applied sciences

MDPI

Article

Segmenting Urban Scene Imagery in Real Time Using an Efficient UNet-like Transformer

Haiqing Xu [1], Mingyang Yu [1,*], Fangliang Zhou [1] and Hongling Yin [2]

[1] School of Surveying and Geo-Informatics, Shandong Jianzhu University, Jinan 250101, China; 2021160105@stu.sdjzu.edu.cn (H.X.); 2021165123@stu.sdjzu.edu.cn (F.Z.)
[2] School of Architecture and Urban Planning, Shandong Jianzhu University, Jinan 250101, China; 12949@sdjzu.edu.cn
* Correspondence: ymy@sdjzu.edu.cn

Abstract: Semantic segmentation of high-resolution remote sensing urban images is widely used in many fields, such as environmental protection, urban management, and sustainable development. For many years, convolutional neural networks (CNNs) have been a prevalent method in the field, but the convolution operations are deficient in modeling global information due to their local nature. In recent years, the Transformer-based methods have demonstrated their advantages in many domains due to the powerful ability to model global information, such as semantic segmentation, instance segmentation, and object detection. Despite the above advantages, Transformer-based architectures tend to incur significant computational costs, limiting the model's real-time application potential. To address this problem, we propose a U-shaped network with Transformer as the decoder and CNN as the encoder to segment remote sensing urban scene images. For efficient segmentation, we design a window-based, multi-head, focused linear self-attention (WMFSA) mechanism and further propose the global–local information modeling module (GLIM), which can capture both global and local contexts through a dual-branch structure. Experimenting on four challenging datasets, we demonstrate that our model not only achieves a higher segmentation accuracy compared with other methods but also can obtain competitive speeds to enhance the model's real-time application potential. Specifically, the mIoU of our method is 68.2% and 52.8% on the UAVid and LoveDA datasets, respectively, while the speed is 114 FPS, with a 1024 × 1024 input on a single 3090 GPU.

Keywords: urban scene imagery; semantic segmentation; vision transformer; linear attention

Citation: Xu, H.; Yu, M.; Zhou, F.; Yin, H. Segmenting Urban Scene Imagery in Real Time Using an Efficient UNet-like Transformer. *Appl. Sci.* **2024**, *14*, 1986. https://doi.org/10.3390/app14051986

Academic Editor: Andrea Prati

Received: 16 January 2024
Revised: 21 February 2024
Accepted: 23 February 2024
Published: 28 February 2024

1. Introduction

In recent years, with the rapid development of satellite and sensor technologies, a large number of remote sensing urban scene images can be easily obtained for a variety of different applications, such as semantic segmentation [1,2], image classification [3,4], and target detection [5]. As a classic computer field method, semantic segmentation aims to take image pixels as the basic segmentation unit to accurately classify the category attributes of each pixel on the image, which leads to more in-depth applications, such as road and building extraction [6,7], land use and land cover (LULC) [8,9], and change detection [10,11]. For example, we can quickly and accurately obtain the basic information of different buildings in the city through automatic building extraction, contributing to the smart cities' construction. Extracting urban green space is conducive to monitoring urban greening and contributing to a more comfortable urban environment. Recently, the deep learning technology has rapidly developed. As a representative branch, CNN has been the mainstream in the segmentation field for many years. CNN-based methods are better at capturing rich local spatial relationships compared with traditional machine learning methods, such as random forests (RF) [12] and support vector machines (SVM) [13]. In addition, the powerful feature representation capability has made them quickly surpass the traditional method to become a mainstream segmentation method in many fields.

Despite the above advantages, the CNN-based methods use the convolution kernel with a fixed receptive field to extract local information, which limits its potential to model long-range dependencies or global spatial information. For the tasks of semantic segmentation, the lack of global contextual information can produce ambiguous and low-accuracy segmentation results. To cope with this problem, some studies [14,15] utilized the attention mechanism to improve the model's potential to model global context. Although the attention mechanism alleviates the above problems to some extent, it still cannot get rid of the local nature of convolution operations and incurs additional computational costs, thus limiting the model's efficiency in different tasks and the potential in real-time applications.

Recently, the emergence of Transformer [16] has provided a new way to solve the above problems. The natural language processing (NLP) tasks first used Transformer to process sequential data, and its great success inspired scholars to explore its application potential in other fields, such as multimodal fusion and computer vision (CV). By eliminating convolution operations and relying solely on efficient multi-layer perceptron (MLP) and multi-head self-attention mechanisms, the Transformer is better at capturing the global context. In the semantic segmentation and image classification tasks, there have been many studies demonstrating that models based on the Transformer structure have strong global information modeling capabilities, which can significantly improve the final classification or segmentation results. Zheng et al. [3] combined two branches (CNN and ViT) to design an efficient multi-branch Transformer for imagery classification in remote sensing, which reduced the model complexity and improved the long-range dependency modeling capability. Wang et al. [17] proposed a U-shaped novel network structure based on the CNN encoder and Transformer decoder for high-resolution urban scene imagery semantic segmentation, and the powerful global information modeling module in the decoder enables it to obtain good results on challenging datasets. Wang et al. [18] introduced the Swin Transformer as the main encoder to extract contextual semantic information; meanwhile, a densely connected feature aggregation module (DCFAM) was designed in the decoder to improve the final accuracy. Zhang et al. [19] designed a Transformer- and CNN-based hybrid network to segment the high-resolution urban imagery, and a spatial pyramid pooling module and an auxiliary boundary detection strategy were applied to obtain better segmentation performance. However, although the self-attention mechanism based on the Softmax function is effective, the quadratic computation complexity, O (n^2), directly related to the length of the input sequence n, leads to great computational costs, which limits the efficiency and real-time application capability of the model. Some studies have solved this problem by limiting the global receptive field to a smaller range, such as using sparse global attention [20,21] or a smaller attention window [22,23]. Although effective, these methods also pose the following two problems: (1) useful information in other regions (outside the window) may be overlooked, and (2) the model's potential to model global context is sacrificed to some extent.

On the other hand, the linear attention mechanism (LNA) [24] provides a new way to address the trade-off between complexity and accuracy performance in Transformer-based models. Specifically, LNA uses a separated kernel function to replace the Softmax function in the self-attention mechanism [25]. As a result, LNA need not compute the pair-wise similarity first as Softmax attention (QK^T). In other words, LNA can adjust the computation order by computing the KV^T first based on the associative law of matrix multiplication (Figure A1), as a result, decreasing the computational complexity from O (N^2d) to O (Nd^2). The channel dimension, d, is usually much smaller than the token number N in modern classical Vision Transformers. Nevertheless, it is still a complex problem to design a module that is as efficient as Softmax self-attention when applying the LNA. Efficient Vit [26] used depth-wise separable convolution to improve the local information modeling capability of the LNA module to achieve a higher classification accuracy. Castling-ViT [27] proposed the idea of a linear angular kernel to model the spectral similarity between each Q and K. Hydra attention [28] used the cosine similarity to simulate the Softmax function to reduce the computational complexity to O (Nd). Nystromformer [29] and SOFT [30]

approximated the Softmax self-attention matrix by matrix decomposition to achieve high efficiency while reducing the computational complexity. There have been many efforts by scholars to outperform the Softmax attention mechanism, yet current LNA-based models face a dilemma between model complexity and performance. Firstly, a simple simulation of the Softmax self-attention function [26] would seriously affect the model performance capability; secondly, elaborately designed and complex kernel function [31] or matrix transformation methods [29,30] would incur additional computation costs and influence the model's real-time application capability. Therefore, there are still significant challenges in utilizing LNA-based models efficiently and effectively.

Currently, the model's segmentation accuracy and efficiency are rarely considered simultaneously in many existing methods, failing to solve the trade-off problem in real-time application situations. In this study, we aim to design a network that can accurately segment urban scene images while ensuring the model's efficiency. Inspired by the advanced work in the Flatten Transformer [25], which can obtain better performance than Softmax attention while incurring lower computation complexity, we designed a new global–local information modeling module (GLIM) based on the proposed WMFSA and constructed a UNet-like network (Flauformer) based on the CNN encoder and Transformer decoder. Furthermore, the window shift operation in the Swin Transformer is effective but incurs additional computation overhead. Therefore, we designed an efficient spatial position transformation module (SPT) to integrate into the GLIM, which can achieve efficient cross-window information interaction with lower computational costs. Meanwhile, a depth-separable convolution was used before the MLP module to enhance the local connection between neighboring windows. Finally, we conducted experiments on four challenging datasets and compared the results with more than ten state-of-the-art models, proving that our model has a better comprehensive performance in the semantic segmentation tasks of urban scene imagery. The main contributions of this paper are summarized as follows:

(1) Aiming at the inability of CNN-based methods to effectively model long-range dependence or global spatial relationships, a global–local spatial information modeling module (GLIM) is proposed, which can simultaneously model global and local spatial information of urban scene distributions.
(2) A spatial position transformation (SPT) module is designed to solve the trade-off problem between the model complexity and global information modeling capability in the existing Vision Transformer method.
(3) A UNet-like Transformer architecture is designed for segmenting remote sensing urban images in real time, which can obtain high-precision segmentation results while maintaining the model's efficiency.

2. Methods

As shown in Figure 1, the proposed Flauformer is an encoding–decoding structure that is different from many Transformer-based encoding and decoding structures, as most of them are pure Transformer structures or the encoder is a Transformer and the decoder is CNN. On the contrary, our encoder consists of simple CNNs to minimize the model complexity, and the decoder is composed of the Transformer, which constitutes a structure based on CNN encoding and Transformer decoding to accomplish the real-time segmentation tasks of high-resolution urban scene imagery. In the following, we describe the details of the proposed method.

