.001 | 82.512 | 78.173 |
| MACU-Net | 63.601 | 76.289 | 81.832 | 77.136 |
| $A^2$-FPN | 64.629 | 77.641 | 82.291 | 77.747 |
| BESNet | 64.113 | 78.231 | 82.539 | 78.241 |
| HBCNet | 63.458 | 77.143 | 81.857 | 77.269 |
| GCDNet | 63.392 | 76.804 | 81.741 | 77.068 |
| CMTFNet | 64.408 | 77.231 | 82.697 | 78.196 |
| FDEG-Net(Ours) | **65.436** | **79.861** | **83.463** | **79.039** |

The black bold indicates the highest, and the underline indicates the second highest.

lack the ability to focus on the current object semantics, making them unsuitable for RSIs. DANet and OCRNet, which introduce self-attention mechanisms, show significant improvements in accuracy.

Recent CNN-based RSI-specific methods utilize multiscale feature fusion and attention mechanisms to effectively improve segmentation results, such as MAResU-Net, $A^2$-FPN, and CMTFNet. However, due to insufficient boundary information, $A^2$-FPN, with the second-highest accuracy on the GID dataset, falls behind our method by 0.807% and 2.220% in MIoU and AF, as shown in Table III. In addition, Transformer-based methods show a strong segmentation ability. UNetFormer and CMTFNet achieve suboptimal accuracy on the Potsdam and Vaihingen datasets, as shown in Tables I and II. Similarly, due to better boundary protection, our method outperforms UNetFormer by 0.544% and 0.305% in K on the Potsdam dataset and AF on the Vaihingen dataset. On the other hand, methods with edge awareness, such as BESNet, HBCNet, and GCDNet, aim to improve accuracy by edge supervision. However, these methods heavily rely on the quality of the edge GT. In contrast, the proposed EGM generates independent and distinct edge features, which can be recycled to provide multiscale clues for locating edges. As a result, our method achieves higher precision compared with the above methods with edge supervision. In short, our proposed FDEG-Net combines context modeling (JSCA) and boundary protection strategy (EGM) simultaneously to tightly couple context and spatial information with clear boundaries. Therefore, FDEG-Net improves the intraclass consistency and interclass discriminability of predictions, resulting in better accuracy performance.

*Qualitative analysis:* We selected 15 different scenarios from the three datasets to visually compare the segmentation results of various models. The segmentation advantage of our proposed model can be observed in red box area of each figure. Fig. 7 presents the results on the Potsdam dataset. Our proposed network outperforms other models in capturing object shapes (in the first and second images) and identifying differences between classes (in the third image). In the fourth and fifth images, due to the dendritic characteristics of trees, locating the interclass edge between the tree and other categories is challenging. Most contrast methods generate fake edges in these scenarios. However, our network, which incorporates the proposed EGM, effectively captures high-frequency edge information resulting from interclass differences, achieving more accurate segmentation results.

