IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 December 2023, accepted 31 December 2023, date of publication 8 January 2024,
date of current version 16 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3350645

APPLIED RESEARCH

Local-Global Feature Capture and Boundary Information Refinement Swin Transformer Segmentor for Remote Sensing Images

RUI LIN, (Graduate Student Member, IEEE), YING ZHANG, XUE ZHU, AND XUEYUN CHEN
School of Electrical Engineering, Guangxi University, Nanning 530004, China

Corresponding author: Xueyun Chen (20140043@gxu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 62061002.

ABSTRACT Semantic segmentation of urban remote sensing images is a highly challenging task. Due to the complex background, occlusion overlap, and small-scale targets in urban remote sensing images, the semantic segmentation results suffer from deficiencies such as similar target confusion, blurred target boundaries, and small-scale target omission. To solve the above problems, a local-global feature capture and boundary information refinement Swin Transformer segmentor (LGBSwin) is proposed. First, the dual linear attention module (DLAM) utilizes spatial linear attention and channel linear attention mechanisms for strengthening global modeling capabilities to improve the segmentation ability of similar targets. Second, boundary-aware enhancement (BAE) adaptively mines the boundary semantic information through the effective integration of high-level and low-level features to alleviate blurred boundaries. Finally, feature refinement aggregation (FRA) establishes information relationships between different layers, reduces the loss of local information, and enhances local-global dependence, thus significantly improving the recognition ability of small targets. Experimental results demonstrate the effectiveness of LGBSwin, with an F1 of 91.02% on the ISPRS Vaihingen dataset and 93.35% on the ISPRS Potsdam dataset.

INDEX TERMS Boundary information, dual linear attention, feature capture, remote sensing images, semantic segmentation.

I. INTRODUCTION

Semantic segmentation [1] is a crucial problem in remote sensing research, and its core objective lies in semantic category recognition on a pixel-by-pixel basis. Currently, remote sensing image semantic segmentation has been widely used in practical application scenarios such as urban planning [2], natural resource management [3], disaster assessment [4], and agricultural production [5], [6]. It can provide highly accurate and efficient application solutions in various fields [7], [8], [9], [10]. However, the high-resolution urban remote sensing images contain a large amount of complex information, which hinders the extraction of global structure and semantic information of targets. Also, occlusion and overlap problems often lead to semantic ambiguity, making it impossible to classify pixels correctly. The issue of tiny targets and blurred boundary information makes it difficult to distinguish between categories. Therefore, extracting semantic information from urban remote sensing images remains a daunting challenge.

The particular characteristics of features, such as tiny size, mutual occlusion, and high similarity, made semantic segmentation utilizing traditional methods laborious and costly. In recent years, the convolutional neural network (CNN) has exhibited exceptional capabilities in computer vision [11], [12]. Fully Convolutional Network (FCN) [13] pioneered architecture that enabled end-to-end CNN-based pixel-level classification. Subsequently, encoder-decoder architectures gained prominence, with UNet [14] utilizing skip connections to capture spatial correlation between coding layers. SegNet

The associate editor coordinating the review of this manuscript and approving it for publication was Tai Fei.

[15] up-sampled low-resolution feature maps to enhance the understanding of spatial information. DeepLabV3+ [16] added a decoder module that utilizes spatial features, substantially enhancing network performance compared to DeepLabV3 [17]. CNN had the advantage of spatial location representation to capture finer-grained local information.

While CNN demonstrated superiority for local information, it was difficult to accurately recognize categories by relying only on local information. Consequently, introducing boundary information facilitated the accurate localization of the target at the pixel level. Currently, numerous approaches [18], [19] have focused on enhancing models' sensitivity to boundary information. HRCNet [20] leveraged HRNet [21] structures to preserve spatial information and effectively addressed the issue of rough edges in feature maps. EDGNet [22] utilized boundary spatial information to facilitate multi-modal information fusion. EGRCNN [23] improved model accuracy using the prior edge information. REFLA [24] performed region-level segmentation based on edge feature and label assistance, which showed better segmentation accuracy. RUnT [25] reduced semantic confusion due to the high similarity between categories by combining edge features with semantic features. EU-Net [26] designed the edge aggregation path for extracting multilevel edge correlation information, which effectively improves the performance of the network. This showed that the importance of boundary information cannot be ignored.

Although many CNN-based methods took boundary information into account and showed excellent performance, some shortcomings existed in long-range dependencies and spatial modeling relationships. It was owing to the limitations of convolutional operations. Modifying the convolution operation to expand the receptive field has become one of the approaches to improve the limitations of convolution. Dilated convolution [27] expanded the receptive field by increasing the dilated rate of the convolutional kernel. Spatial pyramidal pooling [28], which involved pooling at different sizes. VAN [29] proposed decomposing large kernel convolutional operations to capture long-range relations. LSKNet [30] dynamically adjusted its large spatial receptive field to better model the varying contextual nuances of different object types. In addition, several studies have investigated the attention mechanisms to capture long-range dependencies in feature maps. For instance, DANet [31] was proposed by acquiring positional and spatial attention to model extensive contextual information within feature maps effectively. By merging the spatial pyramid structure into the attention mechanism, SPANet [32] considerably enhanced recognition accuracy. Eca-net [33] proposed a method based on the adaptive selection of convolutional kernel size to enable information interaction between channels. The above methods indirectly encoded the global context and focused on capturing global features by aggregating the local features obtained from the CNN. Therefore, it is essential to acknowledge that directly obtaining clear global contextual information in remote sensing images remains crucial.

Transformer-based [34] methods have demonstrated remarkable performance in computer vision with their powerful global information modeling capabilities. By leveraging inter-sequence prediction and employing a multi-head attention mechanism, the ViT [35] captured long-distance dependence and adaptive spatial aggregation capabilities, thus allowing for more powerful and robust representations than CNN to be learned from massive data. Building upon this, Swin Transformer [36] pioneered a hierarchical feature representation scheme that achieved impressive results while maintaining linear computational complexity. UNetFormer [37] developed an efficient attention mechanism to enable the interaction of global and local information. DCSwin [38] further utilized the Swin Transformer as a backbone for extracting contextual information and designed a DCFAM decoder to capture multi-scale relationally enhanced semantic features. ST-UNet [39] designed a novel dual-encoder structure that directs the primary encoder of CNN to capture more diverse features through global features. Despite these significant contributions, none of the abovementioned methods had fully integrated three crucial factors–local information, global contextual information, and boundary information–in semantic segmentation for urban remote sensing images.

This study proposes a novel network framework to consider local information, global contextual information, and boundary information in an integrated way called LGBSwin. We adopt Swin Transformer as an encoder due to its exceptional capabilities in global modeling. The main contributions are as follows:

1) Proposed LGBSwin solves problems in semantic segmentation of urban remote sensing images, such as occlusion and overlapping, size disparity, and intricate background.
2) To capture global contextual information and strengthen long-range dependencies for irregular targets, we construct a dual linear attention module (DLAM).
3) A boundary-aware enhancement (BAE) is developed to extract boundary features and effectively resolve the issue of boundary blurring in remote sensing targets.
4) A feature refinement aggregation (FRA) module is proposed that not only efficiently utilizes the rich semantic information but also retains the local details.