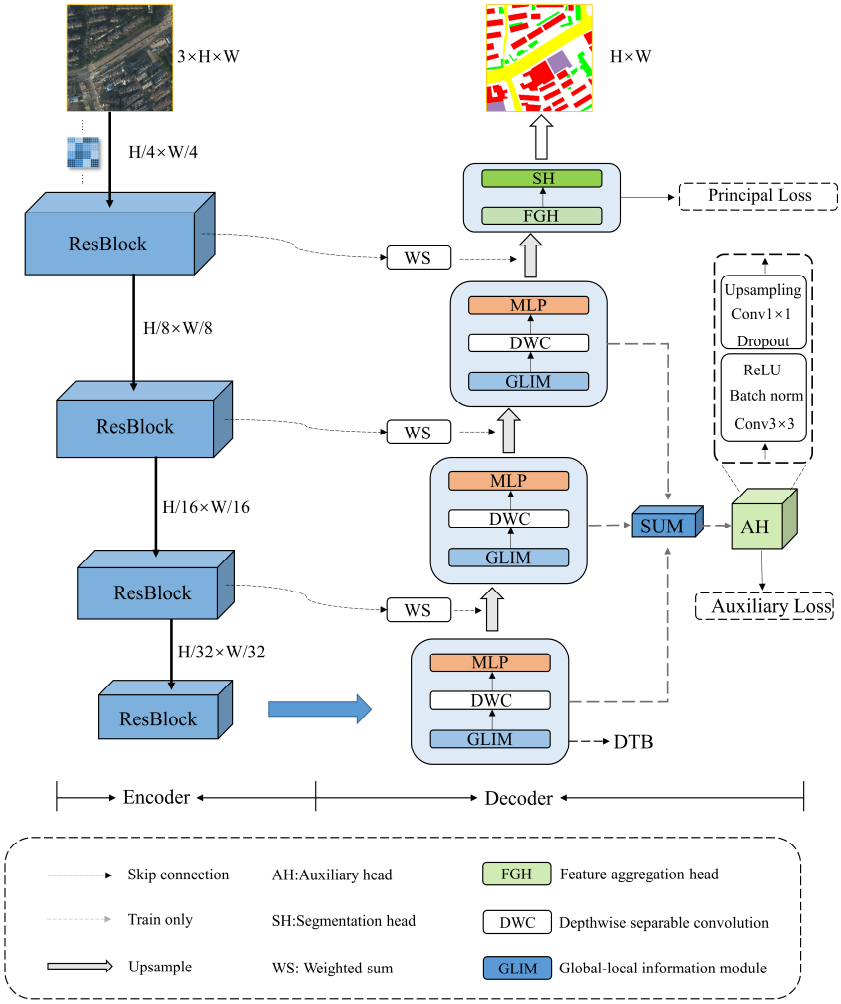


Figure 1. Overall structure of our model.

2.1. Encoder

As a widely used benchmark network, we can see the existence of residual networks [32] in many studies. As one of them, ResNet18 can accurately extract multi-scale semantic features while incurring relatively lower computational costs, which is more favorable to enhancing the model's real-time application ability. Therefore, the ResNet18 is used in our encoder to extract multi-scale semantic information from urban scene images. ResNet18 consists mainly of 4 residual blocks, and the size of each stage's feature map is $1/2$ of the previous stage. In our model, the feature maps generated in the encoder are fused with the counterparts in the decoder by using a 1×1 convolution. Specifically, the feature information generated in the encoder is efficiently fused with the context information generated by GLIM in the decoder by introducing a weighted sum operation. The weighting operation selectively fuses the features that are more beneficial to the final segmentation results; as a result, a more representative aggregated feature is obtained. The formula is shown in Equation (1):

$$AF = \alpha \times EF + (1 - \alpha) \times DF, \tag{1}$$

where AF denotes the aggregated feature after fusion, EF denotes the features at different encoding stages, and DF denotes the features at different decoding stages.

2.2. Decoder

The complexity and diversity (small scale, high similarity, and mutual occlusion) of man-made objects on the ground poses great challenges to the real-time segmentation of remote sensing urban imagery. Meanwhile, aggregating global semantic information is an effective solution. In order to obtain accurate global context information, the existing mainstream methods were modified by adding an attention mechanism to the CNN [2,33] or using Transformer as the main encoder [34]. However, the former cannot eliminate the convolution operation completely and the latter imposes a great computational overhead while losing the features' position information. As shown in Figure 1, we used three DTBs and a feature aggregation module to construct an efficient decoder in the proposed model. Consequently, our model can accurately capture multi-scale global context semantic information while maintaining high efficiency by constructing such a lightweight and hierarchical structure in the decoder.

2.2.1. Focused Linear Attention

First, we analyzed the self-attention in the standard Transformer. Considering N tokens $z \in \mathbb{R}^{N\times C}$, the details of the self-attention calculation in each head are shown in Equation (2):

$$\begin{aligned} Q = zW_Q, K = zW_K, V = zW_V \\ O_i = \sum_{j=1}^{N} \frac{L(Q_i,K_j)}{\sum_{j=1}^{N} L(Q_i,K_j)} V_j \end{aligned}, \tag{2}$$

where W_Q, W_K, and $W_V \in \mathbb{R}^{C\times C}$ are the transformation matrices, and $L()$ denotes the similarity function. In many Transformer architectures, the Softmax attention mechanism is often used to represent this similar relationship. As a result, the algorithm computes the attention between all the query–key pairs, resulting in O (N^2) computation complexity. Due to the computational complexity quadratic to the input, it limits the potential to apply Transformer-based models to real-time scenes. Some studies used sparse global attention or partitioned smaller windows to solve this problem, but these methods sacrificed the model's ability to model global context to some extent. Meanwhile, the linear attention reduces the computational complexity from O (N^2) to O (N) and has the potential to replace the traditional Softmax attention mechanism in Transformer-based models. Specifically, LNA uses a carefully designed kernel function to approximate the similarity function:

$$L(Q,K) = \delta(Q)\delta(K)^T, \tag{3}$$

where we can rewrite the previous formula as:

$$O_i = \sum_{j=1}^{N} \frac{\delta(Q_i)\delta(K_j)^T}{\sum_{j=1}^{N} \delta(Q_i)\delta(K_j)^T} V_j. \tag{4}$$

As a result, we can change the order of self-attention computation based on the matrix multiplication properties, i.e.,

$$O_i = \frac{\delta(Q_i)\left(\sum_{j=1}^{N} \delta(K_j)^T V_j\right)}{\delta(Q_i)\left(\sum_{j=1}^{N} \delta(K_j)^T\right)}, \tag{5}$$

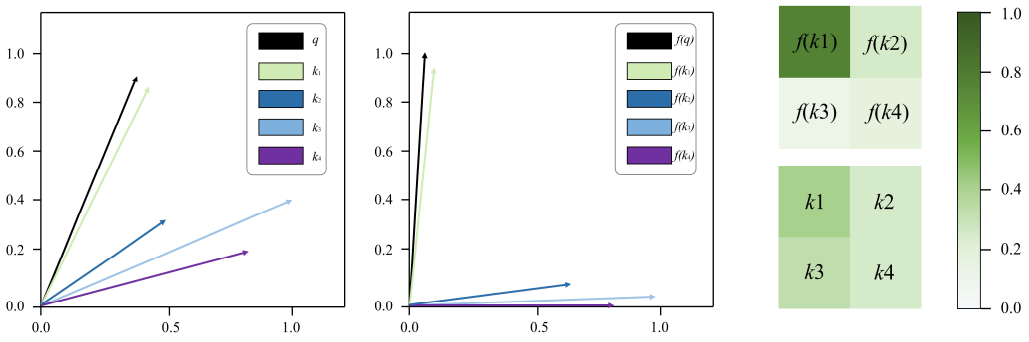
where the computational complexity is reduced to O (N).

However, existing LNA modules are either unable to fully achieve the same performance capability as the Softmax attention mechanism, or they incur additional computational costs from designing complex kernel functions, so directly applying LNA to the Transformer model is not a simple problem. Fortunately, the novel focused linear attention mechanism [25] offers a new approach to solving the problem, which can reduce the self-attention computation complexity while maintaining the performance. On the one hand,

the Softmax attention mechanism has the property of nonlinear weighting, which makes it more likely to focus on the more important image features. However, the traditional LNA has a smoother and more homogeneous attention distribution, which makes it difficult to focus on the information-rich regions of the image. As a remedy, based on the LNA, FLA makes the same query–key pairs closer and different query–key pairs farther away by changing the orientation of each query and key vector. Specifically, a simple function, f_c, is proposed to perform this mapping:

$$\begin{gathered} L(Q_i, K_j) = \delta_c(Q_i)\delta_c(K_j)^T \\ \delta_c(z) = f_c(\mathrm{ReLU}(z)),\ f_c(z) = \frac{\|z\|}{\|z^{**k}\|} z^{**k}, \end{gathered} \tag{6}$$

where z^{**k} denotes z to the power of k, and we found that the norm of the feature remains after mapping ($\|z\| = \|f_c\ (z)\|$), indicating that only the vector direction is changed. As a result, the mapping function f_c can actually affect the attention distribution. As is shown in Figure 2, we can see that f_c makes each vector closer to the axis closest to the vector, and the proximity can be controlled by the parameter k. Therefore, f_c can divide the features into several groups according to the axis closest to them, improving the feature similarity within the same groups and reducing the feature redundancy between different groups.



(**a**) (**b**) (**c**)

Figure 2. The distribution of different vectors for the Softmax attention mechanism and LNA: (**a**) Softmax attention mechanism, (**b**) LNA, and (**c**) numerical visualization.

On the other hand, feature diversity is also the reason for the poor performance of LNA, and one of the deeper reasons behind it is the attention matrix's rank [35,36]. In fact, the rank of the attention matrix in LNA is limited by the number of tokens N and the channel number d:

$$r\left(\delta(Q)\delta(K)^T\right) \leq \min\{r(\delta(Q)),\ r(\delta(K))\} \leq \min\{N, d\}, \tag{7}$$

where the value of d is much smaller than N in the present Vision Transformer, and r denotes the rank. As a result, the rank of the attention matrix tends to change on a smaller scale, which may lead to the multi-line attention maps becoming homogeneous. In addition, the homogeneity of the attention weights has an adverse impact on the final aggregated features. The FLA addresses this problem via a simple strategy. Specifically, a depth-wise separable convolution (*DWC*) module is used to compensate for the deficiency of the self-attention matrix:

$$O = \delta_c(Q)\left(\delta_c(K)^T V\right) + DWC(V). \tag{8}$$

We can better understand the role of the *DWC* by thinking of it as a type of attention. In other words, each query focuses on a few close features instead of all the features, V. This locality ensures that when the LNA encounters a situation where different queries produce the same value, it can still obtain diverse outputs from the local features, thus maintaining the feature representation effectually. From the perspective of rank, the

rank of the attention matrix in LNA may restore to the full rank state when the *DWC* is added, ensuring the feature diversity, as with the Softmax attention mechanism. As a result, the linear complexity ensures that the model can adapt to a larger receptive field while maintaining the same computational complexity and modeling the fine-grained global information.

