Essay

BAFormer: A Novel Boundary-Aware Compensation UNet-like Transformer for High-Resolution Cropland Extraction

Zhiyong Li , Youming Wang , Fa Tian , Junbo Zhang , Yijie Chen , Kunhong Li *

Posted Date: 3 June 2024

doi: 10.20944/preprints202406.0053.v1

Keywords: high-resolution remote sensing image; boundary-aware; cropland; semantic segmentation

 doi:10.20944/preprints202406.0053.v1

Article

BAFormer: A Novel Boundary-Aware Compensation UNet-like Transformer for High-Resolution Cropland Extraction

Zhiyong Li [1,†], Youming Wang [2,†], Fa Tian [2,†], Junbo Zhang [2], Yijie Chen [2], Kunhong Li [1,*]

[1] College of Information Engineering, Sichuan Agricultural University, Ya'an Digital Agricultural Engineering Technology Research Center, Ya'an 625014, China; lzy@sicau.edu.cn

[2] College of Information Engineering, Sichuan Agricultural University, Ya'an 625014, China; wym@stu.sicau.edu.cn (Y.W.); tf@stu.sicau.edu.cn (F.T.); zjb@stu.sicau.edu.cn (J.Z.); chenyijie0813@stu.sicau.edu.cn (Y.C.)

* Correspondence: lkh@sicau.edu.cn (K.L.)

† These authors contributed equally to this work.

Abstract: Utilizing deep learning for semantic segmentation of cropland from remote sensing imagery has become a crucial technique in land surveys. Cropland illustrates diverse morphologies and degrees of fragmentation on the Earth's surface, underscoring the importance of accurately perceiving the complex boundaries of cropland which are crucial for effective segmentation. This paper introduces a UNet-like boundary-aware compensation model BAFormer. Cropland boundaries typically exhibit rapid transformations in pixel values and texture features, often appearing as high-frequency features in remote-sensing images. To enhance the recognition of these high-frequency features as represented by cropland boundaries, the proposed BAFormer integrates a Feature Adaptive Mixer (FAM) and develops a Deep Wide Large Kernel Multi-Layer Perceptron (DWLK-MLP) to enrich the global and local cropland boundaries features separately. Specifically, FAM adaptively mixes high-frequency and low-frequency features through the advantages of convolution and self-attention; DWLK-MLP expands the convolutional receptive field by deeply decomposing large kernel convolutions. The efficacy of BAFormer has been evaluated on the Vaihingen, Potsdam, and LoveDA public datasets, as well as the Mapcup dataset. It has demonstrated advanced performance, achieving mIoU scores of 84.5%, 87.3%, 53.5%, and 83.1% on these datasets respectively. Notably, BAFormer-T, the lightweight iteration of the model, surpasses other lightweight models on the Vaihingen dataset with scores of 91.3% F1 and 84.1% mIoU. The source code is available at https://github.com/WangYouM1999/BAFormer.

Keywords: high-resolution remote sensing image; cropland; boundary-aware

1. Introduction

With the rapid development of remote sensing technology, finer and higher resolution optical remote sensing images are now available [1]. Extracting cropland information from these optical remote-sensing images is crucial for assessing food security and formulating agricultural policies [2]. Currently, the mainstream solution is to use deep learning models, which are effective in identifying plot boundaries and feature types [3]. Although existing deep learning methods have achieved some success in processing cropland data, the boundary segmentation problem still suffers from some inaccuracies due to the highly heterogeneous and fragmented characteristics of cropland [4]. Specifically, this inaccuracy is usually caused by factors such as complex boundary morphology and feature recognition errors, as illustrated in Figure 1. In practical applications, this inaccuracy will be manifested in the phenomenon of boundary error and omission, which will have certain impacts on fields such as land use planning and agricultural production [5]. To alleviate the problem of inaccurate cropland boundary segmentation, it is urgent to enhance the model's perception ability of edge features to improve the recognition accuracy and reliability of segmentation results [6].

 doi:10.20944/preprints202406.0053.v1

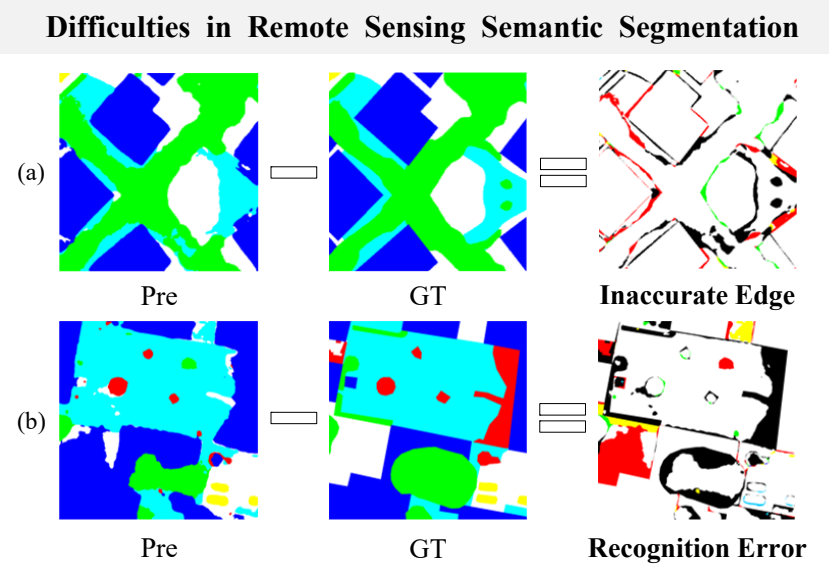


Figure 1. Both (**a**) and (**b**) are obtained by the difference between model Predict(Pre) and Ground Truth(GT).

In recent years, many studies have proposed the combination of deep learning and edge detection to guide the model to better perceive the cropland information, improve the local segmentation score accuracy, and maintain the global morphology land continuity. Existing approaches [2,4,7–16] design-specific network structures based on the characteristics of Cropland to guide the model to focus on key features. However, most of them focus on specific geographic regions or a single cropland type, ignoring the regional differences of cropland parcels, and failing to achieve the purpose of generalized extraction. (2) Feature-based approaches [5,17–26] enhance the representation and understanding of cropland information by supplementing the model with additional features. However, some redundant feature representations do not provide positive images to the model while increasing the computational burden of the model. (3) Loss-based approaches [3,6,27–29] introduce additional metrics to supplement the constraints during the training process. These methods strengthen the constraints on segmentation and optimize boundary continuity. However, these works always categorize boundary and interior pixels into two different classes, which somewhat destroys the consistency of the same boundary pixels and the inter-class variability of different boundary pixels. These methods strengthen the restriction on segmentation and optimize the boundary continuity. However, these works always categorize boundary and interior pixels into two different classes, which somewhat destroys the consistency of the same boundary pixels and the inter-class variability of different boundary pixels. To better solve the problem of edge segmentation in cropland information extraction, it is necessary to comprehensively consider the advantages and disadvantages of different methods, and combine the characteristics of cropland data to explore a more effective way to combine deep learning and edge detection.

In this paper, to alleviate the problem of inaccurate boundary segmentation, we propose a UNet-like boundary-aware compensation model BAFormer. This method compensates for boundary-awareness in three aspects, namely, feature extraction, fusion, and constraints, from the perspective of optimising the edge quality, and effectively improving the quality of edge segmentation. (1) In terms of feature extraction, a feature adaptive mixer (FAM) based on channel mixing and a deep large kernel multilayer perceptron (DWLK-MLP) is proposed, which can effectively enhance the model information flow and expressive ability. FAM uses the advantages of convolution and self-attention to separate the high-frequency and low-frequency features of the image, and it can efficiently extract the details and global information in the image and learn more accurate boundary features. Different from other hybrid structures, we innovatively use the advantages of convolution and self-attention to decompose the high-frequency and low-frequency features of an image and achieve adaptive fusion

 doi:10.20944/preprints202406.0053.v1

by calculating the contribution of frequencies. DWLK-MLP increases the receptive field of convolution by deeply separating the large kernel convolution, and more complete boundary features can be extracted with almost no computational overhead. (2) Regarding feature fusion, the adaptive feature fusion (RAF) strategy based on the perception of spatial and channel semantic relations is proposed. Different from other static feature fusion methods, this method achieves dynamic fusion of multi-scale features by establishing spatial numerical dependencies and channel semantic dependencies between features and weighting the weights from different angles. (3) In terms of boundary constraints, an edge constraint strategy implemented in the deep layer of the network is proposed. The strategy works by back-propagating the error signals to guide the model to focus on and optimize the boundaries from the bottom layer through high-level semantics while supervising the Encoder side a priori and does not require additional auxiliary task overheads. With the above improvements and innovations, BAFormer can better mitigate the problem of inaccurate boundary segmentation in the segmentation task and improve the quality of edge segmentation.

In summary, our main contributions are summarised as follows:

1. We propose the BAFormer framework for edge optimization. The framework comprehensively compensates boundary awareness from multiple perspectives (including feature extraction, feature fusion, and loss constraints), which greatly improves the model's representational capability and edge segmentation quality.
2. A Feature Adaptive Mixer (FAM) is proposed to efficiently extract detailed and global features in an image by mixing high-frequency and low-frequency information, thus enhancing the information flow and representation ability of the model.
3. A Deep Large Kernel Multi-Layer Perceptron (DWLK-MLP) is proposed. The boundary features are enriched with negligible computational overhead by deeply decomposing the large kernel convolution.
4. A Relational Adaptive Fusion (RAF) strategy is proposed, which optimizes feature granularity by dynamically sensing the relationships between features from both spatial and channel semantic perspectives.
5. A Deeply Supervised Edge Constraint Strategy is proposed. The boundary continuity is strengthened by making the model automatically focus on the boundary through deep semantic guidance.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 June 2024 doi:10.20944/preprints202406.0053.v1

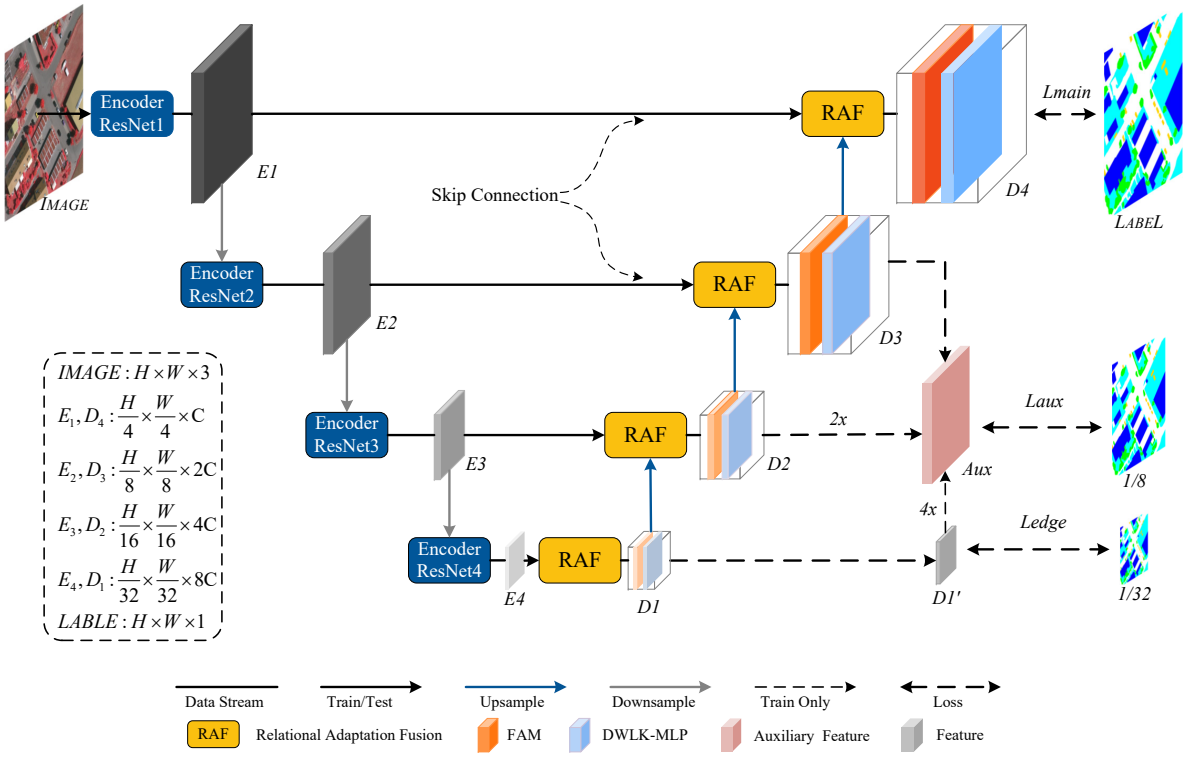


Figure 2. An overview of the BAFormer.