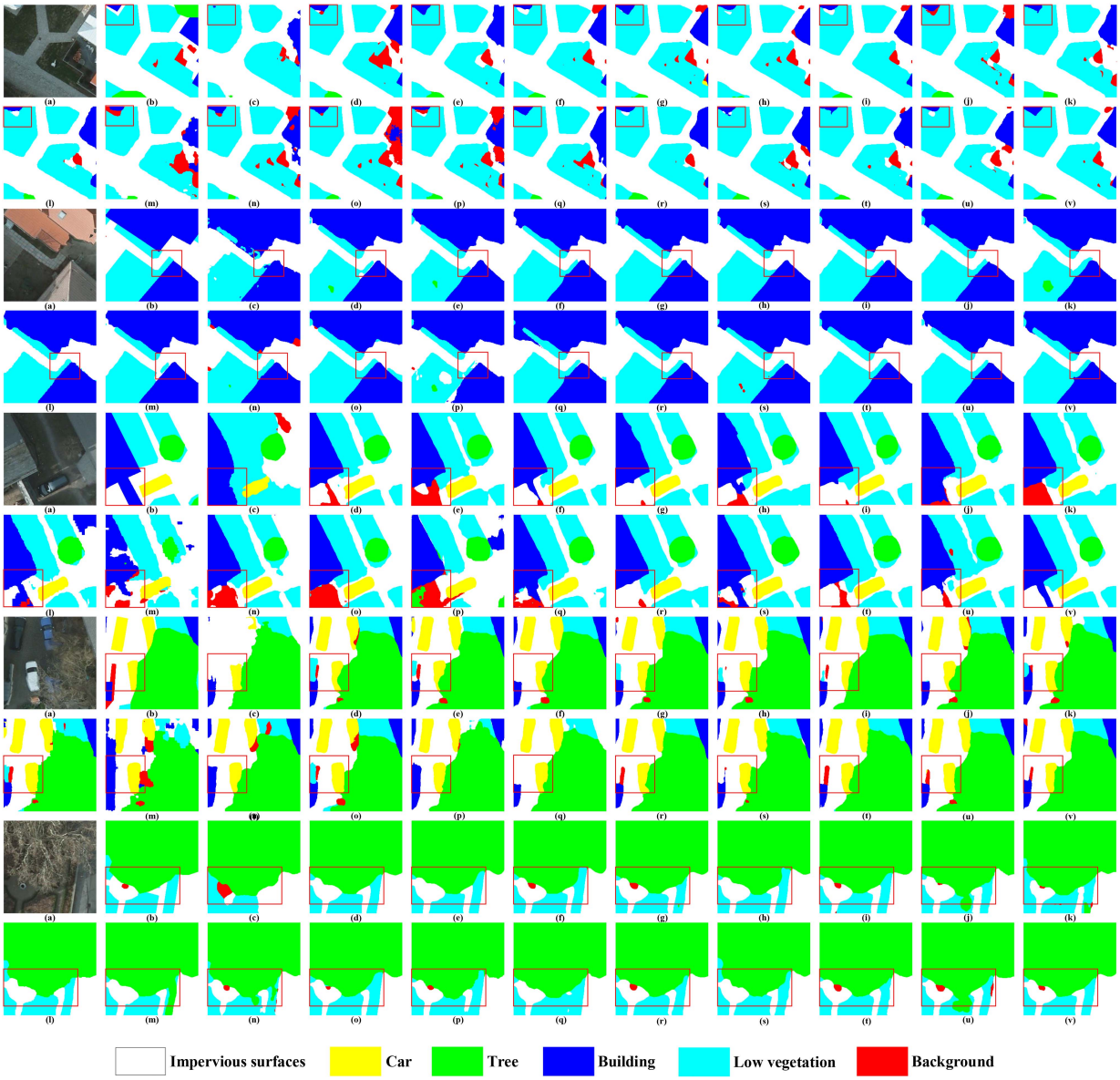


Fig. 7. Segmentation results on the Potsdam dataset. (a) Image. (b) GT. (c) U-Net. (d) PSPNet. (e) DeepLabv3+. (f) DANet. (g) OCRNet. (h) SCAttNet. (i) MANet. (j) EaNet. (k) LANet. (l) BANet. (m) DCST. (n) UNetFormer. (o) MAResU-Net. (p) MACU-Net. (q) $A^2$-FPN. (r) BESNet. (s) HBCNet. (t) GCDNet. (u) CMTFNet. (v) FDEG-Net (Ours).

Fig. 8 showcases the segmentation results on the Vaihingen dataset. Our method demonstrates its strengths in accurately recognizing multiscale objects (in the first and second images) and avoiding interference from shadows (in the third, fourth, and fifth images). This is achieved through the incorporation of the JSCA module in the semantic space, which integrates sparse large kernel convolutions with long-range awareness into the local context. As a result, the network learns the surrounding context information of the object, thereby allowing it to alleviate the negative impact of shadows and occlusions and ultimately produce superior segmentation results.

For the GID, as shown in Fig. 9, benefiting from the discrimination ability of the EGM for the interclass edges, the proposed FDEG-Net can effectively handle objects with striped patterns (such as narrow rivers and roads in the first, third, and fourth images) and rugged contours (contour lines in the second image). In addition, the JSCA module incorporates local information into wider receptive fields to emphasize the current semantic, allowing the model to reduce interference from distant salient objects on the current object. This results in improved segmentation, as seen in the third and fifth images). In comparison, other methods exhibit varying degrees of edge distortions and misclassifications.

*Efficiency analysis:* In addition to accuracy, we have also conducted an efficiency analysis of the aforementioned models, encompassing training time, inference time, parameters, and computational complexity floating point operations (FLOPs). As shown in Table IV, FDEG-Net demonstrates significantly lower time requirements for both training and inference compared with EaNet and UNetFormer, while delivering comparable

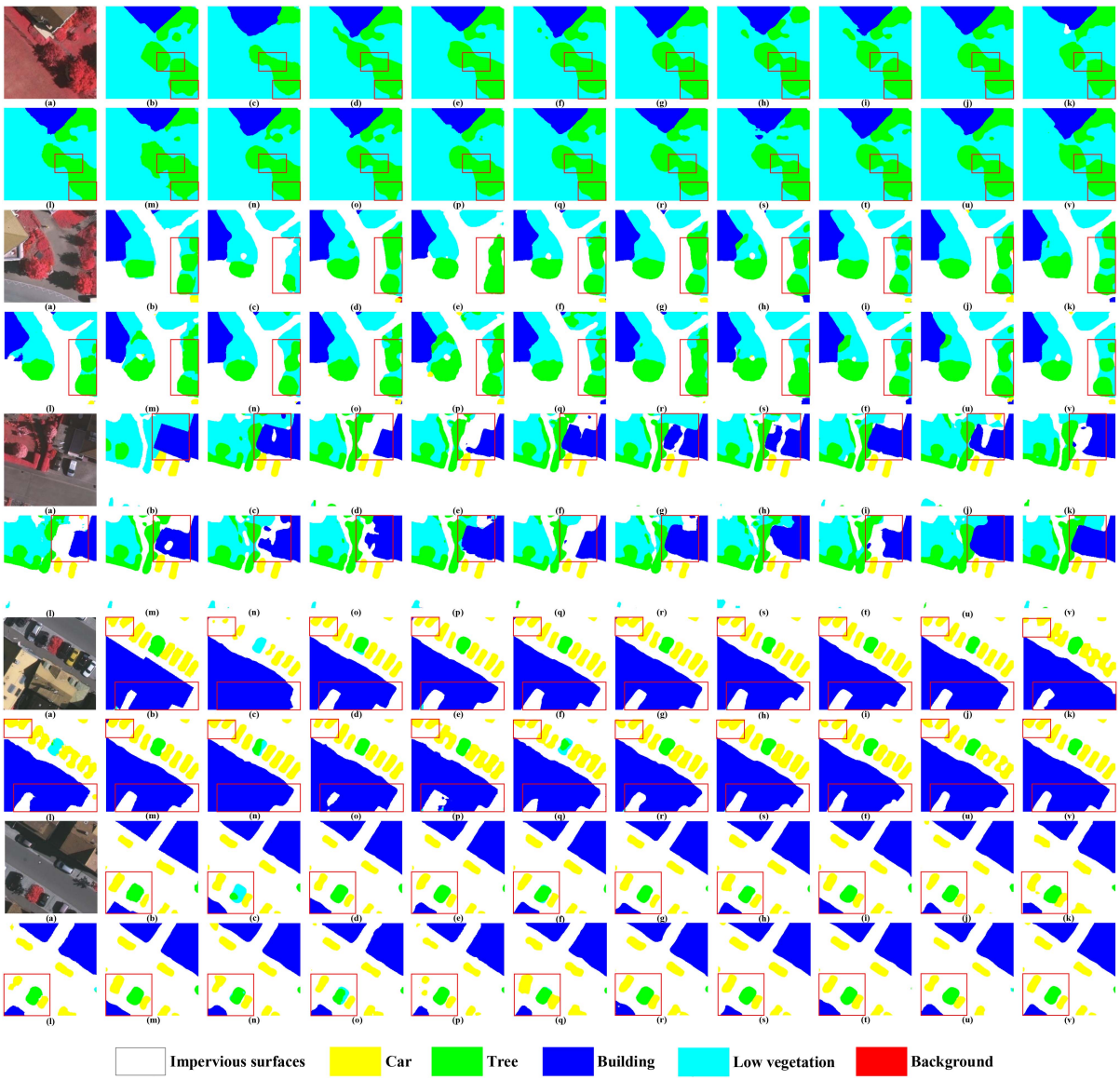


Fig. 8. Segmentation results on the Vaihingen dataset. (a) Image. (b) GT. (c) U-Net. (d) PSPNet. (e) DeepLabv3+. (f) DANet. (g) OCRNet. (h) SCAttNet. (i) MANet. (j) EaNet. (k) LANet. (l) BANet. (m) DCST. (n) UNetFormer. (o) MAResU-Net. (p) MACU-Net. (q) $A^2$-FPN. (r) BESNet. (s) HBCNet. (t) GCDNet. (u) CMTFNet. (v) FDEG-Net (Ours).

performance to MANet. In addition, compared with the BESNet, HBCNet, and GCDNet based on edge supervision, FDEG-Net shows different degrees of advantages. In short, FDEG-Net's memory consumption and FLOPs are reasonable when compared with other RSI-specific methods.