2.2.2. Spatial Position Transformation Operation

The window partition operation in the Swin Transformer is computation-friendly; however, its receptive field is limited in the windows, which has an adverse impact on the model's global information modeling ability. As a result, it limits the model's potential for some tasks that require high-resolution images, such as semantic segmentation and instance segmentation. Figure 3 illustrates the feature information flow based on window self-attention. We found that the output of a window is only related to the window itself, which hinders the cross-window information interaction and weakens the model's feature representation capability. As a remedy, we performed a position transformation operation to maintain the cross-window information interaction, which can be integrated into the window-based multi-head self-attention (WMSA) mechanism, while only introducing a small computational overhead. After that, a spatial alignment operation was used to restore the feature's original spatial position, which is the inverse of the position transformation operation. In addition, multi-head self-attention based on shifted windows uses the shifted window operation to maintain the information interaction between neighboring windows, which is effective but adds a significant computational cost. In this study, inspired by the work of Wu [37], we used a depth-wise separable convolution between the MLP module and GLIM to improve the connection in adjacent windows. As a result, the model can take into account the connection between neighboring windows and cross-windows simultaneously, enhancing its potential to model global context and feature representation.

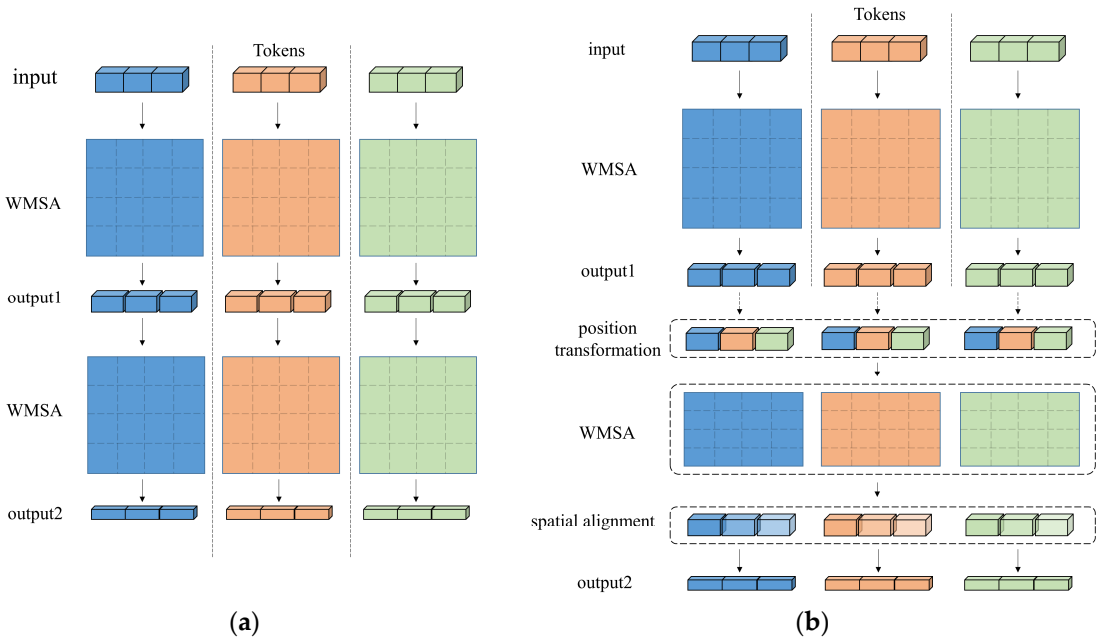


Figure 3. Illustration of the traditional WMSA and the situation after adding SPT: (**a**) standard WMSA and (**b**) WMSA after adding SPT.

2.2.3. Global–Local Information Module

As shown in Figure 4, the GLIM consists of a global path, a local path, a batch normalization layer, a depth-wise separable convolution, and a convolution operation. In the following section, we elaborate the specific process of the GLIM.

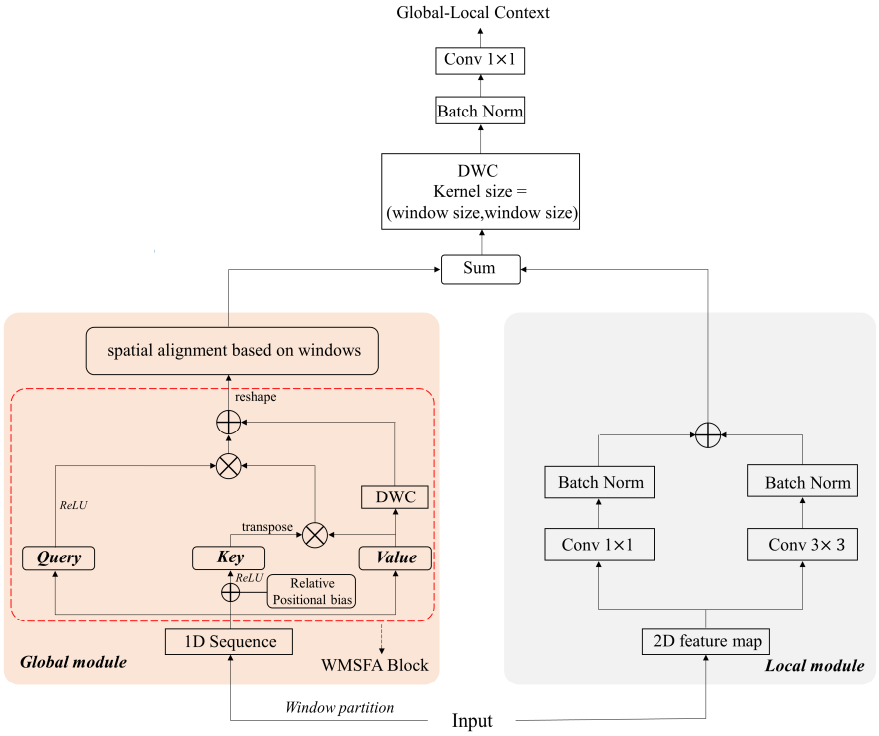


Figure 4. Overall structure of the proposed GLIM.

While global contextual information is important for segmenting complex urban scenes, local context is also crucial to retaining rich spatial details. As shown in Figure 4, for the local branch, we used two groups of parallel convolutions with the kernel sizes of 1 and 3 for local information extraction and, finally, performed an add operation.

The global branch uses a window partition operation and the proposed WMFSA to capture the global information and combines it with the position transformation operation to obtain the final global information. The details about the window partition operation can be found in Figure 5. First, we expanded the number of input channels to three times using a 1×1 convolution; then, the 1D sequence could be obtained by a window partition operation; finally, we merged all the heads to obtain the three pivotal vectors (Q, K, and V) after performing attention computation on each head. The window size and the number of heads were 8, and the channel dimension, C, was 64. For more details about the WMSA, refer to the Swin Transformer [22]. However, conducting self-attention computation in non-overlapping windows is effective, but the lack of connection between different windows destroys the spatial continuity of input features and weakens the model's ability to model global relationships. The Swin Transformer uses the shifted window operation to improve the information interaction between different windows. Although effective, it introduces a great computational cost and limits the model's real-time scene application ability. In this study, we designed the SPT module to replace the shifted window operation to achieve efficient information interaction between windows while only introducing a small computational overhead. This module can be easily integrated into the WMFSA, and more details can be found in Section 2.2.2.

Finally, the global context information generated by the global module was further fused with the local information of the local module. Meanwhile, the final global–local information was obtained after a depth-wise separable convolution, batch normalization, and 1×1 convolution.

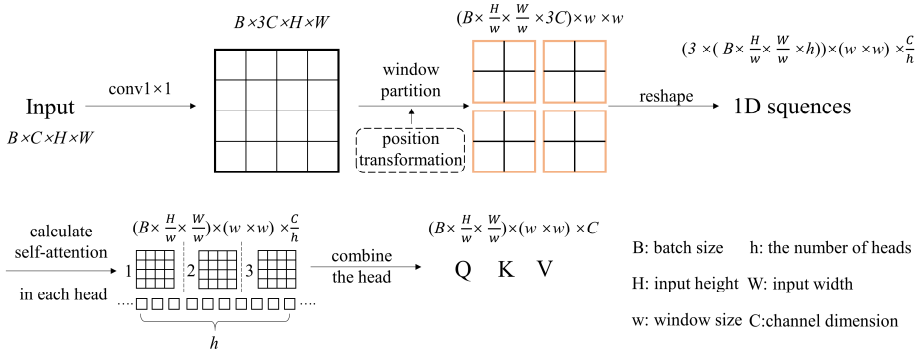


Figure 5. The process of the window partition operation.

2.2.4. Feature Aggregation Head

In the encoding stage, the original input features retain abundant spatial elements, but lack semantic details. In the decoding stage, the output global–local features have fine semantic information, but its spatial resolution is insufficient. Therefore, adding the two features directly may reduce the classification accuracy [38]. In order to deal with this problem, we used a feature aggregation head (FGH) to fuse the two features rationally for more efficient image segmentation.

First, a weighted addition operation was conducted to fuse the two features, and the weights can be updated with model training to take full advantage of the abundant spatial elements and precise semantic information. As shown in Figure 6, the input of FGH is the fusion feature after weighted addition, which is fed into the spatial branch and channel branch, respectively, after processing by a 3 × 3 convolution. Second, the carefully designed paths in the module help to strengthen the channel-based and space-based feature representation. For the spatial path, a depth-wise separable convolution was used to produce the first path feature, $P \in \mathbb{R}^{h \times w \times 1}$, where h and w are the height and width of the feature map. After that, we used a matrix multiplication operation and sigmoid function to obtain the output feature. For the channel path, we used a pooling operation to obtain the second attention feature, $C \in \mathbb{R}^{1 \times 1 \times c}$, where c represents the channel number. Furthermore, the rescale operation utilizes convolution layers (1 × 1) to reduce the channel numbers by a fixed ratio and then restores the channel numbers. Finally, a 1 × 1 convolution was used to classify the refined features of the FGH.