II. METHODOLOGY

A. OVERVIEW

Fig. 1 illustrates the overall architecture of LGBSwin. LGBSwin adopts an encoder-decoder framework. Swin Transformer as the encoder for depth feature extraction and image internal correlation modeling. The output of four stages is processed through a standard 1 × 1 convolution to generate four features (S_1, S_2, S_3, and S_4). We employ the DLAM to capture global multi-scale information. Additionally, BAE is designed to enhance the boundary semantics associated with objects. BAE utilizes low-level features

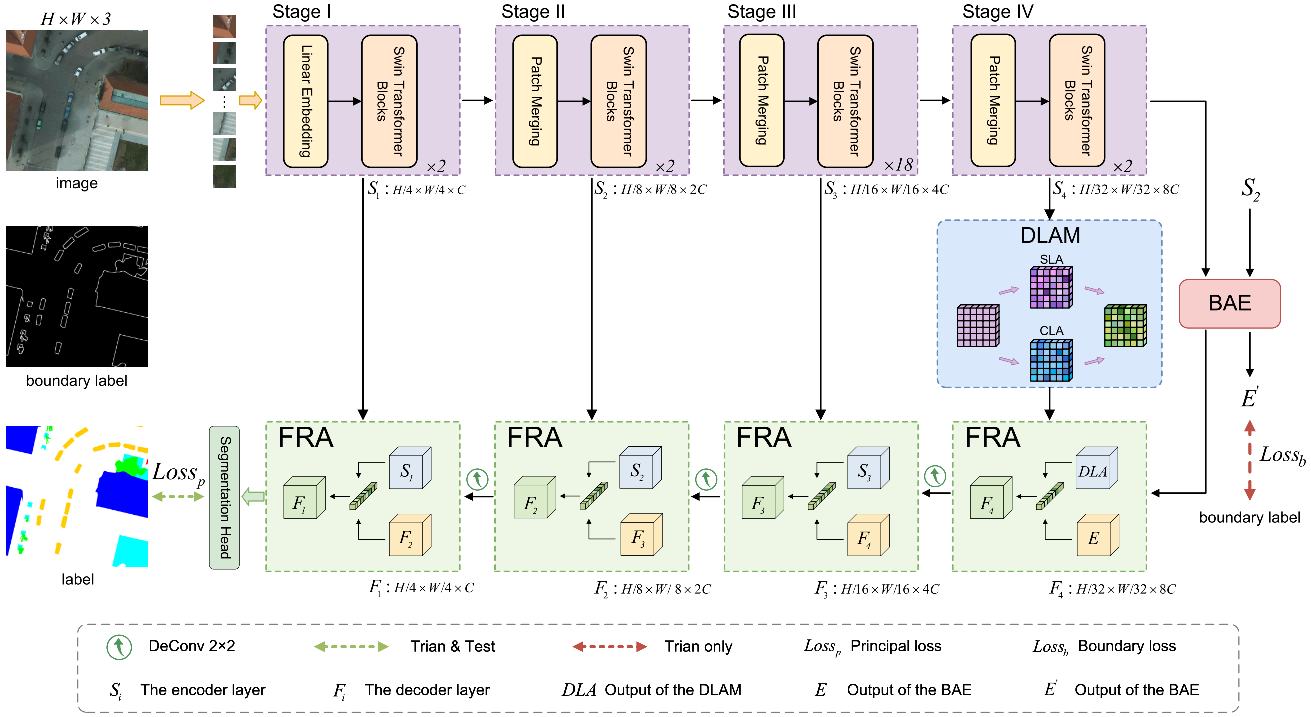


FIGURE 1. The overall architecture of LGBSwin. It contains the channel linear attention mechanism (CLA), the spatial linear attention mechanism (SLA), the dual linear attention module (DLAM), the boundary-aware enhancement (BAE), and the feature refinement aggregation (FRA).

containing local boundary details and high-level features with global location information under boundary supervision. By extracting boundary weights, we incorporate them into the FRA through fusion. In the FRA, we address the problem of neglecting small targets in high-level features by ensuring a compelling fusion of high-level and low-level features.

B. DUAL LINEAR ATTENTION MODULE

The Swin Transformer significantly reduces memory overhead using a multi-head attention mechanism within a confined window. This approach, involving alternating rule and shift window execution, inevitably limited the ability for global modeling to some degree. To better tackle this limitation, we design the dual linear attention module (DLAM), which introduces linear attention in spatial and channel dimensions to consider inter-pixel relationships to augment the global dependency of semantic features. The effective utilization of global dependency enables DLAM to better understand the overall context of urban remote sensing images. The model with DLAM not only captures the details of various feature structures but also comprehends the associations between them, leading to more accurate semantic segmentation. The components of DLAM are shown in Fig. 2.

Specifically, to optimize computational efficiency, a preliminary step is undertaken in which the feature map is treated to a 1 × 1 convolution operation to reduce the number of channels to $c/2$. Subsequently, a combination of two-branch asymmetric convolution [40] and dilated convolution layers [27] is employed to comprehensively gather feature information from objects of varying scales, leveraging distinct receptive fields. Then, the fused feature information is concatenated. The following equation can represent this process:

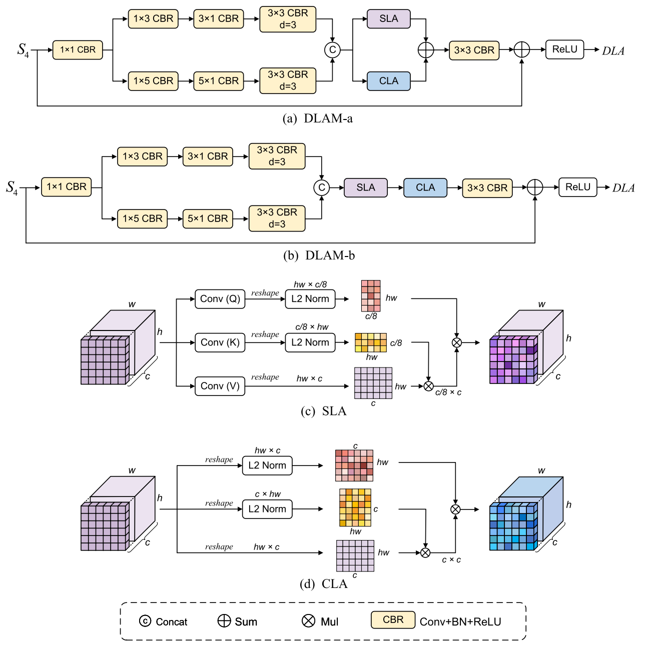
$$T_1 = f_\tau \left(f_\mu \left(f_\theta \left(S_4\right)\right)\right) \tag{1}$$

$$T_2 = f_\tau \left(f_\varphi \left(f_\theta \left(S_4\right)\right)\right) \tag{2}$$

$$T = \mathrm{Cat}\left(T_1, T_2\right) \tag{3}$$

where $S_4 \in \mathbb{R}^{h\times w\times c}$ is the input. $T_1 \in \mathbb{R}^{h\times w\times (c/2)}$ denotes the output of the first branch. $T_2 \in \mathbb{R}^{h\times w\times (c/2)}$ represents the output of the other branch. f_θ denotes the 1 × 1 convolution. f_τ is a 3 × 3 dilated convolution with a dilated rate of 3. f_μ is a composite function that specifically undergoes a 1 × 3 convolution followed by a 3 × 1 convolution. f_φ is a composite function that undergoes 1 × 5 convolution and 5 × 1 convolution in turn. $T \in \mathbb{R}^{h\times w\times c}$ is the result of the concatenation process performed on the outputs from two branches. Cat(·) represents the concatenation along the channel dimension.

Then, the feature map $T \in \mathbb{R}^{h\times w\times c}$ with global multi-scale information is transferred to the spatial linear attention mechanism (SLA) and channel linear attention mechanism (CLA) for attention enhancement. Based on the linear attention mechanism [41], the SLA and the CLA have long-term dependencies in the spatial and channel dimensions. In DLAM-a, the output after 3 × 3 convolution is summed with the original input to obtain the output feature



(a) DLAM-a

(b) DLAM-b

(c) SLA

(d) CLA

FIGURE 2. The components of dual linear attention module (DLAM). DLAM-a and DLAM-b are two various dual linear attention modules. SLA and CLA are the spatial linear attention mechanism and channel linear attention mechanism separately, and they are the components of DLAM-a and DLAM-b. DLAM-a is used in BGSwin.

$DLA(S_4) \in \mathbb{R}^{h\times w\times c}$, which can be expressed as:

$$DLA(S_4) = f_\sigma\ (S_4 \oplus f_\delta(SLA(T) \oplus CLA(T))) \tag{4}$$

where $SLA(T) \in \mathbb{R}^{h\times w\times c}$ and $CLA(T) \in \mathbb{R}^{h\times w\times c}$ represent the SLA and the CLA, respectively. f_δ is a 3 × 3 convolution. f_σ symbolizes the ReLU activation function. $\oplus$ denotes the element-wise addition operation. Note that with DLAM-a in LBGSwin, DLAM defaults to the DLAM-a version if not otherwise specified.