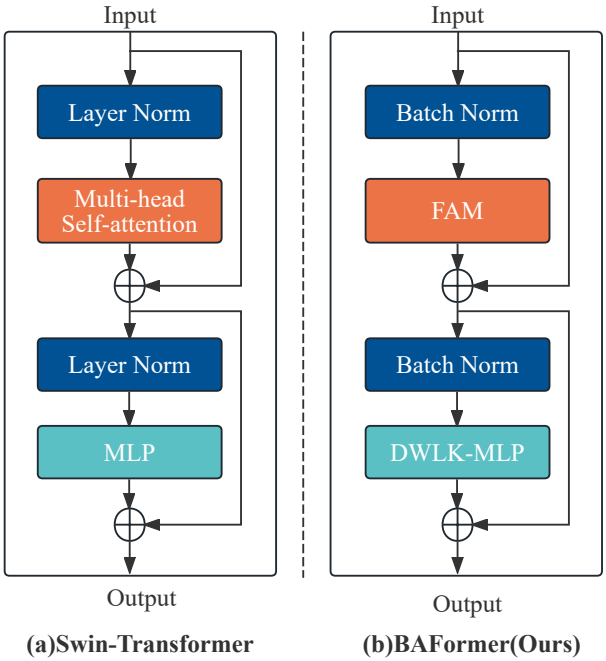


Figure 3. The structure of BABlock. (**a**) represents the Block structure in Swin-Transformer, and (**b**) represents the BABlock structure in BAFormer.

 doi:10.20944/preprints202406.0053.v1

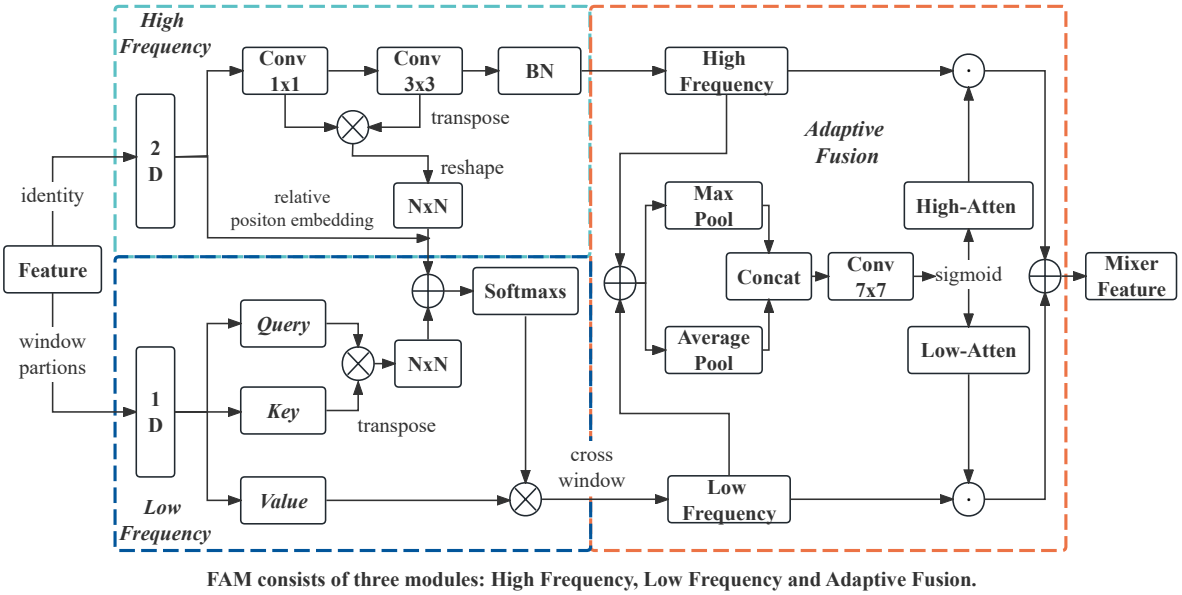


Figure 4. The structure of Feature Adaptive Mixer. 2D is a two-dimensional image, and 1D is a sequence stretched to one dimension. BN stands for Batch Normalization. Hign-Attn represents the weight attention score occupied by the high-frequency features in the mixed information flow, and Low-Attn represents the weight attention score occupied by the low-frequency features in the mixed information flow, and the dimension of both is HxW.

2. Related Work

2.1. Methods Based on Network Design

Various methods have been proposed to set different network structures and modules according to the cropland morphology to achieve better performance. (1) For the inherent finite geometric transformations of convolutional neural networks, methods based on convolutional kernel design have been proposed, represented by the well-known dilated convolution [30] and deformable convolution [7], which show excellent performance in complex monitoring and segmentation tasks. These [2,12,31] can dynamically sense the geometric features of objects to adapt to morphologically changing structures. For example, the MDANet proposed in [32] designs deformable attention module DAM combining sparse spatial sampling strategy and long-range relational modeling capability for capturing the domain structure information of each pixel to enable better adaptation to the structure of HRSI images. (2) CNN-Transformer based hybrid models [8–10,16,33,34], designed to adequately learn diverse target features are proposed. For example, the ASNet network proposed in [8] innovatively integrates Transformer and CNN techniques in a two-branch encoder to capture global dependencies while capturing local fine-grained image features. In [9], Swin-Transformer is embedded into a classical CNN-based UNet to form a novel dual encoder architecture with Swin-Transformer and CNN in parallel to enhance the feature representation of occluded targets, which brings significant performance improvement on the ISPRS-Vaihingen and Potsdam datasets. (3) Proposed a parcel extraction method based on the combination of edge detection and semantic segmentation, which has led to significant progress in image understanding. These [4,11–14] take the task of detecting hard or soft boundaries to guide the model to impose further attention constraints on the boundaries, enabling it to obtain better image decoding capabilities. For example, [14] designed frequency attention to topic emphasize key high-frequency components in the feature map to improve the accuracy of boundary detection. [32] proposed a multi-task joint network MDE-UNet for accurate segmentation by three-branch multi-task learning of deterministic, fuzzy, and primitive boundaries. Inspired by this, this paper designs the model in terms of network architecture as well as boundary guidance. Hybrid model architectures are designed to fully capture complete boundary information. By enhancing boundary awareness

 doi:10.20944/preprints202406.0053.v1

and edge guidance, the model can dynamically focus on the boundary information and automate the optimization during the network learning process. Unlike many assisted boundary guidance approaches that use three segmentation masks for post-processing, the proposed network is further enhanced in the feature extraction, transmission fusion, and constraint guidance processes at the boundaries, resulting in a more efficient and accurate boundary delineation workflow.

2.2. Methods Based on Features Fusion

Feature fusion-based approaches [5,17–26] enhance the representation of cropland information by supplementing additional feature information to the model. Considering the difficulty in labeling the existing high-resolution remote sensing image samples, [17] utilized the existing medium-resolution remote sensing images as a priori knowledge to provide cross-scale relocatable samples for HR images, thus obtaining more effective high-resolution farmland samples. To mitigate feature details due to image downsampling and noise from the same image reduces the network's ability to discriminate between useful and useless information, [19] proposed to compensate for local image features and minimize noise by bootstrapping the feature extraction module to emphasize the learning of useful information. [21] proposed a fully convolutional neural network HRNet-CRF with improved contextual feature representation to optimize the initial semantic segmentation results by morphological post-processing methods to obtain internally homogeneous farmland. [22] proposed a boundary-enhanced segmentation network, HBRNet, with Swing-Transformer as the backbone of the pyramid hierarchy to obtain contextual information while enhancing boundary details. Aiming at the different texture features of plots, [24] proposed a pyramid scene parsing network-statistical texture learning deep learning framework that combines high-level semantic feature extraction with low-level texture feature deep mining to achieve more accurate farmland recognition. [5] proposed to encode parcel features by transformer module and null convolution module, which operates on multi-scale features at the feature extraction order, which in turn improves the ability to capture the details and boundaries of farmland parcels. Different from the above methods, this paper proposes an adaptive feature fusion strategy based on channel semantics and spatial relationship awareness to eliminate the semantic gap between shallow and deep feature fusion from the perspective of spatial values and channel semantics, to obtain finer fine-grained features.

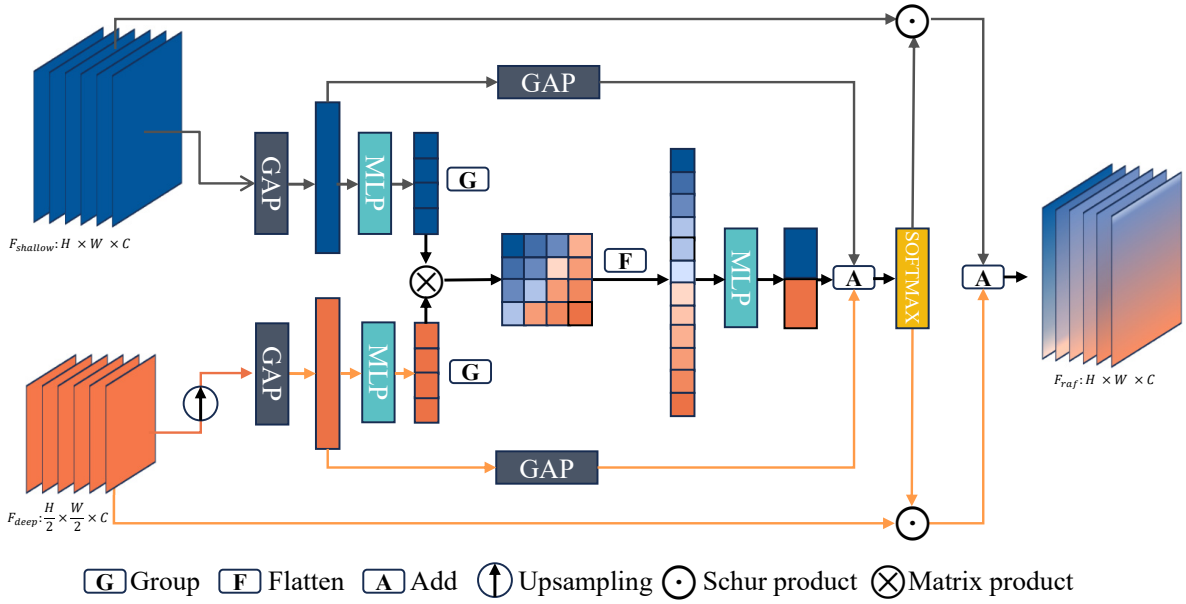


Figure 5. Illustration of the Relational Adaptive Fusion module. GAP stands for global average pooling and MLP stands for multi-layer perceptron variation. Blue and orange represent the feature maps of the shallow and deep layers of the network, respectively.