2.2.5. Auxiliary Head

As shown in Figure 1, the outputs of the three decoding blocks were used as the input of the auxiliary header. In addition, the outputs of the two decoding blocks below were resized to perform the summing operation. The header consists of two parts: the first part contains a 3 × 3 convolution, a normalization operation, and a ReLU activation function, while the second part consists of a dropout operation, a 1 × 1 convolution, and an up-sampling operation. Finally, the header's output was used as the training loss of the whole model, together with the main loss.

2.2.6. Loss Function

In the training stage, we not only used the main loss but also utilized an auxiliary loss to optimize the GLIM, where the method is illustrated in Figure 1. Previous studies [39,40] have demonstrated that the use of a multiple-loss structure is more conducive to the model optimization. In our study, we jointly used the main loss and auxiliary loss to optimize the entire model. The main loss, L_m, consists of a cross-entropy loss, L_{ce}, and a dice loss, L_{dc}, which can be defined as follows:

$$
\begin{aligned}
L_{ce} &= -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_k^{(n)} \log \hat{y}_k^{(n)} \\
L_{dc} &= 1 - \frac{2}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{y_k^{(n)}\hat{y}_k^{(n)}}{y_k^{(n)}+\hat{y}_k^{(n)}} \\
L_m &= L_{ce} + L_{dc}
\end{aligned} \tag{9}
$$

where N represents the sample numbers, and K is the total sample classes.$y_k^{(n)}$ and $\hat{y}_k^{(n)}$ represent the sample's label and the model's predicted value, respectively, $n \in [1, 2, \ldots, N]$. Meanwhile, the cross-entropy loss was utilized as the auxiliary loss function: it takes the fusion features of three decoding blocks as input and consists of a 3×3 convolution layer containing ReLU and batch normalization, followed by a convolution (1×1) and an up-sampling operation to produce the final result. Finally, a weight parameter, β ($\beta = 0.4$), was used to combine the two losses efficiently:

$$L = L_p + \beta \times L_{aux} \tag{10}$$

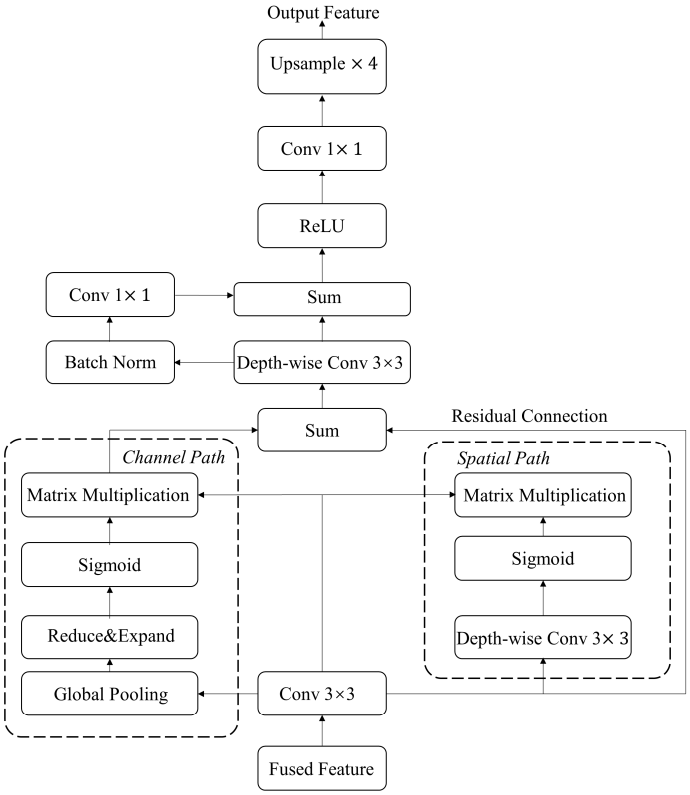


Figure 6. The structure of the FGH.

3. Experiments

3.1. Related Experiment Information

3.1.1. Datasets

Potsdam: The Potsdam dataset consists of 38 high-resolution TOP images with a 6000×6000 pixel size. Each image has four multispectral bands (near infrared, green, blue, and red) as well as a 5 cm ground sampling distance (GSD). The dataset contains one background category (clutter) and five foreground categories (low vegetation, tree, impervious surface, building, and car), and only images in red, green, and blue bands are used. In this paper, we selected the numbers 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 for testing, ID 2_10 for validation, and the remaining 22 images (except for 1 incorrectly labeled image) were used to train the model. Moreover, we only used the TOP image in the experiment, and all the images were further cropped to 1024×1024 px patches.

Vaihingen: The Vaihingen dataset contains 33 high-resolution TOP images (GSD 9 cm) with an average size of 2496 × 2064, and 3 bands (near infrared, red, and green) are utilized. Similar to the Potsdam dataset, this dataset consists of six classes (low vegetation, tree, impervious surface, car, clutter, and building). Image IDs 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, and 38 were used for testing, ID 30 for validation, and the remaining 15 images for training. In our experiment, all the images were further cropped into 1024 × 1024 px patches.

LoveDA: The LoveDA dataset [41] consists of 5987 urban and rural remote sensing images (GSD 0.3 m) with a size of 1024 × 1024 px. The dataset was collected from three cities in China (Wuhan, Changzhou, and Nanjing) and consists of two scenes (rural and urban) with a total of seven categories (forest, road, barren, agriculture, building, water, and background). As a result, the particularity of the dataset (diverse scene styles) poses a great challenge for segmenting it. Furthermore, we used 2540 images for training, 1647 images for validation, and 1800 images for testing. It is worth noting that the official dataset does not provide the image label in the test set, and the online test channel is available.

UAVid: UAVid [42] is a high-resolution unmanned aerial vehicle (UAV) image dataset, which mainly focuses on urban street imagery. The dataset contains images at two resolutions (3840 × 2160 and 4096 × 2160) and eight categories (moving car, static car, building, road, tree, vegetation, human, and clutter). Segmenting the UAVid dataset is a challenging task due to the high resolution, complex scenes, and diverse perspectives. The dataset consists of 42 sequences containing a total of 420 images. In addition, 210 images were used to train, 60 images to validate, and 150 images to test. Each image was further cropped into 1024 × 1024 px patches. Similar to the last dataset, online testing can be selected for accuracy evaluation.

3.1.2. Implementation Details

We used the Pytorch (1.7.0) deep learning framework, with an NVIDIA GTX 3090 GPU. The AdamW optimizer was utilized to accelerate the model convergence. Furthermore, the initial learning rate was set to 1×10^{-4}, and every 5 epochs became 0.98 times the original.

For the UAVid dataset, we used the random flip operation for data augmentation in the training phase, and the batch size was set to 8. In the testing phase, the test time augmentation (TTA) strategies were utilized to enhance the model's performance.

For the LoveDA, Potsdam, and Vaihingen datasets, we randomly cropped the images into a size of 512 × 512. In the training phase, random scale and random flip augmentations were adopted to enhance the sample diversity, and the batch size was set to 16. In the testing process, the augmentation techniques, such as random flip and multi-scale, were used.

3.1.3. Evaluation Metrics

We used two types of metrics to evaluate the model performance. The first one was used to evaluate the model accuracy, including the mean F1 score (F1), mean intersection over union (mIoU), and overall accuracy (OA), and the calculation details are presented in the following equations. The second one was used to evaluate the model's efficiency and potential for real-time scene applications. We used the frames per second (FPS) to evaluate the model speed, the number of model parameters (M) to evaluate the memory requirement, and the floating-point operations count (Flops) to evaluate the model complexity.

$$\begin{array}{l} OA = \frac{TP + TN}{TP + TN + FP + FN} \\ mIoU = \frac{TP}{TP + FP + FN} \\ precision = \frac{TP}{TP + FP} \\ recall = \frac{TP}{TP + FN} \\ F1 = \frac{2 \times precision \times recall}{precision + recall} \end{array}, \tag{11}$$

where TP is the number of correctly classified positive pixels, TN is the number of correctly classified negative pixels, FP is the number of misclassified negative pixels (it is originally

a negative class), and *FN* is the number of misclassified positive pixels (it is originally a positive class).

3.1.4. Comparison Models

We conducted comparison experiments using twelve state-of-the-art models, which can be classified into six categories, including:

(a) CNN-based lightweight semantic segmentation networks: Fast-SCNN [38] and ABCNet [2];
(b) CNN-based attention networks: DANet [43] and MANet [33];
(c) Novel networks specialized in semantic segmentation of remote sensing images: A2FPN [44], EANet [45], and VPANet [46];
(d) Hybrid U-shaped network with Transformer as the encoder and CNN as the decoder: DC-Swin [18] and BANet [47];
(e) Encoders and decoders are all Transformer-based U-shaped networks (pure Transformer structure): SwinUNet [48] and SegFormer [49];
(f) Hybrid U-shaped network with CNN as the encoder and Transformer as the decoder: UNetFormer [17].

3.2. Comparison Experiments

3.2.1. Results on the Vaihingen Dataset

The Vaihingen dataset is one of the most widely used datasets for the semantic segmentation of remote sensing imagery, and many advanced models have achieved good results on this dataset. As shown in Table 1, our method obtained the best F1 score (90.7%) and OA (91.3%), outperforming other lightweight networks. Moreover, our model obtained the highest F1 score on three categories (low vegetation, tree, and impervious surface); specifically, the category impervious surface was at least 1.1% higher than with other models. In addition, as shown in Figure 7, we visualized the comparison experiment results, and the areas that need to be focused on are shown in the red box, which further demonstrates the superiority of the proposed method.

Table 1. Quantitative evaluation results on the Vaihingen test set. The highest values are marked in bold.