In DLAM-b, the output $DLA'(S_4) \in \mathbb{R}^{h\times w\times c}$ can be expressed as:

$$DLA'(S_4) = f_\sigma\ (S_4 \oplus f_\delta(CLA(SLA(P)))) \tag{5}$$

C. BOUNDARY-AWARE ENHANCEMENT

Previous studies [42] on Swin Transformer established layered networks by projecting flattened patches or combining features from adjacent blocks and performing linear operations. However, the above methods tended to result in a loss of fine-grained details and boundary information, which hindered effective segmentation. Furthermore, remote sensing images are prone to feature occlusion, leading to indistinct boundaries. It is well known that low-level features retain rich boundary details of urban scenes but lack semantic content, whereas high-level features offer precise semantic information but possess coarse spatial resolution. In order to effectively mine the object boundary information and thus alleviate the problem of occlusion and overlapping of remote sensing images, we design the BAE, which integrates low-level features from stage II with high-level features from stage IV. It is imperative to emphasize that the lower-level features of stage II are chosen instead of stage I in BAE input. It is taken into account that the stage I features are closer to the original input, have a lot of redundant information, and have a smaller receptive field.

Specifically, we input the low-level feature S_2 to a 1 × 1 convolution to obtain feature $f_\theta\ (S_2)$. In contrast, the high-level feature S_4 is channel-adjusted and upsampled, which can be expressed as $Up_{4\times}\ (f_\theta\ (S_4))$. Then, the concatenate operation is performed on $f_\theta\ (S_2)$ and $Up_{4\times}\ (f_\theta\ (S_4))$. The merged feature is subjected to convolutional operations and a sigmoid activate function to facilitate the fusion between high-level and low-level features, evaluate channel dependency, and adaptively learn important information across channels to obtain the boundary feature E'. On the one hand, E' and the boundary label compute the binary cross-entropy loss to achieve boundary supervision. On the other hand, we multiply E' and S_4 element-by-element, then use the skip connection to obtain $E\ (S_4)$. In summary, the input-output process of BAM can be described using the following equation:

$$E' = f_\varepsilon\ (f_\omega\ (\mathrm{Cat}\ (f_\theta\ (S_2)\ ,\ U\ p_{4\times}\ (f_\theta\ (S_4))))) \tag{6}$$

$$E\ (S_4) = D\left(E'\right) \odot S_4 \oplus S_4 \tag{7}$$

where f_ω denotes successive passes through two 3 × 3 convolution functions and one 1 × 1 convolution layer. f_ε is a sigmoid activate function. $\odot$ represents an element-level multiplication. $D(\cdot)$ denotes downsampling.

D. FEATURE REFINEMENT AGGREGATION

Global semantic information plays a vital role in intricate urban scenes. Introducing boundary information further enhances the segmentation quality. Nevertheless, it would be unwise to overlook local information since it retains substantial spatial details. Notably, urban remote sensing images exhibit significant variations in target scales, necessitating a focus on leveraging local information to delineate object regions across diverse scales accurately. Previous research [43] has demonstrated that different layers encompass distinct information. Consequently, we design the feature refinement aggregation (FRA) to effectively utilize context derived from different layers to better address the problem of target scale variations in urban remote sensing images, especially for small targets. The structure of FRA is illustrated in Fig. 3.

We employ upsampling on the output of the FRA from the previous layer to recover the resolution. Convolution is utilized to modify the channel dimension adaptively. Then perform a concatenation operation between the features from the encoder and the output of the FRA from the previous layer along the channel dimension to achieve a fusion of high-level and low-level semantics. The resulting fused feature representation, denoted as F'_{i-1}, encapsulates both

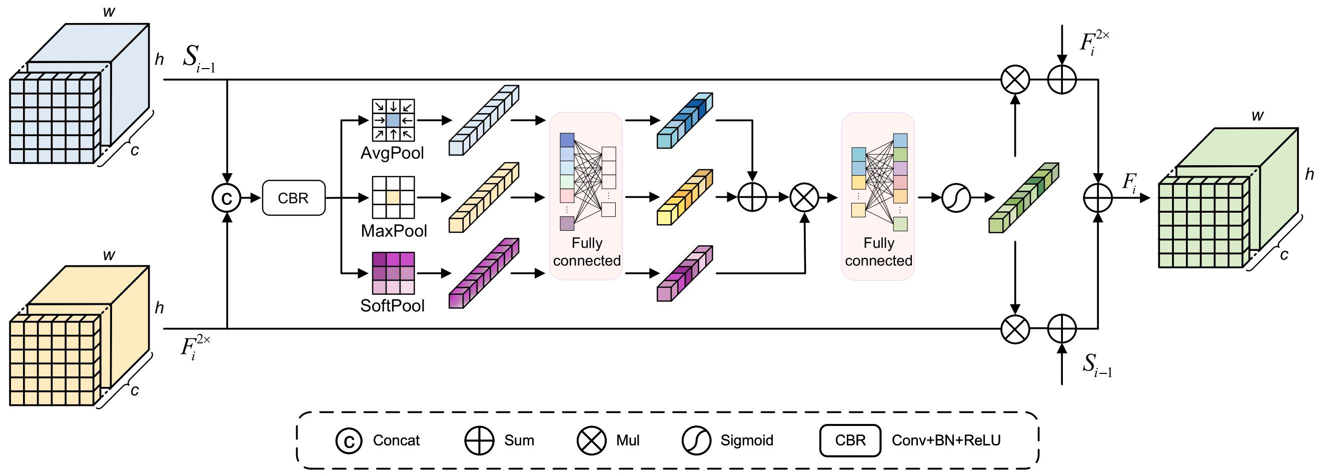


FIGURE 3. The structure of feature refinement aggregation (FRA). S_i stands for the feature from Swin Transformer. F_i is defined as the feature from FRA.

high-level and low-level information:

$$F_4' = f_\theta\left(\mathrm{Cat}\left(DLA\left(S_4\right), E\left(S_4\right)\right)\right) \tag{8}$$

$$F_{i-1}' = f_\theta\left(\mathrm{Cat}\left(S_{i-1}, F_i^{2\times}\right)\right), i \in \{2, 3, 4\} \tag{9}$$

where $\{S_i\}_{i=1}^4$ represents the four-stage output of the Swin Transformer. The four outputs of the FRA are denoted as $\{F_i\}_{i=1}^4$. $DLA\left(S_4\right)$ and $E\left(S_4\right)$ represent the output feature maps of DLAM and BAE, respectively. $F_i^{2\times}$ represents the bilinear interpolation of F_i. Each channel in the feature map performs the role of a feature detector, emphasizing significant image content through channel dependence. We utilize three distinctive pooling strategies to incorporate a more profound knowledge of channel dependence. Primarily, we employ adaptive average pooling and maximum pooling layers to calculate vital features across the channels. These extracted features are subsequently transmitted to a shared fully connected layer. By summing the outputs of these two pooling, denoted as $P_{A\&M}$. Simultaneously, we introduce soft pooling [44] with exponential weights to derive global weights that capture the relative significance of the channels. The global weights are then multiplied element by element with $P_{A\&M}$ to obtain P. The overall process can be concisely captured using the following equation:

$$\begin{aligned} P_{A\&M} &= f_\sigma\left(\psi_1\left(\mathrm{AvgPool}\left(F_i'\right)\right)\right) \\ &\quad \oplus f_\sigma\left(\psi_1\left(\mathrm{MaxPool}\left(F_i'\right)\right)\right) \end{aligned} \tag{10}$$

$$P = f_\varepsilon\left(\psi_2\left(P_{A\&M} \odot \left(f_\sigma\left(\psi_1\left(\mathrm{SoftPool}\left(F_i'\right)\right)\right)\right)\right)\right) \tag{11}$$

where AvgPool(·), MaxPool(·), and SoftPool(·) denote adaptive average pooling, maximum pooling, and soft pooling, respectively. ψ_1 and ψ_2 are defined as fully connected layers of decreasing size and fully connected layers of increasing size, in that order. We then optimize the feature representation by multiplying the P with the high-level and low-level features. Finally, the superimposition of these feature maps generates an output F_i, which can be expressed mathematically as follows:

$$F_4 = DLA\left(S_4\right) \odot P + E\left(S_4\right) \odot P + DLA\left(S_4\right) + E\left(S_4\right) \tag{12}$$

$$\begin{aligned} F_{i-1} &= S_{i-1} \odot P + F_i^{2\times} \odot P + S_{i-1} + F_i^{2\times}, \\ &\quad i \in \{2, 3, 4\} \end{aligned} \tag{13}$$