 doi:10.20944/preprints202406.0053.v1

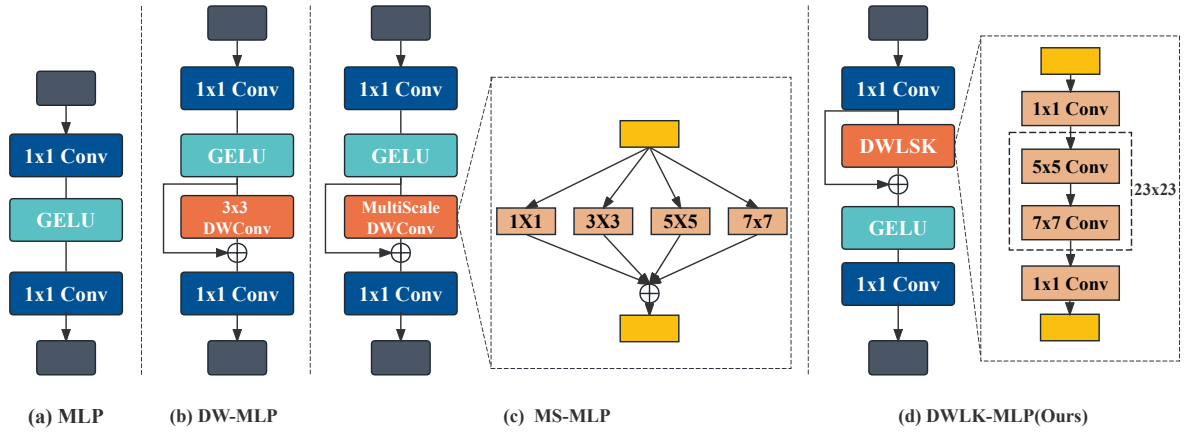


Figure 6. (**a**) Plain MLP that only processes cross-channel information (**b**) Deep residuals for aggregating local tokens DW-MLP (**c**) Deep residuals for aggregating multi-scale tokens MS-MLP (**d**) Our proposed deep large kernel DWLK-MLP.

2.3. Methods Based on Loss Function

Loss function-based methods [3,6,27–29,35], introduce a metric approach to complement strong constraints on morphological boundaries in the training process summarization. SEANet [6] proposes a multi-task loss that constrains irregular agricultural parcels from the mask prediction, edge prediction, and distance map estimation tasks to improve the geometric accuracy of the parcels circling. RBP -MTL [27] jointly models local spatial constraints between each region, boundary, and object through multi-task learning to promote object separability and boundary connectivity for agricultural parcels. [29] proposed a boundary loss in the form of a distance metric on contour space instead of regions, showing that boundary loss can yield significant performance gains while improving training stability. ABL [28] proposed a new active boundary loss algorithm for semantic segmentation that models the boundary alignment problem into a microdirectionally vectorizable prediction problem by incrementally encouraging the alignment of predicted boundaries with the true boundaries problem to improve boundary details. [35] proposed a new conditional edge loss CBL for improving boundary segmentation, specifically by pulling each boundary pixel closer to its unique local class center and pushing it away from its unlike neighbors to enhance pixel intra-class consistency and inter-class variability, which in turn filters out noisy and incorrect information to obtain accurate boundaries. However, these works always classify boundary pixels and internal pixels into two different classes when optimizing the pixel-level boundary classification assistance task, which destroys the consistency of the same class and the inter-class variability of different boundary pixels. In this paper, we propose a deep constraint strategy based on complete boundary perception, which does not need to introduce additional auxiliary task overhead, but only utilizes the rich high-level semantic information in the deep layers of the network to guide the model to self-help focus on the boundary information and strengthen the boundary constraints.

3. Proposed Method

This section will introduce the proposed BAFormer architecture and discuss and analyze its key designs. These key designs include the Feature Adaptive Mixer (FAM), the Deep Wide Large Kernel Multi-Layer Perceptron (DWLK-MLP), the Relationship Adaptive Fusion (RAF) strategy, and the Depth Supervised Edge Constraint Strategy.

3.1. CNN-Based Encoder

In the BAFormer model (as shown in Figure 2), we adopt ResNet-18 as the shallow semantic extractor for the Encoder, to significantly reduce computational costs. This choice has been proven to be efficient and effective in a wide range of semantic segmentation tasks. Specifically, ResNet-18 consists of four residual blocks that can extract shallow (high-frequency) semantic features. In the

 doi:10.20944/preprints202406.0053.v1

process of feature compression, each block will undergo a spatial resolution reduction by half through downsampling, and the number of channels will double to enable deeper semantic information extraction. During each block's process, we employ Skip Connections with identity mapping to connect the shallow semantic features with the corresponding semantic levels in the Decoder. This connection method can adaptively merge the semantic differences between shallow and deep layers, thereby improving the model's performance.

3.2. Transformer-Based Decoder

In the BAFormer model, the Decoder side achieves the abstract feature extraction and reconstruction of the image by stacking four BABlock modules from the bottom upwards, the structure of which is shown in Figure3. To ensure high-quality image recovery, shallow information is dynamically fused by the RAF module before each BABlock.

3.2.1. FAM (Feature Adaptive Mixer)

Convolutional Neural Network (CNN) acts as a high-pass filter that can extract locally salient high-frequency information such as texture and detail [36]. Self-Attention mechanism is a relatively low-pass filter that can extract salient low-frequency information such as global and smooth [37]. Although the traditional pure convolution-based methods can effectively extract rich high-frequency features, they are unable to capture the spatial contextual information of the image. In contrast, methods based on purely self-attentive mechanisms tend to extract only the low-frequency information of the image, and also suffer from computational complexity and poor model generalization. Therefore, how to give full play to the advantages of these two computational paradigms has become a bottleneck for further breakthroughs in model feature extraction capability. From the ideas of information distillation and frequency mixing in image super-resolution reconstruction, we can get some insights. By mixing low-frequency features and high-frequency features, the model's information flow and expression ability can be effectively enhanced [38,39].

 doi:10.20944/preprints202406.0053.v1

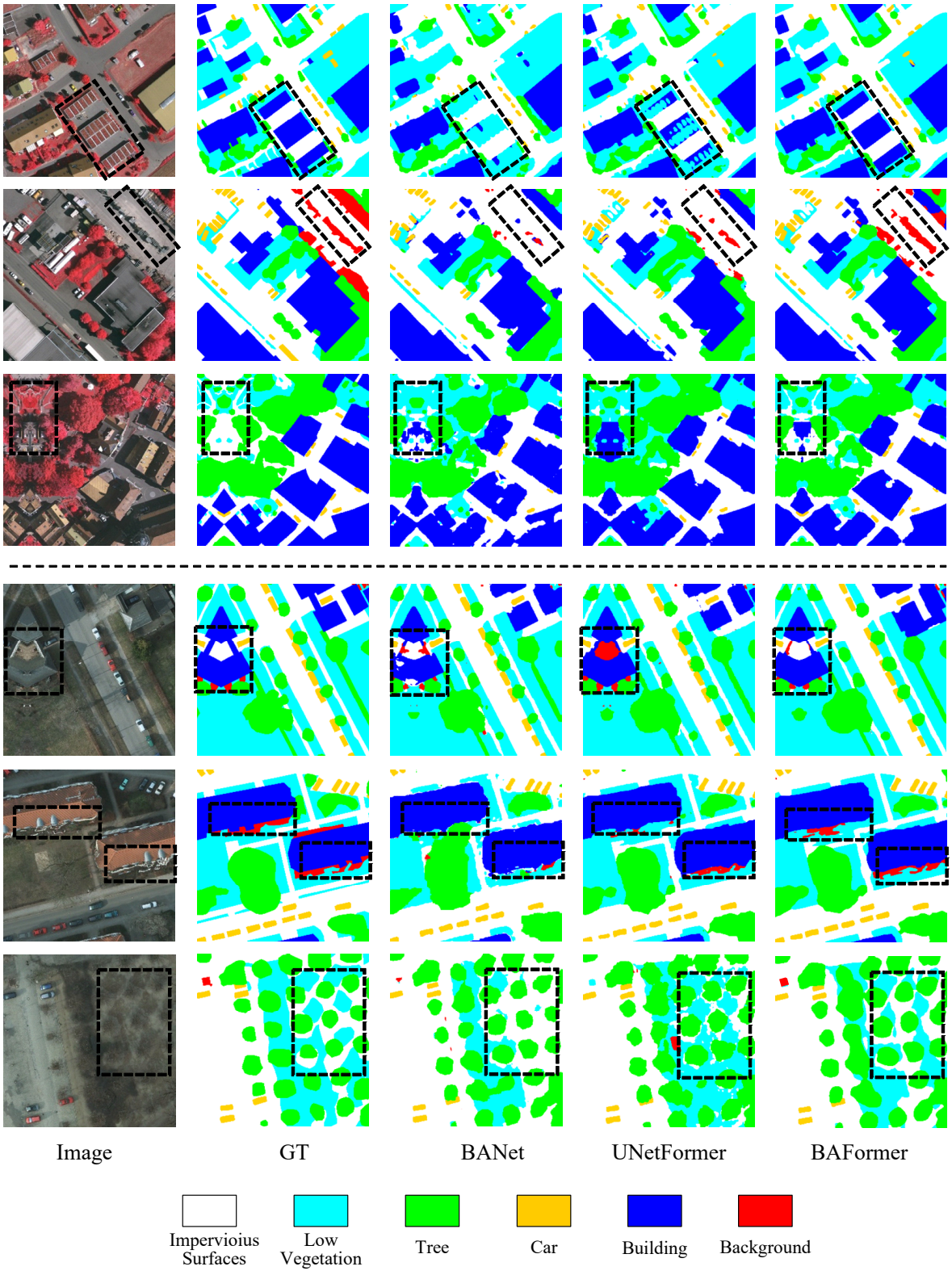


Figure 7. Qualitative comparison under ISPRS Vaihaigen(top), ISPRS Postdam(bottom) test sets. We add some black boxes to highlight the differences to facilitate model comparison.

 doi:10.20944/preprints202406.0053.v1

Table 1. Quantitative comparisons with existing methods were performed on the Vaihingen dataset. The best values in this column are highlighted in bold.

Method	Backbone	Imp.surf	Building	Low.veg	Tree	Car	MeanF1(%)	OA(%)	mIoU(%)
FCN [40]	VGG-16	88.2	93.0	81.5	83.6	75.4	84.3	87.0	74.2
DeepLabv3+ [41]	ResNet-50	88.0	94.2	81.3	87.8	78.1	85.9	88.9	76.3
MAREsU-Net [42]	ResNet-18	92.0	95.0	83.7	89.3	78.3	87.7	89.7	78.7
ABCNet [43]	ResNet-18	92.7	95.2	84.5	89.7	85.3	89.5	90.7	81.3
BANet [44]	ResT-Lite	92.2	95.2	83.8	89.9	86.8	89.6	90.5	81.4
UNetFormer [45]	ResNet-18	92.7	95.3	84.9	90.6	88.5	90.4	91.0	82.7
MANet [46]	ResNet-50	93.0	95.5	84.6	90.0	89.0	90.4	91.0	82.7
Mask2Former [47]	Swin-B	92.9	94.5	85.3	90.4	88.5	90.3	90.8	83.0
DC-Swin [48]	Swin-S	93.6	**96.2**	**85.8**	90.4	87.6	90.7	91.6	83.2
FT-UNetFormer [45]	Swin-B	93.5	96.0	85.6	90.8	90.4	91.3	91.6	84.1
BAFormer-T	ResNet-18	**93.7**	95.7	85.4	90.2	91.0	91.2	91.6	84.2
BAFormer	ResNet-18	**93.7**	96.0	85.7	**90.9**	**91.2**	**91.5**	**91.8**	**84.5**

Table 2. The Potsdam dataset was quantitatively compared with existing methods. The best values in this column are shown in bold.