### D. Ablation Study

*Quantitative analysis of JSCA module and EGM:* To evaluate the effectiveness of the JSCA module and EGM, we conducted extensive ablation experiments using the same hyperparameter settings and runtime environment. The baseline model used for comparison was obtained by removing the JSCA module and EGM from FDEG-Net.

The results in Tables V and VI show that incorporating JSCA improves the baseline model's ability to capture a broader context, resulting in an improvement of approximately 0.2% (OA) across the three datasets, with a slight increase in training time. Notably, the inclusion of JSCA yields a significant increase of 0.967% and 0.601% in MIoU on the Vaihingen and GID datasets. While JSCA enhances the ability of context awareness, object localization depends on the guidance provided by edge information. By combining EGM with the aforementioned components, we observed further improvements compared with the baseline. There are increases of 0.472% and 0.599% in OA and MIoU on the Potsdam dataset, and 0.899% and 1.456% on the Vaihingen dataset. The two metrics also improve by 0.883% and 3.502% on the GID dataset. Notably, these improvements require very little parameter increase. In addition, the EGM performs computations for each channel in the features, including matrix multiplication, addition, and potential matrix reshaping operations. This requires a certain amount of computation and

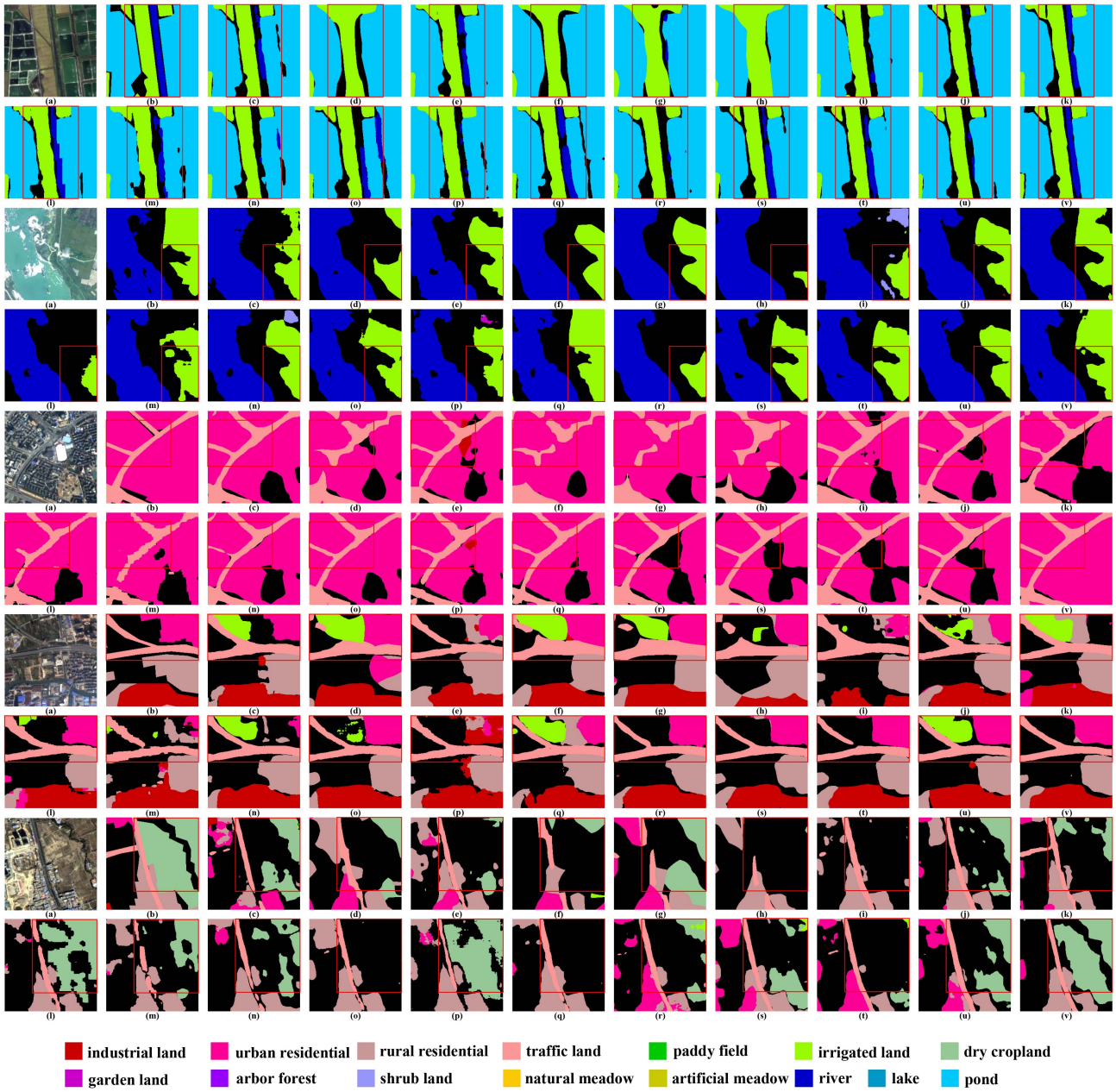


Fig. 9. Segmentation results on the GID dataset. (a) Image. (b) GT. (c) U-Net. (d) PSPNet. (e) DeepLabv3+. (f) DANet. (g) OCRNet. (h) SCAttNet. (i) MANet. (j) EaNet. (k) LANet. (l) BANet. (m) DCST. (n) UNetFormer. (o) MAResU-Net. (p) MACU-Net. (q) $A^2$-FPN. (r) BESNet. (s) HBCNet. (t) GCDNet. (u) CMTFNet. (v) FDEG-Net (Ours).

memory space, which affects the training time. In contrast, JSCA is primarily characterized by a large number of convolution operations. Therefore, JSCA mainly brings an increase in parameters, whereas EGM primarily affects FLOPs and training time. In summary, these results demonstrate the effectiveness of each proposed component in improving RSI semantic segmentation.