Models	Backbone	Lowveg.	Tree	Car	Building	Impsurf.	MeanF1	OA
Fast-SCNN	–	80.1	82.6	62.1	87.2	90.4	79.3	84.5
ABCNet	ResNet18	84.5	89.7	85.3	95.2	92.7	88.5	90.3
DANet	–	82.3	87.4	64.5	93.6	90.1	82.1	87.2
MANet	ResNet50	84.7	90.2	**88.6**	95.4	93.0	90.4	91.0
A2FPN	ResNet18	80.7	87.2	71.5	93.3	93.2	86.3	89.7
EANet	ResNet18	83.2	89.1	80.2	94.5	91.6	87.6	90.1
DC-Swin	Swin-Tiny	85.3	90.3	87.8	**96.2**	93.4	90.5	91.2
BANet	ResT-Lite	83.7	89.9	86.8	95.4	92.3	89.6	90.6
SwinUNet	Swin-Tiny	84.3	88.7	85.3	95.1	93.3	89.1	90.4
VPANet	ResNet50	80.2	85.6	72.4	93.1	92.5	85.7	89.2
SegFormer	MiT-B1	80.7	87.2	71.5	93.3	93.2	86.9	90.3
UNetFormer	ResNet18	84.9	90.1	**88.3**	95.3	92.7	90.4	90.9
Our model	ResNet18	**85.6**	**90.6**	87.9	95.8	**94.5**	**90.7**	**91.3**

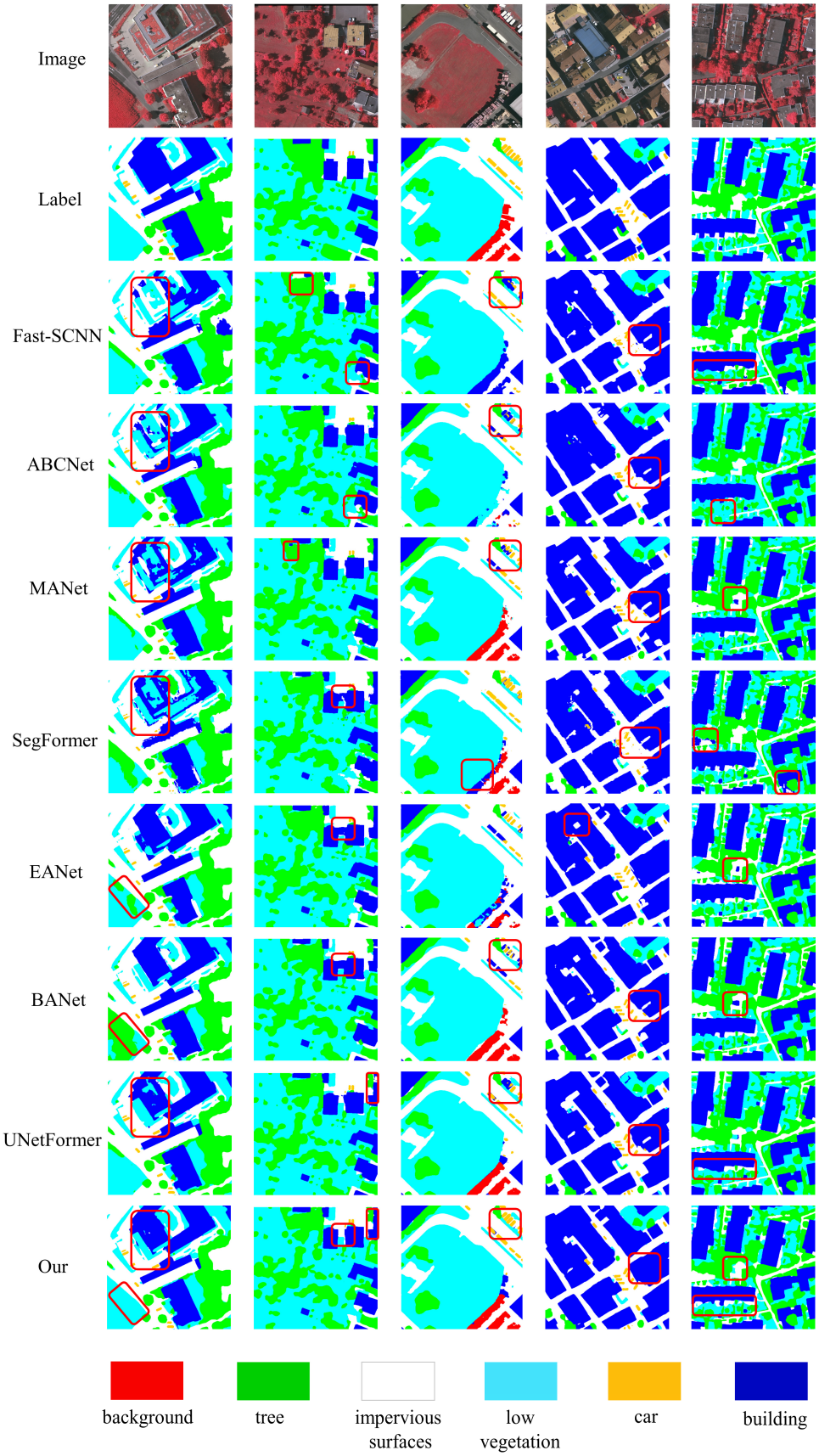


Figure 7. The test results on the Vaihingen dataset.

3.2.2. Results on the Potsdam Dataset

The Potsdam dataset is also one of the most widely used image segmentation datasets in remote sensing. As presented in Table 2, the Flauformer model outperformed all other models on this dataset, with an F1 score and OA of 93.1% and 91.5%, respectively. Figure 8 visualizes the test results of different methods. As shown in the upper part, the segmentation results of our model were purer, and the building edges were smoother compared to other methods. The VPANet and UNetFormer misclassified the buildings in the red box into background categories, and the Fast-SCNN's segmentation results had obvious errors. Although the DC-Swin and SegFormer also obtained good results in this region, they were inefficient and not suitable for real-time segmentation tasks. For the lower half, the inner triangle region mainly consists of low vegetation and background categories. Due to a lack of understanding of the global spatial information, the bare grass was misclassified into the impervious surfaces category by VPANet, DC-Swin, etc. Nevertheless, our method obtained better segmentation results in this region due to taking into account both global and local spatial contexts.

Table 2. Quantitative evaluation results on the Potsdam test set. The highest values are marked in bold.

Models	Backbone	Lowveg.	Tree	Car	Building	Impsurf.	MeanF1	OA
Fast-SCNN	–	85.3	86.9	85.2	94.5	90.8	87.6	88.5
ABCNet	ResNet18	87.5	**89.3**	95.6	96.7	93.6	92.6	91.2
DANet	–	85.6	87.8	84.5	95.3	91.0	88.9	89.1
MANet	ResNet50	87.7	88.1	95.5	96.3	92.9	91.6	90.6
EANet	ResNet18	84.5	85.6	95.2	95.5	92.1	90.5	88.7
DC-Swin	Swin-Tiny	88.1	88.4	96.2	**97.3**	93.5	92.9	91.4
BANet	ResT-Lite	87.5	88.9	96.1	96.8	93.1	92.3	91.0
SwinUNet	Swin-Tiny	87.8	88.3	95.9	96.5	93.2	92.4	91.3
VPANet	ResNet50	86.9	86.4	95.3	96.2	91.5	90.8	90.1
SegFormer	MiT-B1	87.3	87.9	95.8	96.3	92.4	91.1	90.7
UNetFormer	ResNet18	87.6	88.7	**96.8**	96.9	93.4	92.6	91.2
Our model	ResNet18	**88.3**	**88.9**	96.5	97.1	**93.8**	**93.1**	**91.5**

3.2.3. Results on the UAVid Dataset

UAVid is an urban street view image dataset with a high resolution and large size, which is mainly collected by drones in different urban areas. As a result, complex lighting conditions and multiple angles pose a challenge to accurately segment it. As shown in Table 3, our model obtained the highest mIoU on this dataset while maintaining good segmentation accuracy for most subclasses.

Table 3. Quantitative evaluation results on the UAVid test set. The highest values are marked in bold.

Models	Backbone	Building	Tree	Veg.	MovC.	StaC.	Road	Clutter	Human	mIoU
Fast-SCNN	–	84.9	79.3	61.1	60.5	48.3	77.2	65.6	10.6	61.4
ABCNet	ResNet18	86.5	79.7	63.3	69.5	48.3	**81.8**	67.2	14.1	63.6
DANet	–	85.6	78.6	61.7	59.8	47.5	77.4	64.7	9.4	60.8
MANet	ResNet50	85.1	77.6	60.2	67.1	53.4	77.8	64.9	14.9	62.5
A2FPN	ResNet18	86.9	79.3	64.1	70.2	53.7	81.3	67.5	21.4	65.3
EANet	ResNet18	87.1	79.5	63.4	69.4	53.2	80.8	67.2	21.6	64.9
DC-Swin	Swin-Tiny	87.3	80.6	63.4	73.1	56.6	80.9	66.5	30.8	67.3
BANet	ResT-Lite	85.3	78.7	62.4	69.5	52.9	80.5	66.9	21.2	64.5
SwinUNet	Swin-Tiny	85.2	79.5	61.5	70.6	51.8	79.8	65.7	22.6	64.3
VPANet	ResNet50	86.3	78.8	63.4	70.8	54.1	80.3	66.9	20.5	64.7
SegFormer	MiT-B1	86.4	79.7	62.5	72.6	52.7	80.2	66.5	28.3	66.1
UNetFormer	ResNet18	87.5	80.4	63.6	73.4	56.5	**81.3**	68.3	**31.3**	67.6
Our model	ResNet18	**87.6**	**81.4**	**64.2**	**75.3**	**56.8**	81.1	**69.1**	31.1	**68.2**

Specifically, the mIoU was 4.6% and 3.7% higher than the CNN-based network ABCNet and Transformer-based network BANet, respectively. In addition, the accuracy of our method in the human class was close to the UNetFormer based on CNN encoding and Transformer decoding, while achieving the highest scores on the two important categories (moving car and static car). As shown in the red box in Figure 9, the visualization results on this challenging dataset can further prove the superiority of our model. Generally speaking, the extraction results of our method were purer and had better performance on most subclasses.