Method	Backbone	Imp.surf	Building	Low.veg	Tree	Car	MeanF1(%)	OA(%)	mIoU(%)
FCN [40]	VGG-16	88.5	89.9	78.3	85.4	88.8	86.2	86.6	78.5
DeepLabv3+ [41]	ResNet-50	90.4	90.7	80.2	86.8	90.4	87.7	87.9	80.6
MAREsU-Net [42]	ResNet-18	91.4	85.6	85.8	86.6	93.3	88.5	89.0	83.9
BANet [44]	ResT-Lite	93.3	95.7	87.4	89.1	96.0	92.3	91.0	85.3
ABCNet [43]	ResNet-18	93.5	95.9	87.9	89.1	95.8	92.4	91.3	85.5
SwinTF-FPN [49]	Swin-S	93.3	96.8	87.8	88.8	95.0	92.3	91.1	85.9
UNetFormer [45]	ResNet-18	93.6	96.8	87.7	88.9	95.8	92.6	91.3	86.0
MANet [46]	ResNet-50	93.4	96.7	88.3	89.3	96.5	92.8	91.3	86.4
Mask2Former [47]	Swin-B	**98.0**	96.9	88.4	**90.7**	84.6	91.7	**92.5**	86.6
FT-UNetFormer [45]	Swin-B	93.5	97.2	88.4	89.6	96.6	**93.2**	91.6	87.0
BAFormer-T	ResNet-18	93.5	96.8	88.2	89.2	96.4	92.8	91.3	86.4
BAFormer	ResNet-18	93.7	**97.3**	**88.5**	89.7	**96.8**	**93.2**	92.2	**87.3**

To enhance the accuracy of boundary identification, we propose a module called FAM. This method captures more accurate boundary features by enhancing the information flow and expressiveness of the model. It not only solves the single-scale feature problem, but also incorporates the idea of multi-branch structure to filter out important features from rich semantic information. Specifically, FAM includes three main parts: high-frequency branching, low-frequency branching, and adaptive fusion, as shown in Figure 4. It aims to separate high-frequency features and low-frequency features in an image to capture local and global information of the image through the respective advantages of convolutional neural network and self-attention, and adaptively selects the fusion according to the contribution of channel fusion. Unlike traditional hybrid methods, we innovatively combine the high-frequency static affinity matrix extracted by convolution with the dynamic low-frequency affinity matrix obtained based on self-attention, which enhances self-attention's ability to comprehensively capture high-frequency and low-frequency information and feature generalization. In addition, for the characteristics of these two computational paradigms, we carry out adaptive feature selection for multi-frequency mixing in the spatial domain, which can dynamically adjust the fusion effect according to the feature contribution.

The High-Frequency Branch is a simple and efficient module whose main function is to obtain local high-frequency features. Considering that high-frequency information can be obtained by a small convolutional kernel, we obtain local high-frequency feature information by concatenating 1x1 and 3x3 regular convolutions [50]. To enhance the learning and generalization ability of self-attention, we designed to introduce the obtained high-frequency affinity matrix into the low-frequency affinity matrix, which is used to compensate for the lack of feature information of self-attention due to linear modeling. Let the input feature map be $F_i \in \mathbb{R}^{C\times H\times W}$. After confirming the 2D feature map by identity, the equivalent mapping, the feature map size is unchanged by the standard convolution with kernel sizes 1 and 3, and generates the high-frequency features F_h, the high-frequency affinity matrix F_{hm} formulas are as follows:

$$F_h = F_{c2}(F_{c1}(F_i)) \tag{1}$$

$$F_{hm}(i,j) = \phi(F_{c1}(i,j)) \otimes \phi(F_{c2}(i,j))^T \tag{2}$$

 doi:10.20944/preprints202406.0053.v1

Where ϕ denotes the nonempty set and $\otimes$ denotes the matrix multiplication, $F_h \in \mathbb{R}^{C\times H\times W}$, F_{c1} is the feature map obtained by 1×1 convolution, and F_{c2} is the feature map obtained by 3×3 convolution, both of size $\mathbb{R}^{C\times H\times W}$. Finally, through the matrix multiplication of F_{c1} and F_{c2} transpose, the high-frequency affinity matrix $F_{hm} \in \mathbb{R}^{N\times N}$ is obtained according to the window size, where N is the size of the partition window.

Low-Frequency Branch is a key part of capturing global contextual relationships, mainly using a multi-head self-attention mechanism [51]. This method first expands the input feature map $F_i \in \mathbb{R}^{C\times H\times W}$ in the channel dimension threefold through standard 1x1 convolution. After dividing into multiple heads, the window partition operation is applied to divide the 2D feature map into window sizes. Finally, the window feature map is flattened into a 1D sequence $\in \mathbb{R}^{3\times\left(\frac{H}{N}\times\frac{W}{N}\times h\right)\times(N\times N)\times\frac{C}{h}}$. After partitioning the input text into windows of size w and multiple heads of quantity h, it is further divided into three feature vectors: Query(Q), key(K), and Value(V), all with dimensions $\in \mathbb{R}^{\left(\frac{H}{N}\times\frac{W}{N}\times h\right)\times(N\times N)\times\frac{C}{h}}$. During the process of self-attention calculation, a learnable relative position encoding (Position Embedding, PE) is introduced to indicate the positional information of the image sequence. The low-frequency affinity matrix F_{lm} generated through multi-head self-attention is combined with the high-frequency affinity matrix F_{hm} to obtain the neutralized mixed affinity matrix F_{mm}. Finally, after normalization through the sigmoid function, the normalized mixed affinity matrix is multiplied with V to obtain the low-frequency feature map F_l after mixed linear weighting. The formula is described as follows:

$$F_{lm}(i,j) = \phi(Q_{(i,j)}) \otimes \phi(K_{(i,j)})^T \tag{3}$$

$$F_{mm}(i,j) = \phi(F_{hm}(i,j) \oplus \phi(F_{lm}(i,j) \tag{4}$$

$$F_l = Softmax\left(\frac{F_{mm}(i,j)}{\sqrt{d}} + PE_{(i,j)}\right) \otimes V(i,j) \tag{5}$$

where $\oplus$ denotes element level addition, $Softmax$ denotes the normalized activation function, $F_{lm}, F_{mm} \in \mathbb{R}^{N\times N}$, N is the size of the partition window, PE is the learnable positional encoding of window size, $F_l \in \mathbb{R}^{C\times H\times W}$.

High-Low Frequency Adaptive Fusion is a fusion mechanism built on spatial feature mapping. Inspired by the feature rescaling of SK-Net [52], the weights of the contribution values of the hybrid channel occupied by high-frequency features and low-frequency features are learned by designing different pooling methods, so that the network can pick a more appropriate multi-scale feature representation. Specifically, the obtained high-frequency feature $F_h \in \mathbb{R}^{C\times H\times W}$ and low-frequency feature $F_l \in \mathbb{R}^{C\times H\times W}$ are directly added together to fuse and obtain the mixed feature $F_m \in \mathbb{R}^{C\times H\times W}$. Then, the maximum pooling and average pooling are performed on this mixed feature to obtain the high-frequency attention feature map $A_{\mathrm{h}} \in \mathbb{R}^{H\times W}$ and low-frequency attention feature map $A_{\mathrm{l}} \in \mathbb{R}^{H\times W}$, respectively. The two spectral features are connected at the channel level, and the standard convolution smoothing filter with a size of 7×7 is applied to obtain $A \in \mathbb{R}^{2\times H\times W}$. After $Sigmoid$ activation in the fusion dimension, the high-frequency attention feature map $\hat{A}_h \in \mathbb{R}^{H\times W}$ and low-frequency attention feature map $\hat{A}_l \in \mathbb{R}^{H\times W}$ are obtained, and they are individually weighted by element-wise multiplication on the high-frequency feature map F_h and low-frequency feature map F_l. Finally, the weighted feature map results are added together to obtain the output result of the adaptive fusion, $F_o \in \mathbb{R}^{C\times H\times W}$. The relevant formulas are as follows:

$$F_m(i,j) = F_h(i,j) \oplus F_l(i,j) \tag{6}$$

$$A_{\mathrm{h}} = MaxPool(F_m), A_{\mathrm{l}} = AvgPool(F_m) \tag{7}$$

$$A = F_{\mathrm{conv}}^{7\times 7}(Concat(A_{\mathrm{h}}, A_{\mathrm{l}})) \tag{8}$$

$$\hat{A}_{\mathrm{h}}, \hat{A}_{\mathrm{l}} = Sigmoid(A, \mathrm{dim} = 0) \tag{9}$$

 doi:10.20944/preprints202406.0053.v1

$$F_o = \left(F_h \odot \hat{A}_{\mathrm{h}}\right) \oplus \left(F_l \odot \hat{A}_{\mathrm{l}}\right) \tag{10}$$

where *MaxPool* denotes global maximum pooling, *AvgPool* denotes global average pooling, *Concat* denotes channel-level splicing, *Sigmoid* denotes the activation function, and $F_{\mathrm{conv}}^{7\times7}$ denotes convolution with a kernel size of 7×7.