*Qualitative analysis of EGM:* To further demonstrate the effectiveness of EGM in edge guidance, we visualized the heatmap results of models with and without EGM on the Potsdam and GID datasets. In addition, we displayed the features of ten channels on the backbone network and the features extracted by the EGM.

The results for the Potsdam dataset are presented in Fig. 10. In the case of small objects, such as cars, the baseline model equipped with EGM demonstrates better ability to detect gaps between cars. The feature visualizations in the last row reveal that the EGM weakens the smoothing effect and enhances the expression of car edge information. For large-scale objects such as buildings and trees, EGM outperforms the baseline model without it in capturing edge features, thereby demonstrating its robustness to lines and curves. Low vegetation and impervious surfaces often have irregular shapes and shading, but the baseline model with EGM is less affected by these irregularities. This is achieved by converting features to the frequency domain,

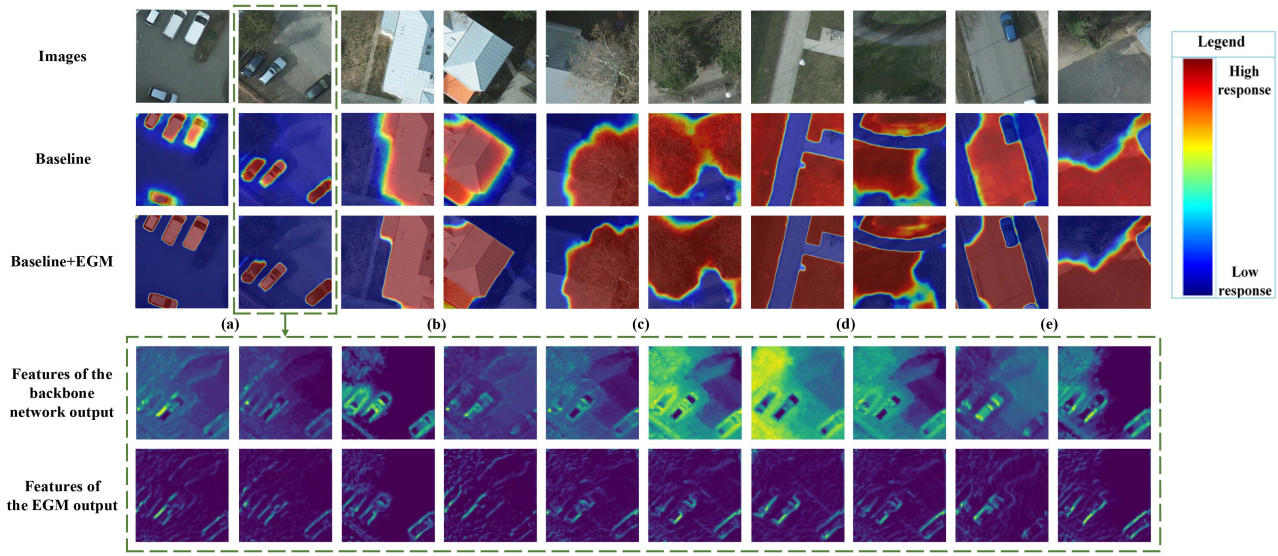


Fig. 10. Visualization of heatmap and features on the Potsdam dataset. Edge extraction results are clearly listed on the last row. (a) Car. (b) Building. (c) Tree. (d) Low vegetation. (e) Impervious surface.

TABLE IV
COMPARISON OF MODEL EFFICIENCY

| Model | Train Time (s) / epoch | Inference Time(s) | Parameters (M) | FLOPs (GMac) |
|---|---|---|---|---|
| U-Net | 56.4 | 26.3 | 34.53 | 65.55 |
| PSPNet | 36.6 | 17.2 | 109.84 | 22.82 |
| DeepLabV3+ | 55.2 | 38.4 | 105.83 | 27.87 |
| DANet | 35.7 | 16.3 | 112.72 | 23.09 |
| OCRNet | 36.5 | 15.2 | 103.81 | 23.88 |
| SCAttNet | 33.8 | 15.4 | 89.85 | 21.56 |
| MANet | 49.1 | 22.3 | 101.72 | 35.37 |
| EaNet | 86.4 | 80.6 | 144.76 | 27.31 |
| LANet | 30.6 | 15.3 | 86.90 | 19.42 |
| BANet | 29.5 | 14.2 | 12.73 | 3.25 |
| DCST | 100.6 | 45.5 | 118.39 | 34.39 |
| UNetFormer | 83.1 | 36.4 | 87.48 | 22.20 |
| MAResU-Net | 43.8 | 18.8 | 93.14 | 25.81 |
| MACU-Net | 31.5 | 14.8 | 5.15 | 2.11 |
| $A^2$-FPN | 36.2 | 16.4 | 90.66 | 44.32 |
| CMTFNet | 65.1 | 23.2 | 95.35 | 46.28 |
| BESNet | 363.2 | 92.0 | 106.85 | 200.84 |
| HBCNet | 57.8 | 24.0 | 116.71 | 40.00 |
| GCDNet | 59.0 | 24.1 | 109.18 | 32.62 |
| FDEG-Net(Ours) | 52.9 | 23.6 | 105.52 | 32.51 |

("M" stands for millions, and FLOPS stands for the number of floating point operations.) The shape of input tensor is 1×3 × 256×256. Time information is taken from experiments on the Vaihingen dataset.

allowing EGM to ignore low-frequency information and extract high-frequency features related to object edges. This adaptation enables it to effectively handle objects with different shapes and environmental interference.

On the GID dataset in Fig. 11, these scenes showcase the intricate edge features exhibited by various objects, with the EGM successfully preserving the integrity of object edges. Although the presence of intraclass variations prompts the EGM to generate additional edge responses, the interference caused to interclass edges remains minimal. To be more precise, intraclass edges and interclass edges are separated from each other within the backbone features, and the high-frequency information expressed by them does not overlap. Consequently, the edges extracted by the EGM remain distinctly visible.