E. LOSS FUNCTION

During the training phase, we implemente a multi-task segmentation architecture by utilizing the final layer of the FRA to generate the segmentation map. Boundary supervision includes the utilization of both the BAE output E' and boundary label, as illustrated in Fig. 1. The multi-task segmentation architecture involves two key components. On the one hand, we use a mask to supervise the final segmentation map, and the principal loss $Loss_p$ is a combination of cross-entropy loss and dice loss, which can facilitate more effective network construction and optimization. This can be defined as:

$$Loss_p = Loss_{ce} + Loss_{dice} \tag{14}$$

On the other hand, the boundary loss $Loss_b$ utilizes the binary cross-entropy loss. In order to achieve a more harmonized weighting between the principal loss and the boundary loss, a balancing factor α is introduced, which serves to multiply the boundary loss. Consequently, the comprehensive representation of the total loss can be succinctly expressed through the following equation:

$$Loss_{total} = Loss_p + \alpha Loss_b \tag{15}$$

where the balancing factor α is set to 0.4.

III. EXPERIMENTAL RESULTS

A. DATASETS

Our proposed network undergoes validation using two publicly available datasets: the ISPRS Vaihingen dataset and the ISPRS Potsdam dataset. The ISPRS Vaihingen dataset comprises 33 tiles with a 9cm/pixel high resolution, accompanied by classification labels. The classification labels include six categories: impervious surfaces, buildings, low vegetation, trees, cars, and background. We apply the canny algorithm to the labels to generate boundary labels. Consistent with the division rules defined by the benchmark organizers, we classify 16 images in the Vaihingen dataset for training and the remaining 17 images for testing. Similarly, the ISPRS Potsdam dataset contains 38 patches of high-resolution (5cm/pixel) and the corresponding classification labels. 24 patches are set as the training set, and the other 14 patches are used for testing.

B. IMPLEMENTATION DETAILS

1) TRAINING SETTING

All experiments are implemented using PyTorch on an NVIDIA Geforce RTX 3090Ti GPU with a batch size of 8, an optimizer set to AdamW, a learning rate of 0.0003, and a weight decay value of 0.0025. To facilitate effective training, the original images and their corresponding labels are cropped to dimensions of 512 × 512. Prior to training, image enhancement techniques are applied to augment the dataset. Image enhancement is performed in the following ways: random rotation, random resizing, flipping along the horizontal or vertical axis, and adding Gaussian noise.

2) EVALUATION METRICS

We employ the following evaluation metrics to assess the performance of LBGSwin: overall accuracy (OA), mean intersection over union (mIoU), and F1 score (F1). These three evaluation metrics are based on a confusion matrix that contains four terms: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). For each category, mIoU is defined as the ratio of the intersection and concatenation of predicted and true values. The above evaluation metrics are calculated as follows:

$$OA = \frac{TP + TN}{TP + FP + FN + TN} \tag{16}$$

$$mIoU = \frac{TP}{TP + FP + FN} \tag{17}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{18}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{19}$$

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{20}$$

C. ABLATION STUDY

To assess the efficacy of the LBGSwin component, we comprehensively evaluate three crucial modules within the network. We use the Swin Transformer as a baseline network for ablation experiments on the Vaihingen datasets and the Potsdam datasets. The experimental configuration specifics and quantitative outcomes are meticulously documented in Table 1. The hyperparameters are kept consistent in all experiments.

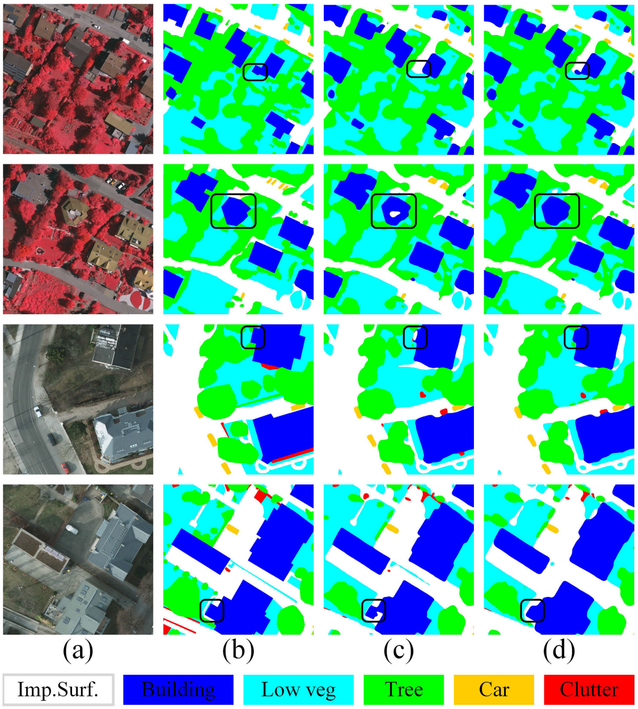


FIGURE 4. Comparison results before and after using DLAM in the baseline. (a) Image. (b) Ground truth. (c) Results of Swin-S. (d) Results of Swin-S+DLAM.

1) THE EFFECTS OF BASELINE

We choose Swin-S: C = 96, window size = 8 × 8, block numbers = $\{2, 2, 18, 2\}$ pre-trained on ImageNet as the backbone of the encoder, which is recovered to the exact resolution as the original input image using an upsampling operation. Table 1 shows the performance of Swin-S as a baseline in the Vaihingen and the Potsdam.

Table 2 presents a comparison of three distinct backbones, namely Swin-T, Swin-S, and Swin-B, while simultaneously maintaining DLAM, BAE, and FRA. The mIoU for Swin-S in the Vaihingen dataset reveals a 1.26% increase compared to Swin-T. Swin-S displays a mIoU metric of 0.73% greater than Swin-B in the Vaihingen. In a similar vein, Swin-S exhibits a higher level of excellence in the Potsdam.

2) THE EFFECTS OF DUAL LINEAR ATTENTION MODULE

Table 1 illustrates that when DLAM is considered in the baseline, there is an increase of 1.01% on the mIoU and 0.6% on F1 in the Potsdam. More intuitively, the comparison of the visual segmentation results is shown in Fig. 4. In the second row, the use of DLAM-a enables the model to discern the segment of the rooftop that exhibits a dissimilar coloration from the rest of the roof and accurately classifies it within the "Building" category. The third and fourth rows present a resemblance in coloration between the balcony and the road, resulting in potential confusion. However, the

TABLE 1. Results of module ablation experiments.