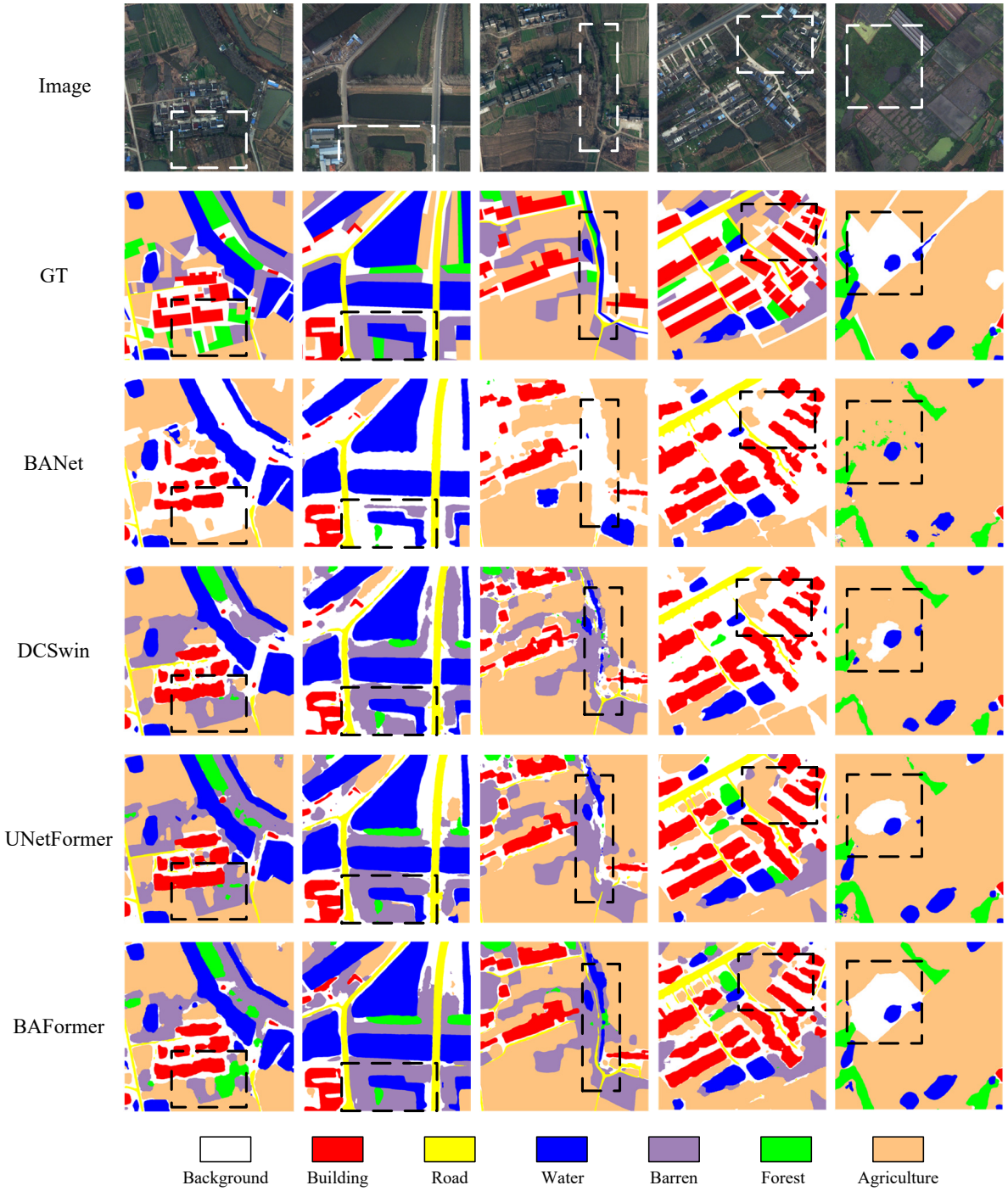


Figure 8. Qualitative comparisons with different methods on the LoveDA validation set.

 doi:10.20944/preprints202406.0053.v1

Table 3. Quantitative comparisons were made between our method and existing methods on the LoveDA dataset. The best values in this column are displayed in bold.

Method	Backbone	Per-Class IoU							mIoU
		Background	Building	Road	Water	Barren	Forest	Agriculture	
FCN [40]	VGG-16	42.6	49.5	48.1	73.1	11.8	43.5	58.3	46.7
DeepLabv3+ [41]	ResNet-50	43.0	50.9	52.0	74.4	10.4	44.2	58.5	47.6
SemanticFPN [53]	ResNet-50	42.9	51.5	53.4	74.7	11.2	44.6	58.7	48.1
FarctSeg [54]	ResNet-50	42.6	53.6	52.8	76.9	16.2	42.9	57.5	48.9
TransUNet [55]	Vit-R50	43.3	56.1	53.7	78.0	9.3	44.9	56.9	48.9
BANet [44]	ResT-Lite	43.7	51.5	51.1	76.9	16.6	44.9	**62.5**	49.6
SwinUperNet [56]	Swin-Tiny	43.3	54.3	54.3	78.7	14.9	45.3	59.6	50.1
DC-Swin [48]	Swin-Tiny	41.3	54.5	56.2	78.1	14.5	47.2	62.4	50.6
MaskFormer [57]	Swin-Base	**52.5**	60.4	56.0	65.9	**27.7**	38.8	54.3	50.8
UNetFormer [45]	ResNet-18	44.7	58.8	54.9	79.6	20.1	46.0	**62.5**	52.4
BAFormer-T	ResNet-18	45.9	57.9	58.2	79.0	19.0	47.3	61.4	52.7
BAFormer	ResNet-18	44.9	**60.6**	**58.6**	**80.4**	21.3	**47.5**	61.5	**53.5**

3.2.2. RAF (Realtionship Adaptive Fusion)

To obtain richer boundary features, fusing feature maps of different scales is considered to be an effective method to improve image effects [58]. Currently, the commonly used fusion methods include spatial numerical summation and channel dimensional splicing. However, shallow and deep features in the network do not play the same contribution in feature fusion. Generally, the shallow features have larger values and the deeper features in the network have smaller values, leading to differences in their spatial contributions. In addition, since shallow and deep features contain different semantic information, there is also some semantic confusion in the channel dimension. How improve the effect of feature fusion has become a new thinking direction to optimize the network performance. Inspired by the perceptual fusion of shallow and deep branches in ISDNet [59], we propose a dynamic fusion strategy RAF based on relational perception. This module obtains more complete boundary information by improving the feature granularity, and its detailed structure is shown in Figure 5.

Unlike other multi-scale static fusion methods, RAF can adaptively adjust the fusion of shallow and deep features according to the network task requirements and data characteristics by explicitly modeling the spatial and channel dependencies between features. While ensuring deep semantic transformation, it can fully use shallow features to achieve higher-quality feature reconstruction. Specifically, the method first learns the relational weight factors in the spatial dimension by modeling the spatial numerical differences between shallow and deep features through global flat pooling. Then, under the feature mapping of spatial modeling, the relationship weight factors on the channel dimension are learned by analyzing the channel relationship matrix after feature variation and compression sensing. Finally, the weighted fusion of features is performed under the dual relationship perception of space and channel. Let the shallow feature map $F_s \in \mathbb{R}^{C\times H_s\times W_s}$, and the deep feature map $F_d \in \mathbb{R}^{C\times H_d\times W_d}$, with $H_s \neq H_d, W_s \neq W_d$, RAF first aligns the spatial sizes of the shallow and deep feature maps, explicitly extracts the feature information, and obtains the two one-dimensional attention vectors $P_s, P_d \in \mathbb{R}^C$ that contain the information of their respective channels. It is expressed by the formula as follows:

$$P_s = GAP(F_s), P_d = GAP(Up(F_d)) \tag{11}$$

Where GAP denotes Global Average Pooling and Up denotes spatially sampled twice. In the second step, spatial dependencies and channel dependencies are modeled step by step. The two one-dimensional attention vectors P_s, P_d are taken as global average pooling to obtain the spatial relationship weight factors S_{ws}, S_{wd}, which are expressed as follows in Equation:

$$S_{ws} = GAP(P_s), S_{wd} = GAP(P_d) \tag{12}$$

When modeling the channel dependencies, considering that there are some differences between the channel semantics, the two one-dimensional attention vectors P_s, P_d are scaled to length r after the multilayer perceptron to obtain two contraction tensors P_{sr}, $P_{dr} \in \mathbb{R}^r$. Then, matrix multiplication is

 doi:10.20944/preprints202406.0053.v1

done based on these two contraction tensors to obtain the channel relevance matrix $R \in \mathbb{R}^{r \times r}$, which is mapped by the straightening and the multilayer perceptron into a channel weight factor C_{ws}, C_{wd} containing only two values with the following formula:

$$R = P_{sr} P_{dr}^{T}, C_{ws}, C_{wd} = \$\big(FlATTEN(R)\big) \tag{13}$$

Where \$ denotes the multi-layer perceptron, R denotes the channel relation matrix, and $FLATTEN$ denotes the operation of straightening the relation matrix. In the third step, dynamic weight fusion. The spatial weight factor and the channel weight factor are summed up, and the weighted values W_s and W_d of the shallow feature map and the deep feature map are obtained by $SOFTMAX$ in the first dimension, which is weighted and summed up with the dot-multiplication weighting of the shallow feature map F_s and the deep feature map F_d respectively to obtain the final fused feature map F_{raf}, with the following formulae:

$$W_s, W_d = SOFTMAX(S_{ws} + C_{ws}, S_{wd} + C_{wd}) \tag{14}$$

$$F_{raf} = (W_s \cdot F_s + F_s) + (W_d \cdot F_d + F_d) \tag{15}$$

3.2.3. DWLK-MLP (Deep Wide Large Kernel Multi-Layer Perceptron)

Enhancing the convolutional perceptual field is an effective means to improve semantic segmentation [60]. Recent studies have shown that the introduction of DW convolution into MLP multilayer perceptrons can effectively integrate the properties of self-attention and convolution, thus enhancing the generalization ability of the model [61]. Compared to ordinary MLP [51], DW-MLP [62] with residual structure introduces a 3x3-sized DW convolution into the hidden layer. This approach is effective in aggregating local information, mitigating the effects of the self-attention paradigm, and improving the generalization ability of the model. However, due to the large number of channels in the hidden layer, a single-scale convolution kernel cannot effectively transform channel information with rich scale features. To solve this problem, a multiscale feedforward neural network MS-MLP [63] has been proposed. He used DW convolution with kernel size [1,3,5,7] to capture multi-scale features. In this way, the performance of the model is enhanced to some extent. However, just using MLP to further transform the multi-scale features to enhance the generalization of the model is limited as it also undertakes the important task of extracting the feature maps for higher-level combination and abstraction.

To further improve the completeness of boundary features, we propose the simple and effective DWLK-MLP module as shown in Figure 6. This module increases the convolutional receptive field by deeply separating the large kernel convolutions, and more complete boundaries can be extracted with almost no computational overhead. Unlike other methods, DWLK-MLP introduces the idea of large kernel convolution, which can take on more advanced abstract feature extraction tasks by creating a large kernel receptive field. Specifically, we introduce a deep large kernel convolution of 23x23 size in front of the activation function. The final result is obtained by summing up the initial feature map with the feature map after the large kernel convolution using jump concatenation. To reduce the number of parameters and computational complexity, we use two depth convolution sequences of 5x5 and 7x7 for decomposition. This approach exploits the lightweight nature of the depth-separable computational paradigm and promotes the fusion of self-attention and convolution to improve network generalization. Numerous experiments have demonstrated that the introduction of deep large kernel convolution before the activation function improves the accuracy and robustness of image recognition more than after the activation function.