*Impact of different label types:* Previous research [7] has shown that utilizing labels with eroded boundaries can improve higher accuracy on the Potsdam and Vaihingen datasets. However, this approach only predicts the boundaries on the side closer to the object's center, which does not fully reflect the model's sensitivity and discrimination toward object boundaries. In this study, we assessed the performance of our proposed method using labels with noneroded boundaries to evaluate its true effectiveness. In addition, to provide a more comprehensive demonstration of our method's ability, we reevaluate our model using labels with eroded boundaries. The results presented in Fig. 12 indicate that our proposed network achieves OA of 91.277% on the Potsdam dataset and 90.545% on the Vaihingen dataset, showcasing competitive results. Other metrics also exhibit significant improvements.

*Impact of different backbone networks:* Differences in the backbone network can significantly impact performance. In order to analyze this, we conducted experiments on different backbone networks, as shown in Fig. 13. After considering the depth and width of the network, we selected ResNet-50, ResNeXt-50, ResNet-101, and ResNeXt-101 and evaluated the performance of our proposed method with these backbones. Obviously, the fourth option, which utilized ResNeXt-101, achieves the highest accuracy. The clear and refined features extracted by the backbone network ResNeXt-101 aid EGM in accurately localizing and extracting edges. As a result, using a strong backbone network enhances the performance of our proposed method.

TABLE V
ABLATION STUDY RESULTS ON THE POTSDAM, VAIHINGEN, AND GID DATASETS

| Model | JSCA | EGM | Potsdam | | | | Vaihingen | | | | GID | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | MIoU | AF | OA | K | MIoU | AF | OA | K | MIoU | AF | OA | K |
| Baseline | - | - | 80.839 | 88.867 | 87.890 | 83.596 | 74.026 | 84.582 | 86.174 | 81.527 | 61.934 | 77.856 | 82.580 | 78.169 |
| Baseline | ✓ | - | 81.047 | 89.125 | 88.091 | 83.983 | 74.993 | 84.869 | 86.434 | 81.956 | 62.535 | 78.231 | 82.799 | 78.482 |
| Baseline | - | ✓ | 81.120 | 89.352 | 88.183 | 84.359 | 74.300 | 85.086 | 86.505 | 82.312 | 63.499 | 78.491 | 82.901 | 78.615 |
| Baseline | ✓ | ✓ | 81.438 | 89.585 | 88.362 | 84.713 | 75.482 | 85.757 | 87.073 | 82.868 | 65.436 | 79.861 | 83.463 | 79.039 |

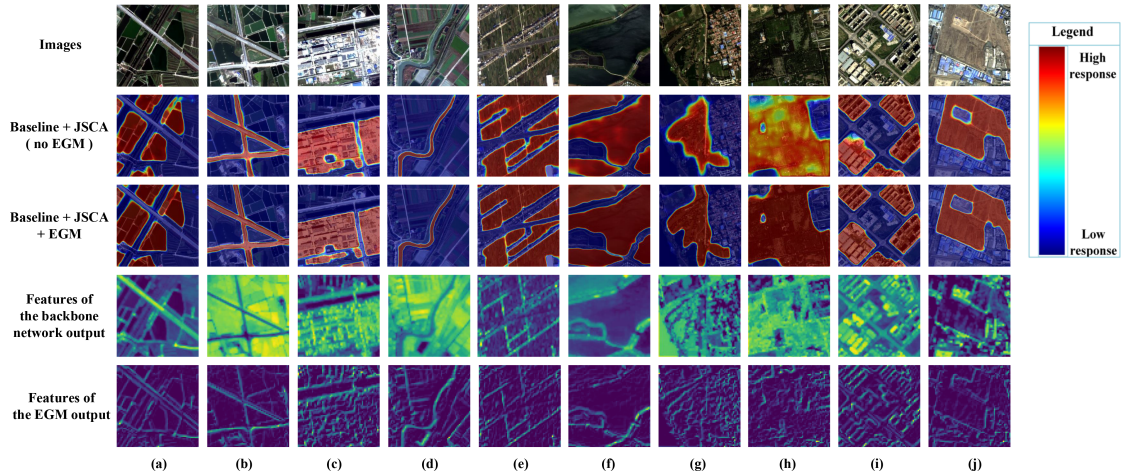


Fig. 11. Visualization of heatmap and features on the GID dataset. Edge extraction results are clearly listed on the last row. (a) Pond. (b) Traffic land. (c) Industrial land. (d) River. (e) Irrigated land. (f) Lake. (g) Artificial grassland. (h) Shrub land. (i) Urban residential. (j) Dry cropland.

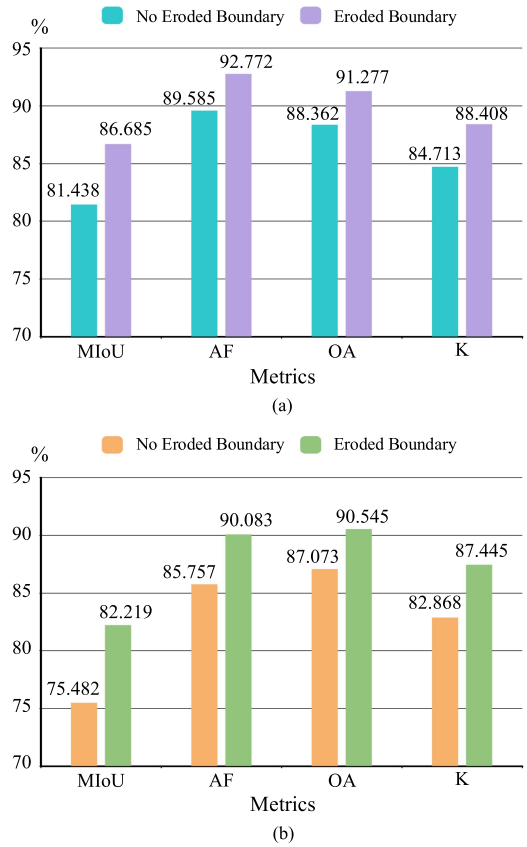


Fig. 12. Experimental results of different label types. (a) Potsdam dataset. (b) Vaihingen dataset.