Model Name	Modules			Vahihingen			Potsdam		
	DLAM	BAE	FRA	mIoU	F1	OA	mIoU	F1	OA
Swin-S				76.66	86.35	89.48	83.52	90.91	90.61
Swin-S+DLAM	√			77.32	86.80	89.74	84.53	91.51	91.03
Swin-S+BAE		√		78.32	87.64	90.18	85.48	91.89	91.24
Swin-S+FRA			√	81.57	89.42	90.31	86.15	92.67	91.38
Swin-S+DLAM+BAE	√	√		81.69	89.77	90.77	87.01	92.94	91.62
Swin-S+DLAM+FRA	√		√	83.44	90.83	91.43	87.49	93.21	91.79
Swin-S+BAE+FRA		√	√	83.11	90.64	91.17	87.36	93.14	91.71
Swin-S+DLAM+BAE+FRA	√	√	√	**83.71**	**91.02**	**91.53**	**87.73**	**93.35**	**91.87**

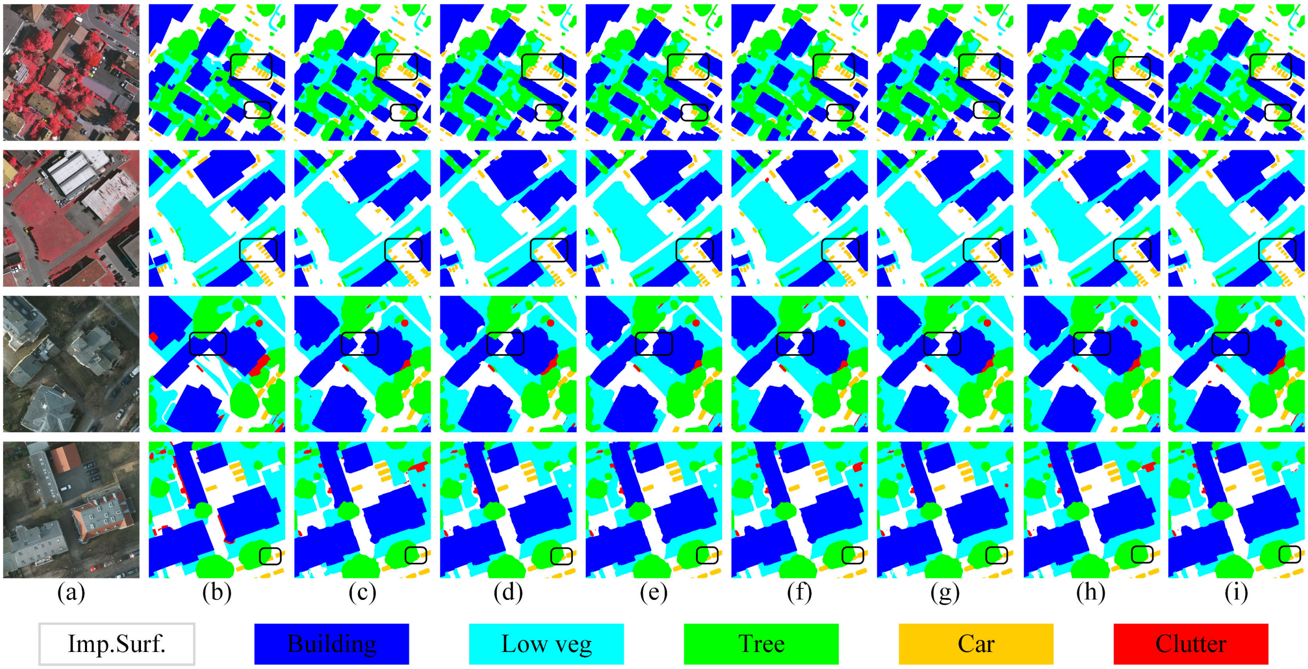


FIGURE 5. Comparison results of attention ablation experiments. (a) Image. (b) Ground truth. (c) Results of CBAM. (d) Results of self-attention. (e) Results of SE block. (f) Results of ECA. (g) Results of SA. (h) Results of DLAM-b. (i) Results of DLAM-a.

TABLE 2. Results of different baseline.

Model Name	Vaihingen			Potsdam		
	mIoU	F1	OA	mIoU	F1	OA
Swin-T+DLAM+BAE+FRA	82.45	90.24	91.27	87.36	93.14	91.72
Swin-S+DLAM+BAE+FRA	**83.71**	**91.02**	**91.53**	**87.73**	**93.35**	**91.87**
Swin-B+DLAM+BAE+FRA	82.98	90.56	91.47	87.62	93.29	91.74

TABLE 3. Results of the attention ablation experiments.

Attention	Vaihingen			Potsdam		
	mIoU	F1	OA	mIoU	F1	OA
CBAM	83.23	90.72	91.50	87.34	93.13	91.69
Self-attention	83.61	90.95	91.55	87.61	93.30	91.78
SE block	83.27	90.73	91.37	87.41	93.17	91.68
ECA	83.44	90.84	91.51	87.25	93.08	91.61
SA	83.32	90.77	91.59	87.20	93.05	91.58
DLAM-b	83.60	90.95	**91.67**	87.69	93.33	91.80
DLAM-a	**83.71**	**91.02**	91.53	**87.73**	**93.35**	**91.87**

model utilizing DLAM-a proves the capability to discern the distinction between the two objects. The results indicate that the employment of DLAM provides the ability to augment the model's capacity to capture global information effectively.

To further explore the role of attention modules, we compared DLAM-a, DLAM-b, and other attention mechanisms. We remove the SLA and CLA from LGBSwin and use other attention modules. The results are shown in Table 3, where DLAM-a is highest in the mIoU and F1 in the Vaihingen and Potsdam, outperforming other attention mechanisms such as CBAM [45], self-attention [34], SE block [46], ECA [33],

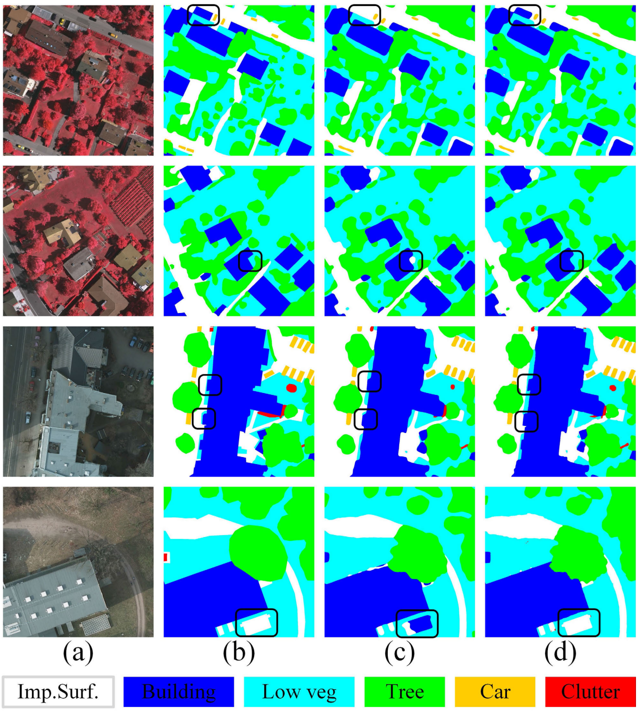


FIGURE 6. Comparison results before and after using BAE in the baseline. (a) Image. (b) Ground truth. (c) Results of Swin-S. (d) Results of Swin-S+BAE.

TABLE 4. Results of different balancing factor α.

α	Vaihingen			Potsdam		
	mIoU	F1	OA	mIoU	F1	OA
0.3	83.46	90.86	**91.57**	87.63	93.29	91.84
0.4	**83.71**	**91.02**	91.53	**87.73**	**93.35**	**91.87**
0.5	83.34	90.79	91.29	87.45	93.19	91.71

and SA [47]. The visualization results are shown in Fig. 5. In the third row, the buildings are similar to the road colors and can be correctly classified using DLAM-a. It is indicated that DLAM enhances the global modeling capability and improves the recognition accuracy of similar categories.

3) THE EFFECTS OF BOUNDARY-AWARE ENHANCEMENT

Table 1 demonstrates that adding BAE to the baseline enhances the mIoU by 1.66% in the Vaihingen and 1.96% in the Potsdam. Fig. 6 displays the efficiency of BAE in capturing boundary details. In the first row, the model introducing BAE uses the boundary information to identify buildings that are similar to roads. In the third row, the model with the addition of BAE extracts more precise building boundary information. The above outcomes show that the sensitivity of BAE to boundary information can boost the ability to recognize target boundaries.