 doi:10.20944/preprints202406.0053.v1

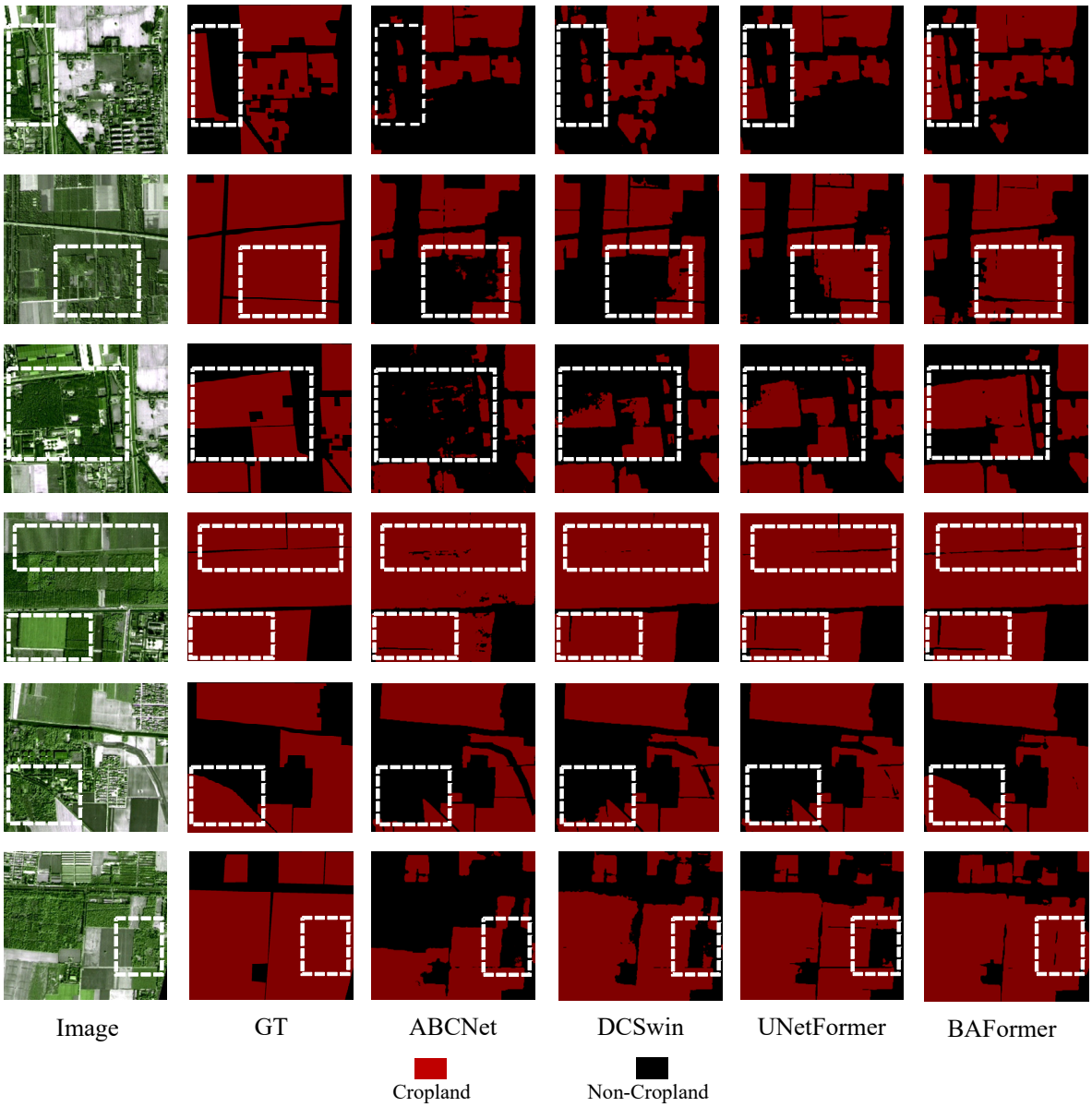
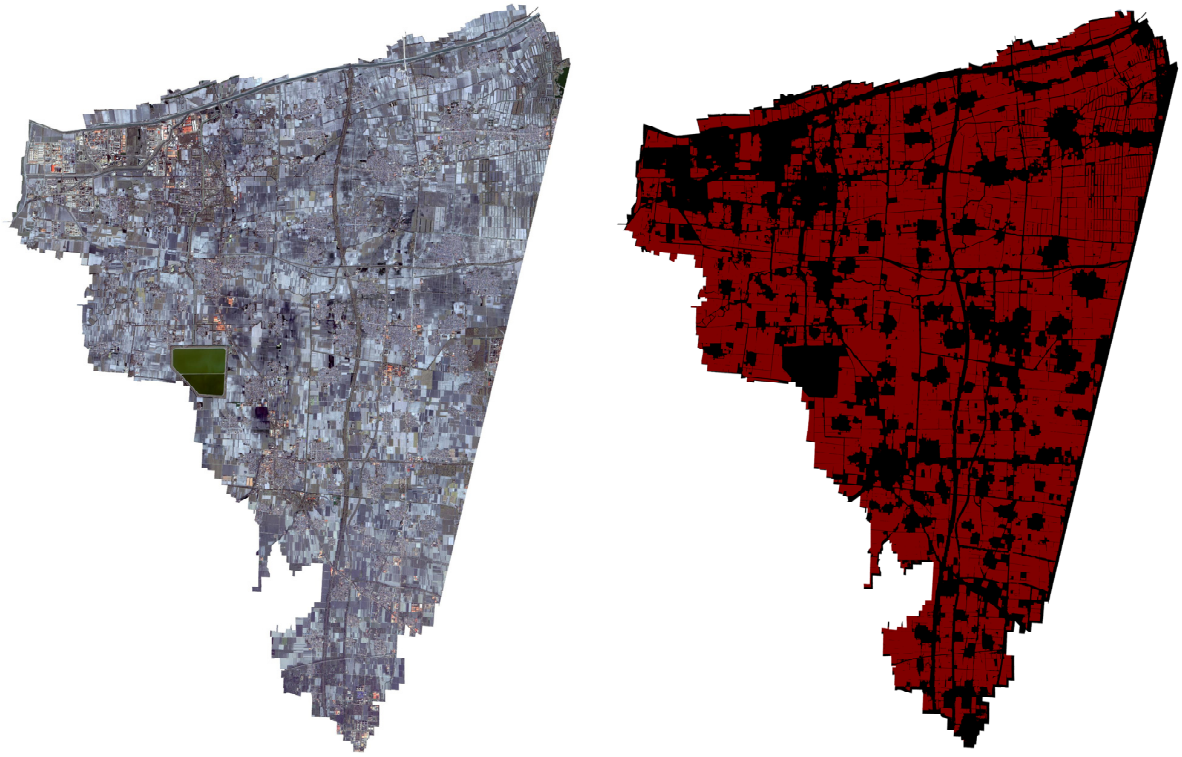


Figure 9. Qualitative comparisons with different methods on the Mapcup test set.

 doi:10.20944/preprints202406.0053.v1



(a) Large Image (b) Inference Image

Figure 10. Inference visualization was performed in a randomly selected region in the north. Red represents Cropland and black represents non-Cropland. (**a**) High-resolution remote sensing large image. (**b**) Visualization of model inference.

Table 4. Comparison of different methods on the Mapcup dataset.

Method	Backbone	CultivatedLand	Non-Cropland	Mean F1(%)	OA(%)	mIoU(%)
FCN [40]	VGG-16	81.6	86.8	84.2	84.7	72.5
DeepLabv3+ [41]	ResNet-50	82.7	87.5	85.1	85.5	74.2
A2FPN [64]	ResNet-18	83.2	87.8	85.5	85.9	74.8
ABCNet [43]	ResNet-18	84.0	88.1	86.1	86.4	75.6
MANet [46]	ResNet-50	86.0	89.3	87.7	87.7	78.1
BANet [44]	ResT-Lite	86.7	89.8	88.3	88.5	79.0
DC-Swin [48]	Swin-S	86.8	89.7	88.2	88.4	79.0
UNetFormer [45]	ResNet-18	87.2	90.3	88.8	88.5	79.6
FT-UNetFormer [45]	Swin-B	88.7	91.0	89.8	90.0	81.6
BAFormer-T	ResNet-18	88.2	90.8	89.5	89.6	81.0
BAFormer	ResNet-18	**89.8**	**91.7**	**90.7**	**90.8**	**83.1**

3.2.4. Edge Constraints with Deep Supervision

The deeper the network layers, the richer the high-level semantic information [65]. By visualizing the layers of the network, we find that shallow feature maps are relatively detailed and highlight local features, while deep feature maps are relatively smooth and highlight global features. To optimize edge detail, we propose an edge constraint strategy using deep supervision at the deeper layers of the network. Unlike other shallow constraint methods, our proposed method starts guiding the model to automatically focus on the boundary features at the deep layer of the network and further consolidates the correct boundary information during the up-sampling process. Specifically, we extract the same number of channels as the number of categories at the deepest feature layer of the model encoder, with the same scale labels applied to the edge cross-entropy constraints to do the supervision. The method effectively enhances boundary supervision by adding only negligible computational complexity without increasing any model parameters. The method significantly improves 0.5% mIoU

 doi:10.20944/preprints202406.0053.v1

on the Vaihingen dataset, achieving encouraging results. The deep edge constraint can be expressed as follows:

$$y = \delta(F_{\text{num\_class}}) \tag{16}$$

$$\mathcal{L}_{\text{edge}} = -\frac{1}{S}\sum_{s=1}^{S}\sum_{p=1}^{P} y_p^{(s)} * \log(\hat{y}_p^{(s)}) \tag{17}$$

where δ is the deepest feature map of the network, $F_{\text{num\_class}}$ refers to the number of channels to take the number of classes of that feature map, and $\mathcal{L}_{\text{edge}}$ refers to the cross-entropy constraints made with the labels.

3.3. Loss Function

The model mainly uses Cross-Entropy Loss ($\mathcal{L}_{\text{ce}}$) and Dice Loss ($\mathcal{L}_{\text{dice}}$). Two joints are acting on the main loss of segmentation $\mathcal{L}_{\text{main}}$ in the prediction graph. To improve the segmentation effect of multi-scale targets, an auxiliary loss constraint $\mathcal{L}_{\text{aux}}$ based on the cross-loss is introduced in the middle layer. To guide the model to focus on the boundaries from the deep layer of the network, we innovate to introduce the depth-supervised edge constraint $\mathcal{L}_{\text{edge}}$. Through the above elaboration, the total loss L of the model is the sum of three kinds of losses, namely, $\mathcal{L}_{\text{main}}$, $\mathcal{L}_{\text{aux}}$, and $\mathcal{L}_{\text{edge}}$. $\mathcal{L}$ the sum of three kinds of losses, the formula is expressed as follows:

$$\mathcal{L}_{\text{ce}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_k^{(n)} \log\left(\hat{y}_k^{(n)}\right) \tag{18}$$

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{y_k^{(n)}\hat{y}_k^{(n)}}{y_k^{(n)} + \hat{y}_k^{(n)}} \tag{19}$$

$$\mathcal{L}_{\text{main}} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{dice}}, \quad \mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{ce}} \tag{20}$$

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{edge}} \tag{21}$$

In the equation, N represents the number of samples, and K represents the number of categories. $y_k^{(n)}$ represents the one-hot encoding of the k-th semantic label in the n-th sample, while $\hat{y}_k^{(n)}$ represents the confidence level of predicting category k in the n-th sample.

4. Experiments

We conducted many experiments to compare the performance of our model with other state-of-the-art segmentation models, validating the effectiveness of our model in several ways. Specifically, we evaluated our model on three public datasets, Vaihingen, Potsdam, and LoveDA, against Mapcup's homegrown dataset to validate its ability to generalize cropland extraction. In addition, we conducted numerous ablation experiments to demonstrate the scientific validity of the module components and parameter settings from multiple perspectives.

4.1. Experimental Setup

4.1.1. Dataset

Vaihingen: The dataset consists of 33 top-view image patches with very fine spatial resolution, averaging 2494 × 2064 pixels in size. Each image patch contains three multispectral bands (near-infrared, red, green), as well as a Digital Surface Model (DSM) and Normalized Digital Surface Model (NDSM) with a ground sampling distance of 9 centimeters (GSD). The dataset includes five foreground classes (impervious surfaces, buildings, low vegetation, trees, and cars) and one background class (clutter). For the specific experiments, we tested using image patches with IDs 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, and 38 for training, and used image patch with ID 30 for validation. The

 doi:10.20944/preprints202406.0053.v1

remaining 15 image patches were used for training. To facilitate semantic segmentation, the image patches were slidingly cropped into patches of size 1024 × 1024 pixels with a sliding size of 512.