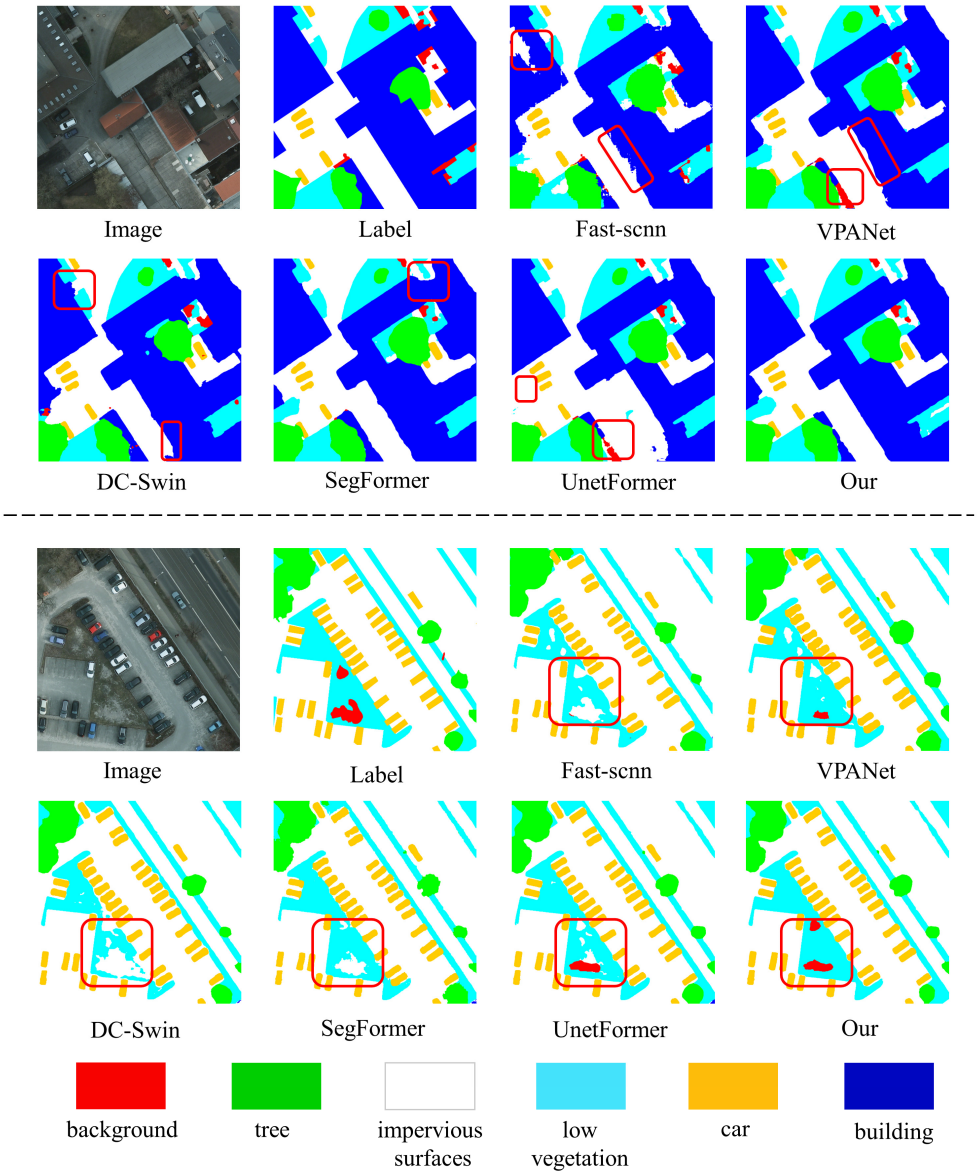


Figure 8. The test results on the Potsdam dataset.

3.2.4. Results on the LoveDA Dataset

The visualization results and quantitative analysis on the LoveDA Dataset can be seen in Figure A2 and Table A1 (in Appendix A), the areas that need to be focused on are presented in the purple box.

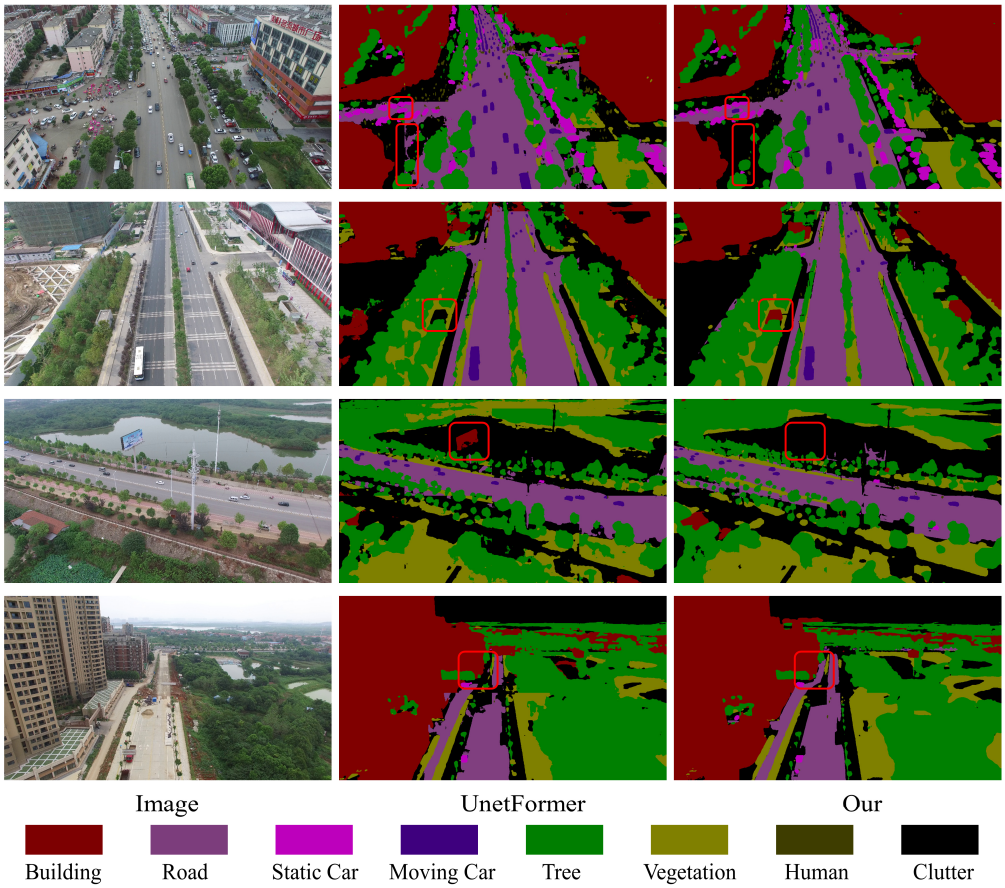


Figure 9. The segmentation results on the UAVid test set.

3.3. Ablation Experiments

3.3.1. Each Component of the Model

In order to evaluate the effect of each module of the proposed model on the final segmentation accuracy, a series of ablation experiments were conducted on the Vaihingen and UAVid datasets. The Baseline is a UNet network using ResNet18 as the backbone, which only considers the local context. In addition, we removed the SPT module and summed the global information in the global branch and the local information in the local branch (Baseline + GLIM-PT) to further evaluate the role of the position transformation operation. As shown in Table 4, the mIoU on the Vaihingen and UAVid datasets increased by 4.2% and 2.7%, respectively, after using the GLIM, and the position transformation operation can further improve the scores. In addition, the use of the FGH improved the mIoU by 0.6% and 1.3%, respectively. The results of the ablation experiments showed that each module of the proposed model is necessary and has positive effects on improving the final segmentation accuracy.

Table 4. Ablation experiments for each module in the model.

Dataset	Strategy	mIoU
Vaihingen	Baseline	75.4
	Baseline + GLIM-PT	79.6
	Baseline + GLIM	81.7
	Baseline + GLIM+ FGH	82.3
UAVid	Baseline	63.1
	Baseline + GLIM-PT	65.8
	Baseline + GLIM	66.9
	Baseline + GLIM+ FGH	68.2

3.3.2. Different Attention Mechanisms

To further demonstrate the effectiveness of the proposed global–local information attention (GLIA) mechanism, we replaced our attention module with other attention mechanisms. As presented in Table 5, the GLIA obtained the highest score (68.2%) on the UAVid test set compared to using other attention mechanisms. In addition, our attention mechanism performed better on the other three metrics. Specifically, our attention mechanism improved the speed by 49 and 46, respectively, when compared to the efficient MSA and SWA in the Transformer, which also proved the superiority of the proposed GLIA.

Table 5. Ablation experiments for different attention mechanisms. The input size was 1024 × 1024 and we used a single 3090 GPU. LNA: linear attention; MSA: multi-head self-attention; SWA: shifted window attention.

Attention Mechanism	Speed	Parameters (M)	Complexity (G)	mIoU
LNA	92	12.5	67.7	66.9
MSA	63	12.7	67.5	67.3
SWA	69	13.2	72.9	67.6
GLIA	114	11.7	47.1	68.2

3.3.3. Different Input Sizes

We repeated the experiments on the UAVid test set using different input sizes (square and rectangular shapes) to evaluate the model's stability. As shown in Table 6, the scores of the network did not vary by more than 0.8% when the input size changed, which proves the robustness of our method. Meanwhile, the highest scores were obtained when the input image size was 1024 × 1024. In addition, the square input performed better than rectangular inputs, and the network accuracy decreased when the input size was too large (2048 × 2048).

Table 6. Ablation experiments for different input sizes.