4) THE EFFECTS OF FEATURE REFINEMENT AGGREGATION

As can be seen from Table 1, the model performance is significantly improved by adding FRA to the baseline. In the Vaihingen, mIoU increases by 4.91%, and F1 rises by 3.07%. In the Potsdam, mIoU boosts by 2.63%, and F1 improves by 1.29%. More specifically, Fig. 7 visualizes the comparison results. The small target size of the cars renders it tough to segment cars in a dense car situation (e.g., the second row). With the addition of FRA, the model can segment the "Car" category more accurately. This shows that introducing local

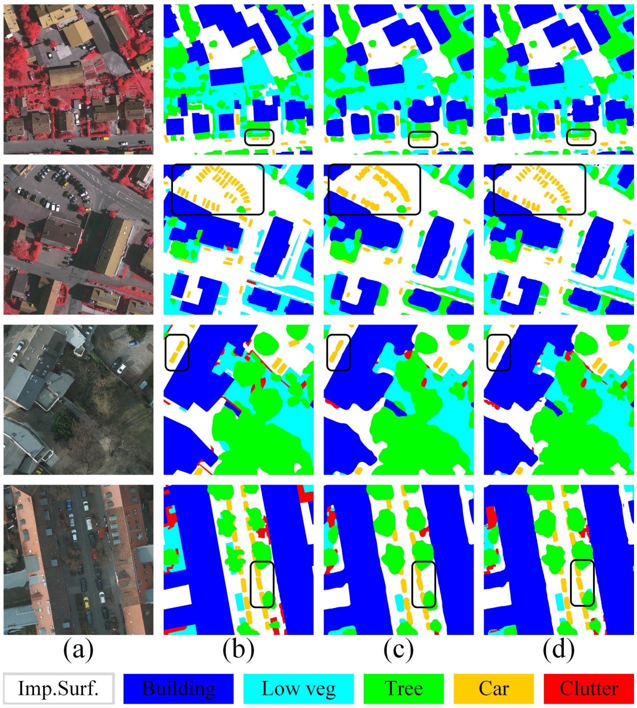


FIGURE 7. Comparison results before and after using FRA in the baseline. (a) Image. (b) Ground truth. (c) Results of Swin-S. (d) Results of Swin-S+FRA.

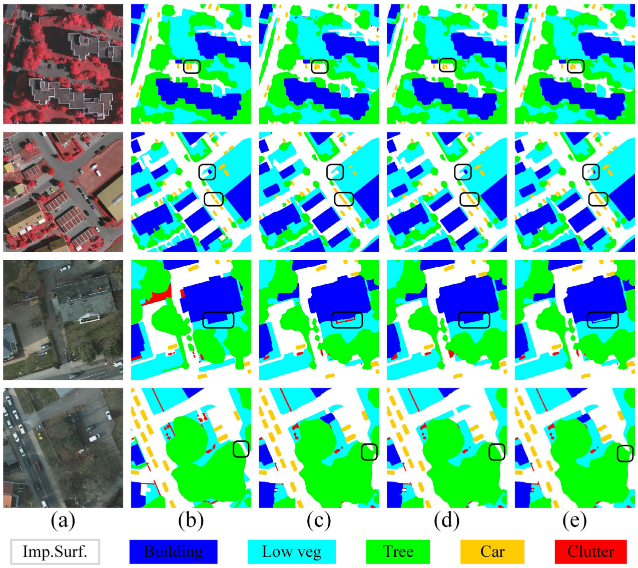


FIGURE 8. Comparison results of different balancing factor α. (a) Image. (b) Ground truth. (c) Results of the α with 0.3. (d) Results of the α with 0.4. (e) Results of the α with 0.5.

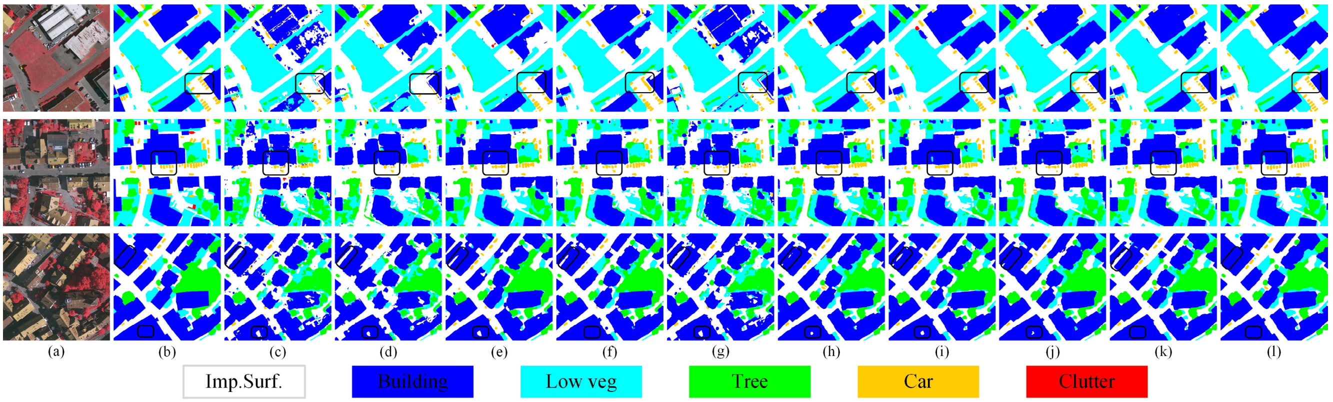


FIGURE 9. Comparison of visualization results on the Vaihingen. (a) Image. (b) Ground truth. (c) FCN. (d) UNet. (e) DeepLabV3. (f) SPANet. (g) DANet. (h) DCSwin. (i) BANet. (j) UNetFormer. (k) LSKNet. (l) Proposed method.

information by FRA can improve the segmentation accuracy for small-scale categories.

In addition, as shown in Table 1, we explore the joint impact between modules. In Vaihingen, adding DLAM and BAE raises mIoU and F1 by 5.03% and 3.42%, respectively. The insertion of DLAM and FRA boosts mIoU by 6.78% and F1 by 4.48%. When BAE and FRA are combined, mIoU increases by 6.45%, and F1 climbs by 4.29%. When all three modules (DLAM, BAE, and FRA) are used simultaneously, mIoU is significantly improved by 7.05%, F1 gains a significant increase of 4.67%, and OA rises by 2.05%.

5) THE EFFECTS OF LOSS FUNCTION

The loss function designed for the LGBSwin, the joint loss $Loss_{total}$ used in this study, has notably enhanced the performance of the network, with the mIoU metric reaching 83.71% and 87.73% for the Vaihingen and the Potsdam, respectively, verifying the effectiveness of the loss function. Table 4 illustrates the performance comparisons of different balancing factor α values. Upon examination of Fig. 8, it becomes apparent that the model exhibits superior performance when the α is set to 0.4, denoted as a result (d). The model demonstrates a high level of accuracy in segmenting dense cars and tiny buildings in both the first and second rows. The model reveals sensitivity to segmentation boundaries in the third row. The presence of "Low veg" obstructing the road in the fourth row is prone to category confusion, and the model enhances the precision of identifying confusing categories.

D. PERFORMANCE COMPARISON

In order to evaluate the efficacy of LBGSwin, we conduct a comparative analysis with established methodologies on two distinct datasets, namely the Vaihingen and the Potsdam. Experiments are conducted under consistent conditions. In the context of the comparison model configuration, we select classical CNN-based semantic segmentation models, including UNet, DeepLabV3, and so on. In addition, our comparison also includes Transformer-based network models, such as DCSwin.

Table 5 presents a comparative analysis of different methods applied to the Vaihingen. The mIoU achieves a value of 83.71%, the F1 score is 91.02%, and the OA is 91.53%. When comparing LBGSwin to the CNN-based model DeepLabV3, it is shown that LBGSwin exhibits a 5.1% gain in mIoU, a 3.13% rise in F1, and a 2.76% improvement in OA. Compared with DCSwin, which also uses Swin Transformer as the backbone, we observe that the IoU of "Imp. Surf." and "Building" are not optimal. However, the advantage of the proposed method lies in the accurate semantic segmentation of the overall scene, as evidenced by the 0.49% increase in mIoU. Compared to LSKNet, which can dynamically adjust the receptive field, LGBSwin's mIoU and F1 are ahead by 0.62% and 0.38%, respectively. Based on the evaluation metrics, it is obvious that our method demonstrates greater performance compared to the other methods. Fig. 9 visualizes the superiority of LBGSwin. In the first row, LGBSwin eliminates car segmentation errors induced by light shading. This demonstrates the global modeling capability of LBGSwin, which effectively improves the segmentation accuracy of occluded objects. LBGSwin performs strongly in small target segmentation within the dense car zone in the second row, effectively using local information. In the third row, LGBSwin successfully segments the shadow-disturbed buildings, verifying the global context capture capability of the model.