Potsdam: The dataset comprises 38 top-view image blocks with extremely high spatial resolution, a ground sampling distance of 5 centimeters, and an image size of 6000×6000 pixels. Similar to the Vaihingen dataset, it covers the same class information. Each image block provides four multispectral bands (red, green, blue, and near-infrared), as well as a digital surface model (DSM) and normalized digital surface model (NDSM). For the experiment, we selected image blocks with the IDs 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6 _13, 6_14, 6_15, and 7_13 for testing, and used the image block with ID 2_10 for validation. The remaining 22 image blocks (excluding images 7_10 due to incorrect annotation) were used for training. Only three bands (red, green, and blue) and the original images were utilized during the processing. During actual training, we sliced the image blocks into patches of size 1024×1024 pixels with a sliding window set at 1024.

LoveDA: The dataset [66] is a collection of 5987 optical remote sensing images, with each image having a resolution of 1024×1024 pixels and a ground sampling distance of 0.3 meters. The dataset covers 7 different land cover classes, including buildings, roads, water, barren land, forests, agriculture, and background. In the entire dataset, there are 2522 images for training, 1669 images for validation, and an additional 1796 images provided by the official for testing. These images were captured from urban and rural scenes in three cities in China (Nanjing, Changzhou, and Wuhan). As a result, this dataset presents multi-scale targets, complex backgrounds, and inconsistent class distributions, posing significant challenges for object recognition.

Mapcup: The dataset is a self-constructed dataset containing 507 high-resolution remote sensing images with a resolution of 1024 × 1024. the ground sampling rate of the images is 0.6 m and covers both cultivated and non-cultivated areas. In the dataset, a total of 373 images were used for training, while 134 were used for validation and testing. In addition, one large high-resolution remote sensing image from actual production was cropped into 1160 images for cartographic inference. These images were acquired from the northern plains region and were finely labeled. They have significant features such as complex scenes, multi-scale targets, and unbalanced data categories, which bring great challenges to the extraction of cropland from high-resolution remote sensing images.

4.1.2. Implementation Details

We used the PyTorch framework for our experiments and all models were implemented on a single Nvidia GTX 3090 GPU. The encoder uses the ResNet-18 pre-trained weights from the timm library, which accelerates model training by migrating the weights. To improve the model convergence speed, we used the AdamW optimizer for training and set the base learning rate to 6e-4 with a cosine learning rate scheduling strategy. To address the complex and variable characteristics of the Potsdam and LoveDA datasets, we performed several data enhancement operations on the input data of size 1024 × 1024 during the training process, including Gaussian blurring, random scaling factors [0.5, 0.75, 1.0, 1.25, 1.5], random horizontal and vertical flipping, and random brightness contrast adjustment. In the training phase, we set 45 epochs to train the model and set the batch size to 4. In the testing phase, we used the test time enhancement (TTA) strategy, which includes operations such as random horizontal and vertical flipping. Considering the lightweight nature of Vaihingen images and the regularity of most targets, we randomly cropped the images into patches of 512 × 512 size for training. In the training phase, we used a random scaling factor [0.5, 0.75, 1.0, 1.25, 1.5], random horizontal vertical flipping, and random rotation for non-destructive data enhancement techniques. The number of training times is set to 105 epochs and the batch size is set to 4. In the testing phase, we use the same data enhancement operations of multi-scale and random flipping as in the above dataset. For the homemade Mapcup dataset, we adopt the same training strategy as Potsdam due to the complexity of feature classes and the multi-scale variation of targets. Meanwhile, all models were compared on this dataset using the same profile and training strategy to ensure the validity and fairness of the results.

 doi:10.20944/preprints202406.0053.v1

4.1.3. Evaluation Indicators

The evaluation of the model follows the ISPRS benchmark. To assess the effectiveness of the model, we adopt three commonly used semantic segmentation metrics: Intersection over Union (IoU), Overall Accuracy (OA), and F1 score. Among them, IoU is the most common and important evaluation metric, as it intuitively reflects the model's ability to separate foreground and background. A higher IoU value indicates more accurate segmentation results. F1 score is the harmonic mean of precision and recall, and it is used to comprehensively evaluate the accuracy and recall performance of the model. Precision measures the proportion of true positives among the samples predicted as positives by the model, while recall measures the proportion of correctly predicted positive samples among all true positive samples. Therefore, the F1 score considers both metrics and provides a better assessment of the overall performance of the model. As for OA, it mainly measures the prediction accuracy of the model at the pixel level. We define TP as true positives, TN as true negatives, FP as false positives, and FN as false negatives. Based on these definitions, we can calculate the value of OA using the corresponding formulas. Through the evaluation of these three metrics, we can comprehensively measure the performance of the model and effectively evaluate and compare different models. The use of these metrics not only makes the evaluation results more objective and accurate, but also increases people's trust in the effectiveness of the model. The relevant formulas can be expressed as follows:

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

$$Recall = \frac{TP}{TP + FN} \tag{23}$$

$$F1 = 2x \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{24}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{25}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{26}$$

4.2. Experimental Result

4.2.1. Results on Vaihingen Dataset

In order to validate the effectiveness of our proposed method, we conducted many comparative experiments. The results of the quantitative comparison with state-of-the-art models on the ISPRS Vaihingen test set are shown in Table 1. We compare classical semantic segmentation algorithms such as FCN [40], and DeepLabV3+ [41], and advanced CNN-encoder-based semantic segmentation algorithms such as MAResU-Net [42], ABCNet [43], BANet [44], UNetFormor [45], MANet [46], and others. In addition, we also compare advanced large models such as DC-Swin [48], Mask2Former [47], and FT-UNetformer [45] based on the Transformer-encoder. As shown in the experimental results, our proposed BAFormer achieves the best score on the Vaihingen test set, reaching 91.5% MeanF1, 91.8% OA, and 84.5% mIoU, which demonstrates its advanced performance. mIoU index is the most important evaluation metric reflecting the performance of the model's foreground and background segmentation. Compared with the traditional classical semantic segmentation algorithms, BAFormer achieves a significant improvement in mIoU by 10.3% over FCN and 8.2% over DeepLabV3+. Compared with the state-of-the-art MAResU-Net, ABCNet, BANet, UNetFormor, and MANet models that use Resnet-18 blocks as Encoder, BAFormer achieves a 1.8% mIoU improvement over the best-performing UNetFormer network, proving the effectiveness of our method. Compared with the Transformer-encoder-based DC-Swin, FT-UNetformer, and Mask2Former state-of-the-art macro

 doi:10.20944/preprints202406.0053.v1

models, our proposed lightweight model BAFormer-T also outperforms the best-performing FT-UNetformer macro models, achieving encouraging results. Notably, our approach achieves a score of 90.9% on the " Car " class, outperforming most of the state-of-the-art models by about 1-2% points. This is because our proposed DWLK-MLP can facilitate the fusion of CNN and Self-Attention while deepening the capture of boundary information through the large kernel sensing field, which improves the recognition accuracy of small targets.

The qualitative comparison of the ISPRS Vaihingen test set is shown in Figure 7. Three representative typical examples were selected to compare the effectiveness of our model, and some black dashed boxes were added as areas of focus for comparison. As demonstrated in the first row of results, there are three buildings with some shadows at intervals, with similar colors and textures to the low-level vegetation, and there is a large inter-class similarity. The same convolution-based semantic segmentation methods such as BANet and UNetFormer misclassify this building area as low-level vegetation, while our model makes an accurate prediction. This shows that our proposed BAFormer method can fully learn the global spatial contextual relationships and correctly classify the inter-class differences by modeling the long-distance dependency extraction, not only limited to the local prediction of local color and texture features. As shown in the second line of results, a piece of opaque foreground road is placed on a piece of clutter with texture and shape significantly different from the surrounding road, with high intra-class variability. As the BANet, the UNetFormer network misclassifies and omits most of the area of this grocery as an opaque water foreground, whereas our BAFormer can correctly classify it by significantly extracting the interaction representation of foreground and background. This interaction situation also illustrates the importance of foreground-background category balance in the remote sensing semantic segmentation problem, which is one of the key issues that previous methods have failed to solve. As shown in the results in the third row, under a complex area surrounded by low-level vegetation, the category targets have been distorted and deformed due to the irresistible factors related to sensor imaging and image processing, which can no longer be distinguished by the naked eye, while the classification results obtained by our proposed method are the closest to the real labels, and more accurate boundary details are obtained. Overall, our proposed BAFormer method learns global contextual correlations well, fully extracts inter-class differences and intra-class variation features, models the interactive dependence of foreground and background, and obtains segmentation maps with higher accuracy and richer details.

4.2.2. Results on Potsdam Dataset

On the ISPRS Potsdam test set, we performed quantitative comparisons with state-of-the-art models, and the comparison results are shown in Table 2. In our experiments, we compared FCN [40], DeepLabV3+ [41], MAResU-Net [42], ABCNet [43], BANet [44], UNetFormor [45], MANet [46], Mask2Former [47], SwinTF-FPN [49], and FT-UNetFormer [45] excellent methods. From the results, our proposed BAFormer achieves impressive results on this dataset, obtaining advanced scores of 87.3% mIoU, 93.2% F1, and 92.2% OA, which have already surpassed the state-of-the-art Mask2Former, and FT-UNetFormer segmentation models. The best accuracy is achieved in the most easily distinguishable " Car " class and " Building " class, and the mIoU is improved by 1-2% compared to other advanced convolution-based models. This proves that our method has better feature recognition ability when dealing with this small-scale or large-scale feature target. Meanwhile, BAFormer also obtains the best accuracy results in the " Low.veg " category, which is the most difficult to identify, which fully proves the powerful feature learning and characterization ability of the model. In summary, our method performs well on the ISPRS Potsdam test set, not only outperforming many state-of-the-art semantic segmentation algorithms, but also achieving higher accuracy on specific categories.

Similarly, we performed qualitative experimental comparisons on the ISPRS Potsdam test set, and the results of the local visualization plots are shown in Figure 7. We selected three representative samples for qualitative analysis. As shown in the results of the fourth row, there is a region of opaque water (square) in the center of a piece of building, and it has been difficult for the naked eye

 doi:10.20944/preprints202406.0053.v1

to distinguish whether the region is a building, an opaque water surface, or a background. Other convolution-based models BANet and UNetFormer either misclassify the background or predict fuzzy and rough boundaries. In contrast, our model's predictions are almost perfectly close to the labels and achieve better segmentation results. This result shows that our proposed method can fully understand the long-range dependencies of spatial context and make correct and accurate predictions. As shown in the result in the fifth line, a ring of fuzzy noise appears around the building in the dashed box, which belongs to the background class of intra-class mutation. All other models identify this intra-class mutation type as the building ontology, and only our method BAFormer can fully identify the anomalous samples within the class and make correct predictions, indicating that our proposed method is effective in dealing with highly similar targets between classes. The sixth row of results shows some discrete, but not significantly different, low-level vegetation and trees in the dashed box, and only our BAFormer can adequately characterize the interclass differences, obtaining a more detailed and accurate edge recognition. To briefly summarise, our proposed BAFormer can extract richer full-local detail features, resulting in more accurate visualization results.

Table 5. It is compared with the current state-of-the-art lightweight network on the Vaihingen dataset. All experiments were performed on a single NVIDIA GTX 3090 GPU using 1024× 1024 inputs to test its complexity and model parameters.