TABLE VI
ABLATION STUDY RESULTS ON THE MODEL EFFICIENCY

| Model | JSCA | EGM | Train Time (s) / epoch | Inference Time (s) | Parameters (M) | FLOPs (GMac) |
|---|---|---|---|---|---|---|
| Baseline | - | - | 43.2 | 18.6 | 90.39 | 28.96 |
| Baseline | ✓ | - | 44.1 | 20.3 | 105.37 | 30.47 |
| Baseline | - | ✓ | 50.6 | 21.7 | 90.54 | 31.01 |
| Baseline | ✓ | ✓ | 52.9 | 23.6 | 105.52 | 32.51 |

## V. DISCUSSION

The frequency-driven edge guidance network proposed in this article achieves more accurate and robust results in semantic segmentation of RSIs. This mainly benefits from the influence of the two components in the network, JSCA and EGM. The stronger the feature learning and representation capabilities of these components, the better the overall segmentation results. However, through our experiments, we can find that the wavelet transform lacks sensitivity to adjacent objects with high inter-class similarity, leading to potentially false segmentation. This is because the process of wavelet transform is based on pixel intensity. For edges with strong fuzziness and uncertainty, the pixel intensity of features is relatively weak. In this case, the wavelet transform may have potential losses of pixel conversion. In addition, the training efficiency of the proposed method can be further improved, which is crucial for evaluating the model performance.

On the other hand, the main advantage of EGM is that it defines edge positions based on pixel intensity, making it easily

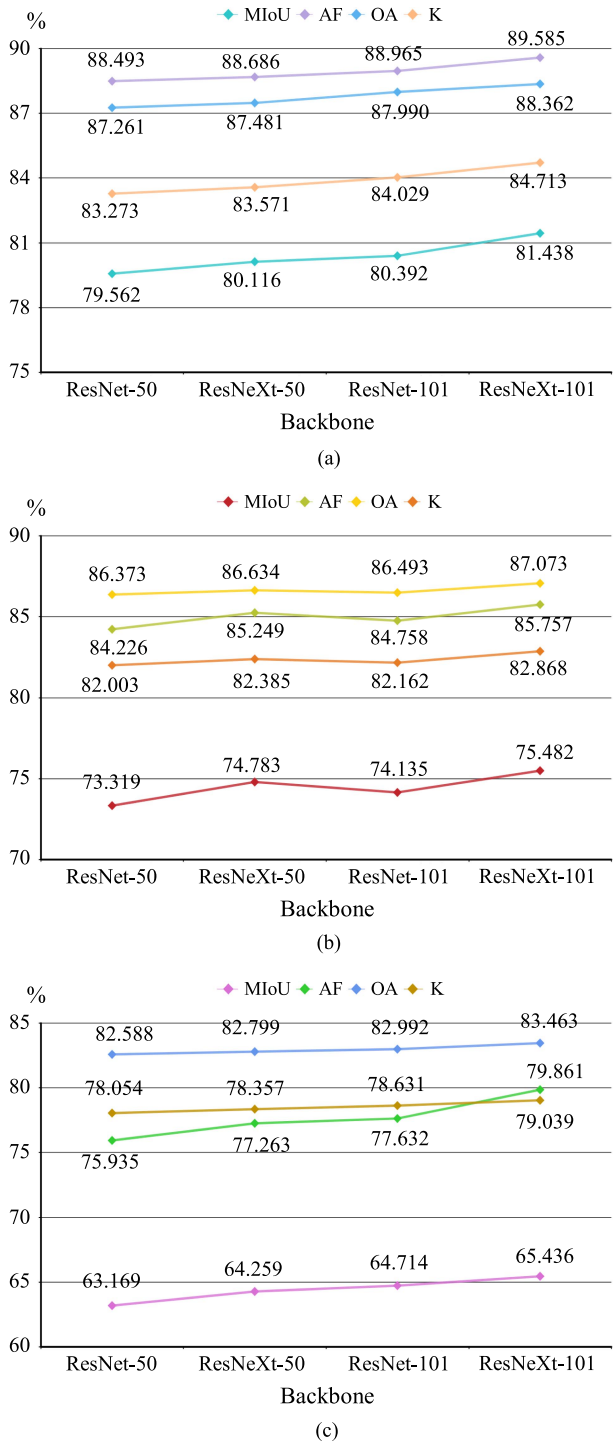


Fig. 13. Experimental results of different backbones on the three datasets. (a) Potsdam dataset. (b) Vaihingen dataset. (c) GID dataset.

interpretable. In addition, it targets feature maps rather than input images. These advantages provide opportunities for other potential types of inputs, such as hyperspectral images, synthetic aperture radar images, and DSM. By further analyzing and studying the frequency domain properties of these multisource images, we can gain more clues to the key features of objects. In the future, we plan to explore multimodal techniques to fully utilize multisource frequency domain features to address the limitations of the proposed method, ultimately improving semantic segmentation.

## VI. Conclusion

In this study, we propose FDEG-Net, a network for RSI semantic segmentation. Specifically, the designed JSCA module enhances the CNNs' ability to model long-range dependencies, effectively addressing object semantic confusion caused by scale differences and complex environments. To improve the segmentation of interclass edges, we design a frequency-driven edge extraction module EGM based on wavelet transform theory to explore the high-frequency characteristics reflected by edges. The EGM flexibly generates edge features of ground objects without introducing additional edge labels. Independent edge features guide the refinement of spatial information to obtain accurate segmentation results. Experimental results on the Potsdam, Vaihingen, and GID datasets demonstrate that FDEG-Net has achieved maximum improvements of 0.807%, 2.220%, and 0.544% in MIoU, AF, and K, respectively, compared with CNN-based methods and Transformer-based methods. A detailed visual analysis of edge extraction intuitively shows that EGM has a clear interpretability advantage. Furthermore, frequency domain properties can provide valuable clues to features of interest, thus improving pixel-level tasks. In future research, we intend to explore supplementary methods, such as incorporating multimodal information, to improve the accuracy of edge segmentation.

## References

[1] G. Moser, S. B. Serpico, and J. A. Benediktsson, "Land-cover mapping by Markov modeling of spatial–contextual information in very-high-resolution remote sensing images," in *Proc. IEEE*, vol. 101, no. 3, pp. 631–651, Mar. 2013, doi: 10.1109/JPROC.2012.2211551.

[2] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," in *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 144, doi: 10.3390/rs10010144.

[3] 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, doi: 10.1109/TGRS.2020.2964675.