Table 6 illustrates the results of comparing different methods on the Potsdam dataset. The mIoU, F1, and OA metrics reached 87.83%, 93.35%, and 91.87%, correspondingly. Compared to SPANet using spatial pyramid attention, LBGSwin demonstrates a notable increase in mIoU by 7.23% and an improvement in F1 by 4.22%. Compared to UNetFormer, which uses a lightweight ResNet 18 as its encoder, LBGSwin demonstrates better performance with a 1.33% increase in mIoU and a 0.72% increase in F1. Compared with DCSwin, LBGSwin's excellent performance in other categories makes up for the disadvantage in "Imp.

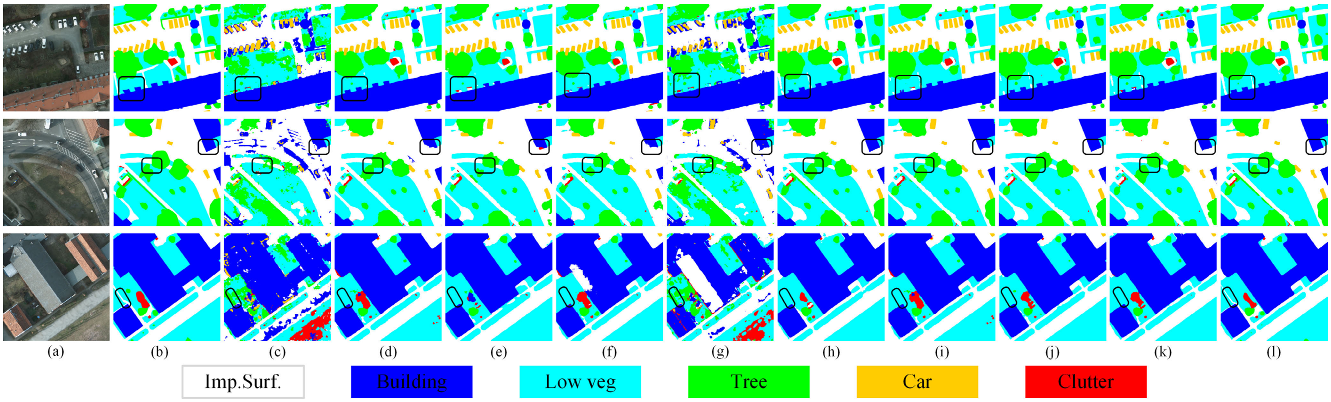


FIGURE 10. Comparison of visualization results on the Potsdam. (a) Image. (b) Ground truth. (c) FCN. (d) UNet. (e) DeepLabV3. (f) SPANet. (g) DANet. (h) DCSwin. (i) BANet. (j) UNetFormer. (k) LSKNet. (l) Proposed method.

TABLE 5. Comparison of different methods on the Vaihingen dataset.

Method	IoU					Evaluation metrics		
	Imp.Surf.	Building	Low veg	Tree	Car	mIoU	F1	OA
FCN [13]	74.78	75.96	60.00	74.53	46.60	66.38	79.18	83.01
UNet [14]	75.63	77.75	61.96	74.77	38.68	65.76	78.29	83.72
DeepLabV3 [16]	82.73	87.09	70.47	79.91	72.84	78.61	87.89	88.77
SPANet [32]	86.80	91.73	73.81	82.62	80.50	83.09	90.64	91.15
DANet [31]	68.11	72.33	57.35	72.54	38.27	61.72	75.46	80.34
DCSwin [38]	**87.77**	**92.70**	74.55	82.48	80.93	83.22	90.71	**91.63**
BANet [48]	86.53	91.00	72.61	81.62	78.35	82.02	89.98	90.68
UNetFormer [37]	86.36	91.21	72.70	81.56	75.14	81.39	89.58	90.67
LSKNet [30]	86.80	91.73	73.81	82.62	80.50	83.09	90.64	91.15
Ours	87.62	92.33	**74.73**	**82.70**	**81.20**	**83.71**	**91.02**	91.53

TABLE 6. Comparison of different methods on the Potsdam dataset.

Method	IoU					Evaluation metrics		
	Imp.Surf.	Building	Low veg	Tree	Car	mIoU	F1	OA
FCN [13]	64.84	62.39	55.64	46.75	53.42	56.61	72.07	72.26
UNet [14]	82.53	87.40	75.26	77.45	89.11	82.55	90.34	88.50
DeepLabV3 [16]	81.85	85.14	72.66	74.50	89.09	80.65	89.15	87.23
SPANet [32]	81.80	85.04	72.98	74.26	88.91	80.60	89.13	87.07
DANet [31]	65.19	65.45	55.94	49.52	60.33	59.28	74.26	73.60
DCSwin [38]	**88.57**	94.86	79.57	81.40	93.08	87.50	93.21	91.84
BANet [48]	86.10	93.10	76.94	79.95	92.16	85.65	92.14	90.61
UNetFormer [37]	87.46	93.73	78.48	79.82	93.00	86.50	92.63	91.11
LSKNet [30]	86.57	92.38	77.50	80.18	92.27	85.78	92.23	90.62
Ours	88.46	**94.90**	**79.82**	**81.76**	**93.72**	**87.83**	**93.35**	**91.87**

TABLE 7. Comparison of the computational costs.

Method	Backbone	FLOPs(G)	Params(M)	mIoU(%)
FCN [13]	ResNet50	21.59	23.52	66.38
UNet [14]	-	160.83	17.26	65.76
DeepLabV3 [16]	ResNet50	26.64	43.24	78.61
SPANet [32]	ResNet50	60.36	34.03	83.09
DANet [31]	ResNet50	27.78	47.64	61.72
DCSwin [38]	Swin-S	70.15	66.90	83.22
BANet [48]	ResT-Lite	13.06	12.73	82.02
UNetFormer [37]	ResNet18	11.74	11.68	81.39
LSKNet [30]	LSKNet-S	15.73	14.35	83.09
Ours	Swin-S	55.23	66.53	83.71

Surf.”, enabling mIoU to achieve optimal performance. This comprehensiveness advantage stems from the model’s ability to effectively mine complex urban information and differentiate between multiple classes of features. BANet [48] is a bilateral structure that combines Transformer and convolution, and LBGSwin is still 1.21% higher on the F1 metric. Compared with LSKNet, the F1 and OA of the proposed method increase by 1.12% and 1.25%, respectively. The results mentioned above illustrate the advancements achieved by the LBGSwin. Fig. 10 displays the segmentation capability of LBGSwin. In the first and second rows, LBGSwin avoids the interference of “Low veg” and roads, and successfully recognizes the “Building” region. In the third row, “Low veg” is exceedingly similar to roads, and LGBSwin correctly divides the “Low veg” category. This indicates that LBGSwin leverages boundary information to boost segmentation accuracy for similar categories.

Calculating costs is crucial for evaluating the network. Table 7 demonstrates the comparison of the computational costs of the involved methods. Specifically, the table contains mIoU for the Vaihingen dataset, FLOPs, and parameters. Compared to UNet, LGBSwin has a significant advantage in terms of computational cost. This indicates that the proposed method utilizes computational resources more efficiently while maintaining high performance. Compared to DCSwin, which also uses Swin-S as a backbone, our method performs better in terms of both computational cost and performance. Although our parameters are lower than DCSwin, some models still perform better in this regard, and future work could consider optimizing the model structure to reduce the number of parameters. Overall, the proposed method balances computational cost and performance in urban remote sensing semantic segmentation tasks, showing potential superiority. LGBSwin is highly competitive in scenarios with high requirements for semantic segmentation accuracy.