Input Size	Building	MovC.	StaC.	Human	mIoU
512 × 512	87.3	75.0	56.5	30.4	67.9
512 × 1024	86.9	74.8	56.3	30.6	67.4
1024 × 1024	87.6	75.3	56.8	31.1	68.2
2048 × 2048	87.1	75.2	55.9	29.8	67.7

3.3.4. Different Encoders

In many related studies, most used Transformer-based encoders, which are effective for accuracy improvement but introduce a large computational overhead. In this study, in order to verify the effect of different encoders on the final segmentation accuracy, we replaced the ResNet18 encoder with different lightweight Transformer encoders (ViT-Tiny [50], Swin-Tiny [22], and CoaT-Mini [51]) to conduct ablation experiments on the UAVid dataset. Specifically, the use of a lightweight Transformer encoder only achieved a limited accuracy improvement, but it introduced a large computational cost that limited the model's real-time scene application potential, demonstrating that using the ResNet18 as the encoder is a reasonable choice. The details can be seen in Table 7.

Table 7. Ablation experiments for different encoders on the UAVid dataset. The input size was 1024 × 1024 and we used a single 3090 GPU.

Encoder	Speed	Parameters (M)	Complexity (G)	mIoU
ViT-Tiny	32	8.5	35.4	67.5
CoaT-Mini	14	10.5	159.8	68.7
Swin-Tiny	31	28.1	104.3	68.9
ResNet18	114	11.7	47.1	68.2

4. Discussion

4.1. Visualization

As shown in Figure 10, we compared the experimental results of different attention mechanisms on the UAVid dataset to further clarify their specific roles. We found that LNA had the worst results and FLA obtained the best overall performance, proving that our method can achieve equivalent or better results than the Softmax attention mechanism. In addition, we visualized the attention effect of the segmentation head and found that the attention was equally distributed in the important regions of the image, which ensures that the model can accurately segment different classes. The visualization results further showed that the segmentation effects of our method outperformed traditional LNA, and it can achieve equivalent results to the Softmax attention mechanism while only introducing a small computational complexity.

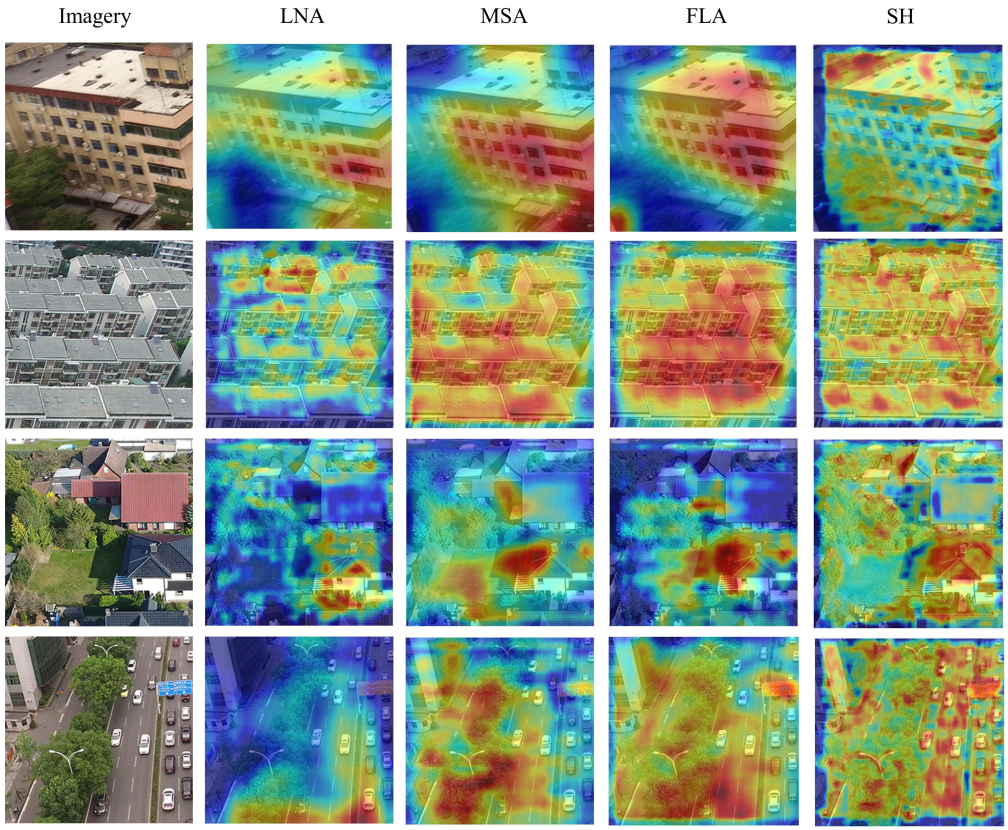


Figure 10. The heatmap visualization of different attention mechanisms. LNA: linear attention; MSA: multi-head self-attention; FLA: focused linear attention; SH: segmentation heads of our model.

4.2. Model Efficiency

Complexity and efficiency are extremely important to enhance the model's potential for real-time scene applications. Therefore, we compared the parameters, complexity, and speed metrics of each model on the UAVid dataset (1024 × 1024). As shown in Table 8, the speed (114) of our model was second only to the UNetFormer and Fast-SCNN, and it also achieved the highest accuracy performance (mIoU = 68.2). Specifically, the accuracy of our model was improved by 6.8% compared with the fastest shallow CNN network, Fast-SCNN, and it also outperformed other Transformer-based networks. In addition, we replaced the shifted window attention mechanism with the global–local information attention module to maintain the information interaction of different windows, which can further improve the model's efficiency (Table 5). The efficient balance between model efficiency and accuracy proves the potential of our method for real-time scene applications,

such as the ability to segment the remote sensing images in real time for environmental monitoring and land law enforcement inspections in urban areas.

Table 8. The speed comparison of each model on the UAVid dataset. The input size was 1024 × 1024 and we used a single 3090 GPU.

Models	Backbone	Complexity (G)	Parameters (M)	Speed	mIoU
Fast-SCNN	–	3.55	1.17	260	61.4
ABCNet	ResNet18	62.51	13.39	108	63.6
DANet	–	110.85	47.44	67	60.8
MANet	ResNet50	310.97	35.86	27	62.5
A2FPN	ResNet18	167.33	22.82	42	65.3
EANet	ResNet18	388.19	81.52	30	64.9
DC-Swin	Swin-Tiny	184.28	45.61	27	67.3
BANet	ResT-Lite	52.71	12.72	13	64.5
SwinUNet	Swin-Tiny	182.25	41.34	19	64.3
VPANet	ResNet50	143.21	24.15	56	64.7
SegFormer	MiT-B1	63.24	13.71	32	66.1
UNetFormer	ResNet18	46.98	11.68	121	67.6
Our model	ResNet18	47.10	11.69	114	68.2

5. Conclusions

In this study, a CNN-based encoding and Transformer-based decoding U-shaped network (Flauformer) was proposed for efficiently segmenting remote sensing urban images. Meanwhile, we designed the WMFSA module based on the focused linear attention and window-based, multi-head self-attention mechanisms, which constitutes a major part of the global–local information modeling module. The local branch and global branch in the GLIM ensured that our model could capture both local and global contexts, which is conducive to obtaining more accurate segmentation results. Specifically, the FLA and SPT modules were integrated into the global branch to extract global information efficiently, while the local branch used two sets of parallel convolutions to extract local information. As a result, the dual-branch structure in the decoder is the main reason that the proposed Flauformer achieved good performance. In addition, we conducted comparison and ablation experiments on four challenging datasets, demonstrating the potential of our model in real-time scene applications. Meanwhile, we also used the Transformer as the encoder. Although the model accuracy was improved slightly, it incurred great computational overhead, which demonstrates the rationality of using ResNet18 as the encoder. However, the method was not accurate enough to recognize some small objects, and the segmentation accuracy among different categories was not balanced. In addition, some areas with different lighting conditions were easily misclassified, such as the grass and pavements covered by building shadows. In the future, we will try to develop more efficient and effective models to address these challenges and enhance the potential of Transformer-based methods in the semantic segmentation tasks of remote sensing urban imagery, making contributions to urban planning, environmental protection, and sustainable development.

Author Contributions: Conceptualization, M.Y. and H.X.; methodology, H.X.; validation, M.Y. and H.Y.; software, H.X.; investigation, F.Z.; formal analysis, H.X.; data curation, H.X.; resources, M.Y.; writing—original draft preparation, H.X.; visualization, H.X.; writing—review and editing, M.Y.; supervision, M.Y.; funding acquisition, M.Y.; project administration, M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 41801308, and the China National Key R&D Program during the 13th Five-year Plan Period, grant number 2019YFD1100800.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Potsdam and Vaihingen datasets are available at Benchmark on Semantic Labeling (isprs.org), accessed on 14 July 2023; the UAVid dataset is available at UAVid Semantic Segmentation Dataset, accessed on 16 July 2023; the LoveDA dataset is available at CodaLab—Competition (upsaclay.fr), accessed on 18 July 2023.

Acknowledgments: The authors thank the anonymous reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

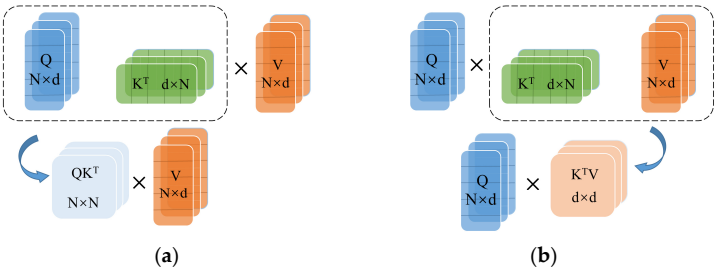


Figure A1. Illustration of the different attention mechanisms: (**a**) Softmax attention and (**b**) linear attention.