[4] 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, doi: 10.1109/CVPR.2015.7298965.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Computer-Assist. Interv.–MICCAI, 2015, 18th Int. Conf.*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[6] 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, doi: 10.1016/J.ISPRSJPRS.2019.07.007.

[7] R. Dong, X. Pan, and F. Li, "DenseU-Net-based semantic segmentation of small objects in urban remote sensing images," *IEEE Access*, vol. 7, pp. 65 347–65 356, 2019, doi: 10.1109/ACCESS.2019.2917952.

[8] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2021, doi: 10.1109/TGRS.2020.2994150.

[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," vol. 2, 2017, *arXiv:1706.05587*, doi: 10.48550/arXiv.1706.05587.

[10] 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, doi: 10.1109/CVPR.2017.660.

[11] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3063381.

[12] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3065, doi: 10.3390/rs13163065.

[13] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2022.3143368.

[14] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "EfficientViT: Memory efficient vision transformer with cascaded group attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14 420–14 430, doi: 10.1109/CVPR52729.2023.01386.

[15] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11 963–11 975, doi: 10.1109/CVPR52688.2022.01166.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[17] 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, doi: 10.1007/978-3-030-01234-2_49.

[18] X. Zheng, L. Huan, G.-S. Xia, and J. Gong, "Parsing very high resolution urban scene images by learning deep convnets with edge-aware loss," *ISPRS J. Photogrammetry Remote Sens.*, vol. 170, pp. 15–28, 2020, doi: 10.1016/j.isprsjprs.2020.09.019.

[19] S. Pan, Y. Tao, C. Nie, and Y. Chong, "PEGNet: Progressive edge guidance network for semantic segmentation of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 4, pp. 637–641, Apr. 2021, doi: 10.1109/LGRS.2020.2983464.

[20] X. Sun, M. Xia, and T. Dai, "Controllable fused semantic segmentation with adaptive edge loss for remote sensing parsing," *Remote Sens.*, vol. 14, no. 1, 2022, Art. no. 207, doi: 10.3390/rs14010207.

[21] A. Li, L. Jiao, H. Zhu, L. Li, and F. Liu, "Multitask semantic boundary awareness network for remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, doi: 10.1109/TGRS.2021.3050885.

[22] J. Jin, W. Zhou, R. Yang, L. Ye, and L. Yu, "Edge detection guide network for semantic segmentation of remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5000505, doi: 10.1109/LGRS.2023.3234257.

[23] B. Sui, Y. Cao, X. Bai, S. Zhang, and R. Wu, "BIBED-Seg: Block-in-block edge detection network for guiding semantic segmentation task of high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1531–1549, 2023, doi: 10.1109/JSTARS.2023.3237584.

[24] J. Zheng, A. Shao, Y. Yan, J. Wu, and M. Zhang, "Remote sensing semantic segmentation via boundary supervision aidedmulti-scale channel-wise cross attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4405814, doi: 10.1109/TGRS.2023.3292112.

[25] Y. Ni, J. Liu, J. Cui, Y. Yang, and X. Wang, "Edge guidance network for semantic segmentation of high resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 9382–9395, 2023, doi: 10.1109/JSTARS.2023.3316307.

[26] 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, doi: 10.1016/J.ISPRSJPRS.2017.11.009.

[27] S. M. Azimi, P. Fischer, M. Körner, and P. Reinartz, "Aerial LaneNet: Lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2920–2938, May 2019, doi: 10.1109/TGRS.2018.2878510.

[28] X. Li, T. Li, Z. Chen, K. Zhang, and R. Xia, "Attentively learning edge distributions for semantic segmentation of remote sensing imagery," *Remote Sens.*, vol. 14, no. 1, 2021, Art. no. 102, doi: 10.3390/rs14010102.

[29] Y. Duan, F. Liu, L. Jiao, P. Zhao, and L. Zhang, "SAR image segmentation based on convolutional-wavelet neural network and Markov random field," *Pattern Recognit.*, vol. 64, pp. 255–267, 2017, doi: 10.1016/j.patcog.2016.11.015.

[30] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3252–3261, Sep. 2018, doi: 10.1109/JSTARS.2018.2860989.

[31] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "RESUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 162, pp. 94–114, 2020, doi: 10.1016/j.isprsjprs.2020.01.013.

[32] Y. Wang, B. Liang, M. Ding, and J. Li, "Dense semantic labeling with atrous spatial pyramid pooling and decoder for high-resolution remote sensing imagery," *Remote Sens.*, vol. 11, no. 1, 2018, Art. no. 20, doi: 10.3390/rs11010020.

[33] G. Chen et al., "Fully convolutional neural network with augmented atrous spatial pyramid pool and fully connected fusion path for high resolution remote sensing image segmentation," *Appl. Sci.*, vol. 9, no. 9, 2019, Art. no. 1816, doi: 10.3390/APP9091816.

[34] C. Zheng, J. Nie, Z.-H. Wang, N. Song, J. Wang, and Z. Wei, "High-order semantic decoupling network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5401415, doi: 10.1109/TGRS.2023.3249230.

[35] M. Yao, Y. Zhang, G. Liu, and D. Pang, "SSNet: A novel transformer and CNN hybrid network for remote sensing semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3023–3037, 2024, doi: 10.1109/JSTARS.2024.3349657.

[36] Y. Fu, X. Zhang, and M. Wang, "DSHNet: A semantic segmentation model of remote sensing images based on dual stream hybrid network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4164–4175, 2024, doi: 10.1109/JSTARS.2024.3355943.

[37] W. Zhao, J. Cao, and X. Dong, "Multilateral semantic with dual relation network for remote sensing images segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 506–518, 2024, doi: 10.1109/JSTARS.2023.3330731.

[38] J. Hou, Z. Guo, Y. Feng, Y. Wu, and W. Diao, "SPANet: Spatial adaptive convolution based content-aware network for aerial image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2192–2204, 2023, doi: 10.1109/JSTARS.2023.3244207.

[39] 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, doi: 10.1109/CVPR.2018.00813.

[40] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154, doi: 10.1109/CVPR.2019.00326.

[41] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607713, doi: 10.1109/TGRS.2021.3093977.

[42] R. Li, L. Wang, C. Zhang, C. Duan, and S. Zheng, "A2-FPN for semantic segmentation of fine-resolution remotely sensed images," *Int. J. Remote Sens.*, vol. 43, no. 3, pp. 1131–1155, 2022, doi: 10.1109/TGRS.2021.3093977.