IV. CONCLUSION

In this study, we propose LGBSwin as a novel approach for acquiring features in remote sensing images to enhance semantic segmentation. The DLAM incorporates spatial linear attention mechanisms and channel linear attention mechanisms to enhance global modeling capabilities. The BAE is designed to tackle the issue of boundary ambiguity by leveraging boundary information. The FRA establishes the interaction between local and global information, alleviating semantic ambiguity resulting in the loss of local information. Furthermore, a combinatorial loss function is employed to optimize the LGBSwin network. Experimental results show that the proposed method achieves better performance on two public datasets in terms of both numerical metrics and visual results. In future research, we will further achieve more balanced and superior segmentation performance for all categories while maintaining a high level of mIoU.

REFERENCES

[1] P. He, L. Jiao, R. Shang, S. Wang, X. Liu, D. Quan, K. Yang, and D. Zhao, ''MANet: Multi-scale aware-relation network for semantic segmentation in aerial scenes,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[2] X. Lu, L. Jiao, L. Li, F. Liu, X. Liu, S. Yang, Z. Feng, and P. Chen, ''Weak-to-strong consistency learning for semisupervised image segmentation,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[3] B. Guo, J. Zhang, and X. Li, ''River extraction method of remote sensing image based on edge feature fusion,'' *IEEE Access*, vol. 11, pp. 73340–73351, 2023.

[4] W. Qiao, L. Shen, J. Wang, X. Yang, and Z. Li, ''A weakly supervised semantic segmentation approach for damaged building extraction from postearthquake high-resolution remote-sensing images,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, Mar. 2023.

[5] M. H. Asad and A. Bais, ''Crop and weed leaf area index mapping using multi-source remote and proximal sensing,'' *IEEE Access*, vol. 8, pp. 138179–138190, 2020.

[6] H. S. Ullah, M. H. Asad, and A. Bais, ''End to end segmentation of canola field images using dilated U-Net,'' *IEEE Access*, vol. 9, pp. 59741–59753, 2021.

[7] X. Li, Z. Zhang, S. Lv, M. Pan, Q. Ma, and H. Yu, ''Road extraction from high spatial resolution remote sensing image based on multi-task key point constraints,'' *IEEE Access*, vol. 9, pp. 95896–95910, 2021.

[8] S. Wang, X. Hou, and X. Zhao, ''Automatic building extraction from high-resolution aerial imagery via fully convolutional encoder–decoder network with non-local block,'' *IEEE Access*, vol. 8, pp. 7313–7322, 2020.

[9] H. Gao, L. Cao, D. Yu, X. Xiong, and M. Cao, ''Semantic segmentation of marine remote sensing based on a cross direction attention mechanism,'' *IEEE Access*, vol. 8, pp. 142483–142494, 2020.

[10] D. Feng, Z. Zhang, and K. Yan, ''A semantic segmentation method for remote sensing images based on the Swin transformer fusion Gabor filter,'' *IEEE Access*, vol. 10, pp. 77432–77451, 2022.

[11] X. Liu, L. Jiao, L. Li, L. Cheng, F. Liu, S. Yang, and B. Hou, ''Deep multiview union learning network for multisource image classification,'' *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 4534–4546, Jun. 2022.

[12] X. Liu, L. Li, F. Liu, B. Hou, S. Yang, and L. Jiao, ''GAFnet: Group attention fusion network for PAN and MS image high-resolution classification,'' *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10556–10569, Oct. 2022.

[13] J. Long, E. Shelhamer, and T. Darrell, ''Fully convolutional networks for semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[14] O. Ronneberger, P. Fischer, and T. Brox, ''U-Net: Convolutional networks for biomedical image segmentation,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Oct. 2015, pp. 234–241.

[15] 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.

[16] 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. (ECCV)*, Sep. 2018, pp. 801–818.

[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, ''Rethinking atrous convolution for semantic image segmentation,'' 2017, *arXiv:1706.05587*.

[18] J. Ji, X. Lu, M. Luo, M. Yin, Q. Miao, and X. Liu, ''Parallel fully convolutional network for semantic segmentation,'' *IEEE Access*, vol. 9, pp. 673–682, 2021.

[19] C. You, L. Jiao, X. Liu, L. Li, F. Liu, W. Ma, and S. Yang, ''Boundary-aware multi-scale learning perception for remote sensing image segmentation,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023

[20] Z. Xu, W. Zhang, T. Zhang, and J. Li, ''HRCNet: High-resolution context extraction network for semantic segmentation of remote sensing images,'' *Remote Sens.*, vol. 13, no. 1, p. 71, Dec. 2020.

[21] K. Sun, B. Xiao, D. Liu, and J. Wang, ''Deep high-resolution representation learning for human pose estimation,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.

[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, pp. 1–5, 2023.

[23] B. Bai, W. Fu, T. Lu, and S. Li, ''Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.

[24] R. Shang, M. Liu, L. Jiao, J. Feng, Y. Li, and R. Stolkin, ''Region-level SAR image segmentation based on edge feature and label assistance,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[25] H. Xu, X. Cui, C. Li, Z. Tian, J. Liu, and J. Yang, ''RUnT: A network combining residual U-Net and transformer for vertebral edge feature fusion constrained spine CT image segmentation,'' *IEEE Access*, vol. 11, pp. 55692–55705, 2023.

[26] R. Zhao, W. Chen, and G. Cao, ''Edge-boosted U-Net for 2D medical image segmentation,'' *IEEE Access*, vol. 7, pp. 171214–171222, 2019.

[27] F. Yu and V. Koltun, ''Multi-scale context aggregation by dilated convolutions,'' 2015, *arXiv:1511.07122*.

[28] K. He, X. Zhang, S. Ren, and J. Sun, ''Spatial pyramid pooling in deep convolutional networks for visual recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[29] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, ''Visual attention network,'' *Comput. Vis. Media*, vol. 9, no. 4, pp. 733–752, Jul. 2023.

[30] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, ''Large selective kernel network for remote sensing object detection,'' 2023, *arXiv:2303.09030*.

[31] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, ''Dual attention network for scene segmentation,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.

[32] J. Guo, X. Ma, A. Sansom, M. McGuire, A. Kalaani, Q. Chen, S. Tang, Q. Yang, and S. Fu, ''Spanet: Spatial pyramid attention network for enhanced image recognition,'' in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.

[33] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, ''ECA-Net: Efficient channel attention for deep convolutional neural networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ''An image is worth 16×16 words: Transformers for image recognition at scale,'' Jun. 2020, *arXiv:2010.11929*.

[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ''Swin transformer: Hierarchical vision transformer using shifted windows,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[37] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, ''UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery,'' *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022.

[38] 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.

[39] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, ''Swin transformer embedding unet for remote sensing image semantic segmentation,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[40] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1911–1920.

[41] 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.

[42] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.

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

[44] A. Stergiou, R. Poppe, and G. Kalliatakis, "Refining activation downsampling with SoftPool," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10357–10366.

[45] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

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

[47] Q.-L. Zhang and Y.-B. Yang, "SA-Net: Shuffle attention for deep convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 2235–2239.

[48] 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, p. 3065, Aug. 2021.

RUI LIN (Graduate Student Member, IEEE) received the B.E. degree from Hainan University, China, in 2021. She is currently pursuing the M.S. degree with Guangxi University. Her research interests include semantic segmentation, computer vision, and deep learning.

YING ZHANG received the B.S. degree from the Central South University of Forestry and Technology, China, in 2021. She is currently pursuing the M.S. degree with Guangxi University. Her research interests include object detection, computer vision, and deep learning.

XUE ZHU received the B.E. degree from Chang'an University, Xi'an, China, in 2021. She is currently pursuing the M.S. degree with Guangxi University. Her research interests include object detection, computer vision, and deep learning.

XUEYUN CHEN is currently an Associate Professor and a Ph.D. Supervisor with the School of Electrical Engineering, Guangxi University. His research interests include target detection and recognition in remote sensing images, semantic segmentation, face detection, road detection, and automatic driving.