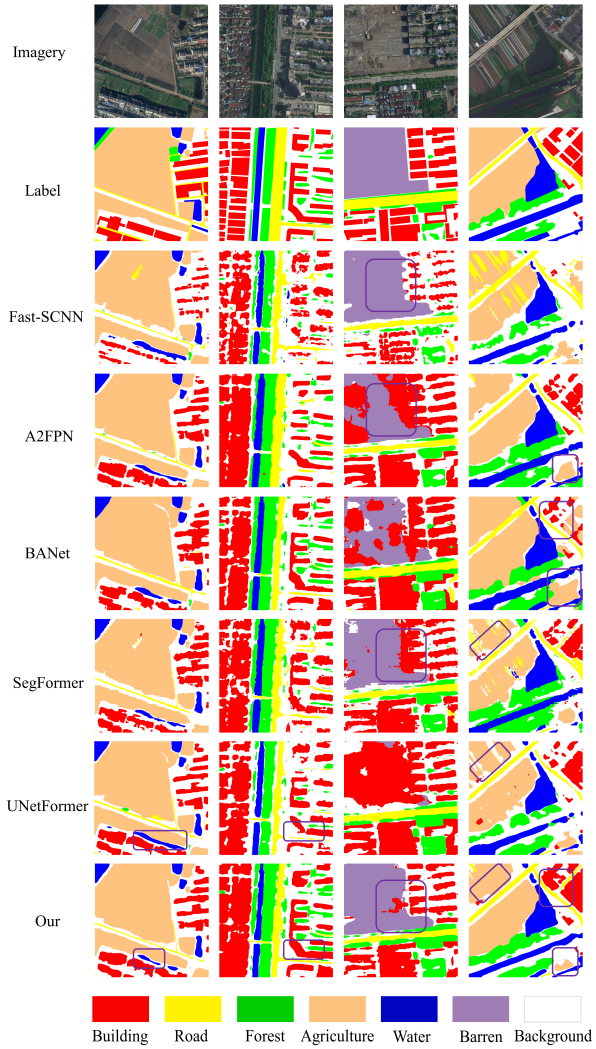


Figure A2. The experimental results of the LoveDA validation set.

Table A1. Quantitative evaluation results of the LoveDA test set. The highest values are marked in bold.

Models	Backbone	AgriC.	Building	Water	Barren	Road	Forest	Background	mIoU
Fast-SCNN	–	58.7	51.1	75.6	11.8	51.5	42.3	40.4	47.9
ABCNet	ResNet18	62.1	53.9	78.1	15.1	55.4	**46.9**	42.3	50.5
MANet	ResNet50	61.3	54.2	77.4	14.5	54.6	45.7	42.5	50.1
A2FPN	ResNet18	60.7	54.1	76.8	13.9	53.7	44.3	41.8	49.3
EANet	ResNet18	60.9	54.5	77.1	14.2	54.2	43.8	41.5	49.5
DC-Swin	Swin-Tiny	62.3	54.3	78.2	14.3	56.5	47.4	41.4	50.8
BANet	ResT-Lite	62.5	51.4	76.8	16.7	51.2	44.5	43.6	49.5
SwinUNet	Swin-Tiny	61.8	51.6	76.3	17.3	52.1	45.2	43.2	49.7
VPANet	ResNet50	60.2	53.5	75.6	14.3	54.1	43.7	40.5	48.6
SegFormer	MiT-B1	62.6	53.6	77.8	13.6	**56.7**	46.7	41.2	50.6
UNetFormer	ResNet18	62.2	58.7	79.5	20.3	54.6	46.1	**44.5**	52.3
Our model	ResNet18	**63.7**	**59.5**	**79.8**	**21.1**	55.4	46.5	43.6	**52.8**

References

1. Zhang, C.; Sargent, I.; Pan, X.; Li, H.; Gardiner, A.; Hare, J.; Atkinson, P.M. Joint Deep Learning for land cover and land use classification. *Remote Sens. Environ.* **2019**, *221*, 173–187. [CrossRef]
2. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Wang, L.; Atkinson, P.M. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 84–98. [CrossRef]
3. Zheng, F.; Lin, S.; Zhou, W.; Huang, H. A Lightweight Dual-branch Swin Transformer for Remote Sensing Scene Classification. *Remote Sens.* **2023**, *15*, 2865. [CrossRef]
4. Shen, J.; Yu, T.; Yang, H.; Wang, R.; Wang, Q. An Attention Cascade Global–Local Network for Remote Sensing Scene Classification. *Remote Sens.* **2022**, *14*, 2042. [CrossRef]
5. Cheng, G.; Yao, Y.; Li, S.; Li, K.; Xie, X.; Wang, J.; Yao, X.; Han, J. Dual-Aligned Oriented Detector. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
6. Shamsolmoali, P.; Zareapoor, M.; Zhou, H.; Wang, R.; Yang, J. Road segmentation for remote sensing images using adversarial spatial pyramid networks. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4673–4688. [CrossRef]
7. Griffiths, D.; Boehm, J. Improving public data for building segmentation from Convolutional Neural Networks (CNNs) for fused airborne lidar and image data using active contours. *ISPRS J. Photogramm. Remote Sens.* **2019**, *154*, 70–83. [CrossRef]
8. Li, R.; Zheng, S.; Duan, C.; Wang, L.; Zhang, C. Land cover classification from remote sensing images based on multi-scale fully convolutional network. *Geo-Spat. Inf. Sci.* **2022**, *25*, 278–294. [CrossRef]
9. Marcos, D.; Volpi, M.; Kellenberger, B.; Tuia, D. Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 96–107. [CrossRef]
10. Wu, C.; Du, B.; Zhang, L. Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection. *arXiv* **2022**, arXiv:2201.06030. [CrossRef]
11. Xu, X.; Yang, Z.; Li, J. AMCA: Attention-guided Multi-scale Context Aggregation Network for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5908619. [CrossRef]
12. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* **2005**, *26*, 217–222. [CrossRef]
13. Guo, Y.; Jia, X.; Paull, D. Effective Sequential Classifier Training for SVM-Based Multitemporal Remote Sensing Image Classification. *IEEE Trans. Image Process.* **2018**, *27*, 3036–3048. [CrossRef] [PubMed]
14. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
15. Yang, X.; Li, S.; Chen, Z.; Chanussot, J.; Jia, X.; Zhang, B.; Chen, P. An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 238–262. [CrossRef]
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2017; Volume 30.
17. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
18. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
19. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20. [CrossRef]
20. Xia, Z.; Pan, X.; Song, S.; Li, L.E.; Huang, G. Vision transformer with deformable attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4794–4803.

21. Wang, W.; Xie, E.; Li, X.; Fan, D.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.
22. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
23. Hassani, A.; Walton, S.; Li, J.; Li, S.; Shi, H. Neighborhood attention transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 6185–6194.
24. Katharopoulos, A.; Vyas, A.; Pappas, N.; Fleuret, F. Transformers are rnns: Fast autoregressive transformers with linear attention. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 12–18 July 2020; pp. 5156–5165.
25. Han, D.; Pan, X.; Han, Y.; Song, S.; Huang, G. FLatten Transformer: Vision Transformer using Focused Linear Attention. *arXiv* **2023**, arXiv:2308.00442.
26. Cai, H.; Gan, C.; Han, S. Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. *arXiv* **2022**, arXiv:2205.14756.
27. You, H.; Xiong, Y.; Dai, X.; Wu, B.; Zhang, P.; Fan, H.; Vajda, P.; Lin, Y. Castling-vit: Compressing self-attention via switching towards linear-angular attention during vision transformer inference. *arXiv* **2022**, arXiv:2211.10526.
28. Bolya, D.; Fu, C.Y.; Dai, X.; Zhang, P.; Hoffman, J. Hydra attention: Efficient attention with many heads. In Proceedings of the European Conference on Computer Vision, Aviv, Israel, 23–27 October 2022; pp. 35–49.
29. Xiong, Y.; Zeng, Z.; Chakraborty, R.; Tan, M.; Fung, G.; Li, Y.; Singh, V. Nystromformer: A nystrom-based algorithm for approximating self-attention. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; pp. 14138–14148.
30. Lu, J.; Yao, J.; Zhang, J.; Zhu, X.; Xu, H.; Gao, W.; Xu, C.; Xiang, T.; Zhang, L. Soft: Softmax-free transformer with linear complexity. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 21297–21309.
31. Choromanski, K.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J.; Mohiuddin, A.; Kaiser, L.; et al. Rethinking attention with performers. *arXiv* **2020**, arXiv:2009.14794.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
33. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
34. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
35. Yu, T.; Khalitov, R.; Cheng, L.; Yang, Z. Paramixer: Parameterizing mixing links in sparse factors works better than dot-product self-attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 691–700.
36. Ren, H.; Dai, H.; Dai, Z.; Yang, M.; Leskovec, J.; Schuurmans, D.; Dai, B. Combiner: Full attention transformer with sparse computation cost. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 22470–22482.
37. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. *arXiv* **2021**, arXiv:2103.15808.
38. Poudel, R.P.; Liwicki, S.; Cipolla, R. Fast-scnn: Fast semantic segmentation network. *arXiv* **2019**, arXiv:1902.04502.
39. Zhu, Z.; Xu, M.; Bai, S.; Huang, T.; Bai, X. Asymmetric non-local neural networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 593–602.
40. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *arXiv* **2020**, arXiv:2004.02147. [CrossRef]
41. Lyu, Y.; Vosselman, G.; Xia, G.-S.; Yilmaz, A.; Yang, M.Y. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *165*, 108–119. [CrossRef]
42. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A Remote Sensing Land Cover Dataset for Domain Adaptive Semantic Segmentation. *arXiv* **2021**, arXiv:2110.08733.
43. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
44. Li, R.; Wang, L.; Zhang, C.; Duan, C.; Zheng, S. A2-FPN for semantic segmentation of fine-resolution remotely sensed images. *Int. J. Remote Sens.* **2022**, *43*, 1131–1155. [CrossRef]
45. Zheng, X.; Huan, L.; Xia, G.-S.; Gong, J. Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss. *ISPRS J. Photogramm. Remote Sens.* **2020**, *170*, 15–28. [CrossRef]
46. Wang, X.; Kang, M.; Chen, Y.; Jiang, W.; Wang, M.; Weise, T.; Tan, M.; Xu, L.; Li, X.; Zou, L.; et al. Adaptive Local Cross-Channel Vector Pooling Attention Module for Semantic Segmentation of Remote Sensing Imagery. *Remote Sens.* **2023**, *15*, 1980. [CrossRef]
47. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images. *Remote Sens.* **2021**, *13*, 3065. [CrossRef]
48. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv* **2021**, arXiv:2105.05537.

49. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
50. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
51. Xu, W.; Xu, Y.; Chang, T.; Tu, Z. Co-Scale Conv-Attentional Image Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 9981–9990.