[43] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603018, doi: 10.1109/TGRS.2021.3065112.

[44] J. Zhang, M. Shao, Y. Qiao, and X. Cao, "Enhancing efficient global understanding network with CSWin transformer for urban scene images segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 10 230–10 245, 2023, doi: 10.1109/JSTARS.2023.3328559.

[45] L. Wang et al., "UNetformer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 190, pp. 196–214, 2022, doi: 10.1016/j.isprsjprs.2022.06.008.

[46] H. Wu, P. Huang, M. Zhang, W. Tang, and X. Yu, "CMTFNet: CNN and multiscale transformer fusion network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 2004612, doi: 10.1109/TGRS.2023.3314641.

[47] Y. Mo, H. Li, X. Xiao, H. Zhao, X. Liu, and J. Zhan, "Swin-Conv-DSPP and global local transformer for remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5284–5296, 2023, doi: 10.1109/JSTARS.2023.3280365.

[48] F. Waldner and F. I. Diakogiannis, "Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network," *Remote Sens. Environ.*, vol. 245, 2020, Art. no. 111741, doi: 10.1016/j.rse.2020.111741.

[49] L. Chen, Z. Qu, Y. Zhang, J. Liu, R. Wang, and D. Zhang, "Edge-enhanced GCIFFNet: A multiclass semantic segmentation network based on edge enhancement and multiscale attention mechanism," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4450–4465, 2024, doi: 10.1109/JSTARS.2024.3357540.
[50] J. Li et al., "Feature guide network with context aggregation pyramid for remote sensing image segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9900–9912, 2022, doi: 10.1109/JSTARS.2022.3221860.
[51] Y. Xu and J. Jiang, "High-resolution boundary-constrained and context-enhanced network for remote sensing image segmentation," *Remote Sens.*, vol. 14, 2022, Art. no. 1859, doi: 10.3390/rs14081859.
[52] F. Chen, H. Liu, Z. Zeng, X. Zhou, and X. Tan, "BES-Net: Boundary enhancing semantic context network for high-resolution image semantic segmentation," *Remote Sens.*, vol. 14, 2022, Art. no. 1638, doi: 10.3390/rs14071638.
[53] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408820, doi: 10.1109/tgrs.2022.3144894.
[54] J. Cui, J. Liu, J. Wang, and Y. Ni, "Global context dependencies aware network for efficient semantic segmentation of fine-resolution remoted sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 2505205, doi: 10.1109/LGRS.2023.3318348.
[55] Y. Ni, J. Liu, J. Cui, Y. Yang, and X. Wang, "Edge guidance network for semantic segmentation of high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 9382–9395, 2023, doi: 10.1109/JSTARS.2023.3316307.
[56] S. Dong, Y. Zhuang, H. Chen, T. Zhang, L. Li, and T. Long, "Full semantic constructed network for urban use classification from very high-resolution optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5606820, doi: 10.1109/TGRS.2022.3225144.
[57] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1492–1500, doi: 10.1109/CVPR.2017.634.
[58] P. Wang et al., "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1451–1460, doi: 10.1109/WACV.2018.00163.
[59] Z. Song, J. Yang, D. Zhang, S. Wang, and Z. Li, "Semi-supervised dim and small infrared ship detection network based on Haar wavelet," *IEEE Access*, vol. 9, pp. 29 686–29 695, 2021, doi: 10.1109/ACCESS.2021.3058526.
[60] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1532–1546, Sep. 2000, doi: 10.1109/83.862633.
[61] X.-Y. Tong et al., "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, 2020, Art. no. 111322, doi: 10.1016/j.rse.2019.111322.
[62] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. Comput. Vis.–ECCV 2020: 16th Eur. Conf.*, 2020, pp. 173–190, doi: 10.1007/978-3-030-58539-6_11.
[63] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 905–909, May 2021, doi: 10.1109/LGRS.2020.2988294.
[64] R. Li, C. Duan, S. Zheng, C. Zhang, and P. M. Atkinson, "MACU-Net for semantic segmentation of fine-resolution remotely sensed images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2020, Art. no. 8007205, doi: 10.1109/LGRS.2021.3052886.

**Jinsong Li** received the B.E. degree in computer science and technology from Shandong University, Jinan, China, in 2021. He is currently working toward the master's degree in computer technology with Qingdao University of Science and Technology, Qingdao, China.

His research interests include computer vision, remote sensing, and deep learning.

**Shujun Zhang** received the Ph.D. degree in artificial intelligence in virtual marine environment from Ocean University of China, Qingdao, China, in 2007.

She was a Postdoctoral Researcher with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China. She is currently an Associate Professor with Qingdao University of Science and Technology, Qingdao. Her research interests include computer vision, image processing, machine learning, and virtual reality.

Dr. Zhang is a Member of the Chinese Computer Federation (CCF) and a Communication Evaluation Expert for the China Academic Degrees and Graduate Education Development Center (CDGDC).

**Yukang Sun** received the B.E. degree in mechanical electronics engineering from the Nanjing Institute of Technology, Nanjing, China, in 2020. He is currently working toward the master's degree in software engineering with the Qingdao University of Science and Technology, Qingdao, China.

His research interests include computer vision, image processing, and deep learning.

**Qi Han** received the B.E. degree in information engineering in 2020 from the Qingdao University of Science and Technology, Qingdao, China, where he is currently working toward the master's degree in computer technology.

His research interests include computer vision, image processing, and deep learning.

**Yuanyuan Sun** received the Ph.D. degree in remote sensing information technology from Zhejiang University, Hangzhou, China, in 2018.

She is currently a Lecturer with the Qingdao University of Science and Technology, Qingdao, China. Her research interests include applied remote sensing, computer vision, pattern recognition, and data mining.

**Yimin Wang** received the Ph.D. degree in modelling physical location based factors of photovoltaic viability from University of Sheffield, Sheffield, U.K., in 2014.

She was a Postdoctoral Researcher with the Solar Physics and Space Plasma Research Centre (SP2RC), School of Mathematics and Statistics, University of Sheffield, Sheffield, U.K. She is currently an Associate Professor with Qingdao University of Science and Technology, Qingdao, China. Her research interests include computer vision, machine learning, and self-supervised learning.

Dr. Wang is a committee member of the International Federation of Automatic Control (IFAC).