Method	Backbone	Memory(M)	Params(M)	Complexity(G)	mIoU(%)
DANet [44]	ResNet-18	611.1	12.6	39.6	68.8
BiSeNet [67]	ResNet-18	970.6	12.9	51.8	69.1
Segmenter [68]	ViT-Tiny	933.2	13.7	63.3	73.6
BoTNet [69]	ResNet-18	710.5	17.6	49.9	74.3
FANet [70]	ResNet-18	971.9	13.6	86.8	75.6
ShelfNet [71]	ResNet-18	579.0	14.6	46.7	78.3
SwifNet [72]	ResNet-18	835.8	11.8	51.6	79.9
ABCNet [43]	ResNet-18	1105.1	14.0	62.9	81.3
MANet [46]	ResNet-50	1169.2	12.0	51.7	82.7
UNetFormer [45]	ResNet-18	1003.7	11.7	46.9	82.7
BAFormer-T	ResNet-18	1067.3	12.8	51.3	84.1
BAFormer	ResNet-18	2668.3	35.5	142.0	84.5

Table 6. Ablation on the Vaihingen dataset investigates the effect of different input sizes on model stability. The experiments were carried out on a single NVIDIA GTXX 3090 GPU using the BAFormer-T model.

Input_Size	Imp.surf	Building	Low.veg	Tree	Car	MeanF1(%)	OA(%)	mIoU(%)
512x512	93.19	95.83	85.36	90.89	89.39	90.93	91.48	83.63
768x768	93.56	95.93	85.42	90.72	90.14	91.15	91.60	83.94
1024x1024	93.67	95.89	85.33	90.79	90.87	91.30	91.63	84.16
2048x2048	93.40	95.79	85.60	90.73	89.88	91.08	91.54	83.80

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 June 2024 doi:10.20944/preprints202406.0053.v1

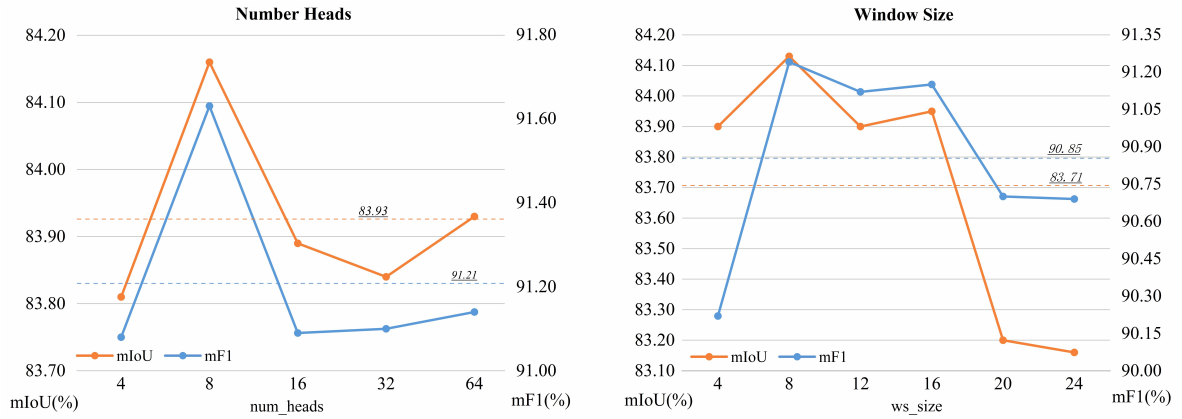


Figure 11. Ablation study on the number of model multi-heads and window size on the Vaihingens dataset.

4.2.3. Results on LoveDA Dataset

To further assess the validity of the model, we performed a comprehensive evaluation in both quantitative and qualitative terms by making many comparisons with state-of-the-art models on the LoveDA dataset. The quantitative results are shown in Table 3, while the qualitative results are exhibited in Figure Figure8. In our experiments, we compare the classical semantic segmentation algorithms represented by FCN [40] and DeepLabV3+ [41], as well as the current mainstream state-of-the-art convolutional lightweight methods such as SemanticFPN [53], FarctSeg [54], TransUNet [55], BANet [44], UNetFormer [45], and Transformer-encoder based SwinUperNet [56], DC-Swin [48], MaskFormer [57] large model semantic segmentation methods. From the experimental results, our proposed BAFormer achieves the highest mIoU of 53.5%, which exceeds the mIoU of the state-of-the-art UNetFormer model by 1.1%, and also exceeds the mIoU of the Transformer-based MaskFormer by 2.7%. BAFormer achieves the highest mIoU of 53.5% in the three significant feature segmentation methods, namely Road, Water, and Forrest, which are the three categories of salient features, achieved the best segmentation results. Still, the recognition of Background and Barren, which are two complex and variable non-salient feature categories, is not satisfactory. This phenomenon suggests that our proposed method enhances the learning and identification ability of salient feature targets, but still has some shortcomings compared to mask-based processing of complex and variable non-salient feature targets.

4.2.4. Results on Mapcup Dataset

In order to evaluate the effectiveness of the models for segmentation in real production environments, we further conducted experimental tests on a homemade Mapcup dataset, and the quantitative results are shown in Table 4. In this experiment, we compared several excellent state-of-the-art models, including FCN [40], DeepLabV3+ [41], A2FPN [64], ABCNet [43], MANet [46], BANet [44], DC-Swin [48], UNetFormer [45], and FT-UNetFormer [45]. From the experimental results, BAFormer achieves satisfactory results, obtaining 90.7% F1, 90.8% OA, and 83.1% mIoU, which is attributed to the model's hybrid extraction and selection of high-frequency features and low-frequency features. Especially in the segmentation of the "Cropland" category, the characterization ability of BAFormer is significantly enhanced. Compared with the state-of-the-art convolution-based UNetFormer model, the mIoU value of BAFormer is improved by 2.5%; compared with the state-of-the-art FT-UNetFormer model based on Transformer, the mIoU value of BAFormer is improved by 1.5%. Our proposed model not only achieves the best segmentation results among models with the same volume but also outperforms larger models like DC-Swin and FT-UNetFormer, achieving a better balance between model parameters and accuracy. This result fully demonstrates the advancement of our proposed BAFormer model.

The qualitative results on the Mapcup test set are shown in Figure9. We have selected six visualization results for presentation. Looking at the results in the first, second, and third rows, we can

 doi:10.20944/preprints202406.0053.v1

see that different types of cropland exist within the dashed box, including rice cropland with different lengths, cropland full of low-level vegetation, fruit and vegetable gardens, and recreational land. These different types of cropland differ somewhat in color and texture, and there is a high degree of category mutability between them. In contrast, our BAFormer model can learn the similarities between different mutants well enough to fully characterize mutants of the same category. This result fully demonstrates that even in the face of complex and variable scenes, our proposed method has strong feature learning and fitting characterization capabilities, and can be widely applied to real-world environments with complex and variable scenes. In the fourth line of results, we can observe the existence of a road between the croplands in the dashed box of the image, which is almost indistinguishable from the naked eye. However, our method learns sufficiently about the variability between categories to adequately characterize and distinguish those less significant category differences. This is also demonstrated in the results in rows 5 and 6, where some of the croplands covered with low-level vegetation are texturally distinct from the surrounding agricultural croplands, and the intra-category variability between them is so great that it is difficult to make a distinction based on texture and shape alone. However, our model can make full use of global contextual features and combine them with spatially detailed features to correctly classify them. Overall, our proposed BAFormer approach can fully learn the inter- and intra-category differences and similarities when dealing with complex and changing real-world scenarios and has a strong feature-fitting capability to cope with complex contextual content.

To further observe the performance of the model in the production environment, we randomly selected some areas for inference visualization, and the results are shown in Figure10. By analyzing the inference results of the visualization, we find that BAFormer can effectively eliminate the interference of the feature background, has accuracy in the boundary identification of various complex targets, and achieves a satisfactory segmentation quality.

4.3. Ablation Experiment

4.3.1. Each Component of BAFormer

In BAFormer, we effectively improve the accuracy of the model for boundary-aware segmentation by compensating the three aspects of feature extraction, feature fusion, and loss constraints. To verify the effectiveness of each module even further, we conducted extensive ablation experiments on the Vaihingen dataset, and the results are shown in TABLE 8. It should be noted that to make a fairer comparison, we uniformly adopt the same stochastic enhancement strategy and test-time enhancement strategy, and the relevant hyperparameters are consistent with the benchmark model UNetFormer by default.

Table 7. Ablation of large core selection in DWLK-MLP. K represents the size of the convolution kernel and D represents the dilation rate of the convolution.

Kernel_Size	(K,D) Sequence	Flops(G)	Paras(M)	Mapcup mIoU(%)	Vaihingen mIoU(%)
11	(3,1) → (5,2)	18.40	12.74	82.77	84.16
23	(5,1) → (7,3)	18.62	12.78	83.11	84.47
29	(3,1) → (5,2) → (7,3)	18.67	12.79	82.86	84.22
35	(5,1) → (11,3)	19.02	12.85	82.54	84.08

 doi:10.20944/preprints202406.0053.v1

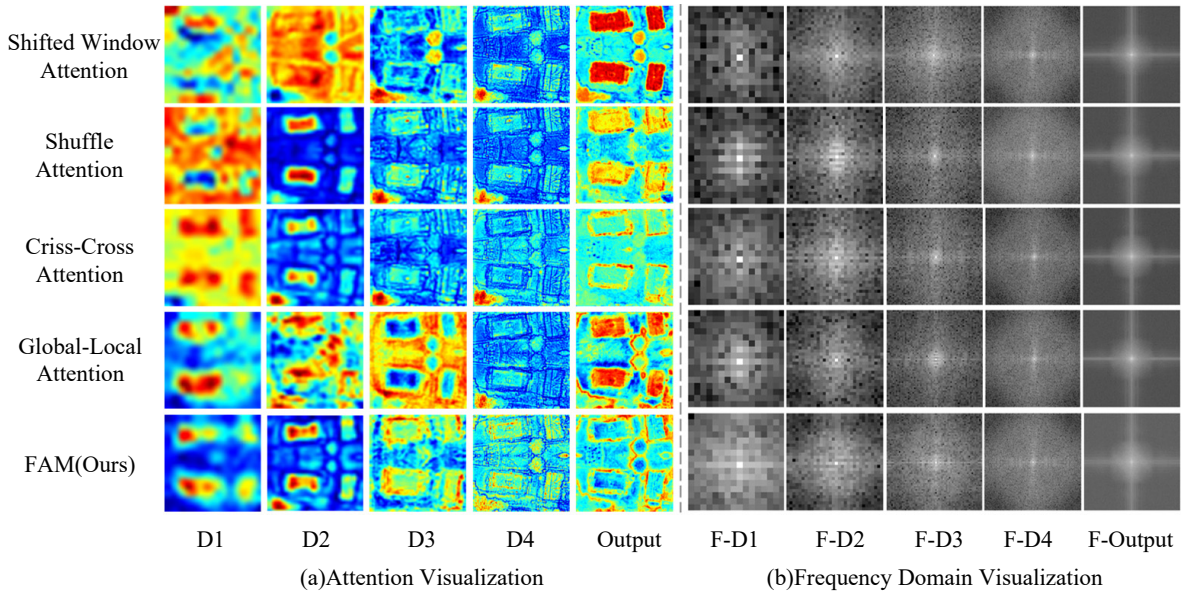


Figure 12. Feature Adaptive Mixer(FAM) feature map visualization.

4.3.2. Selection of Large Kernel Convolution

The DWLK-MLP module is designed to enhance the completeness of the boundary extraction by deeply decomposing the large kernel sensory fields. To further explore the effect of using large kernel convolution on the module, we performed a correlation validation on Mapcup and Vaihingen datasets, and the results are shown in TABLE 7. The experimental results show that choosing a large kernel convolution with a size of 23 yields the best accuracy, whereas a further increase in the kernel size instead reduces the accuracy and increases the model complexity. Therefore, we set the size of the large kernel convolution to 23 by default and use the sequence of small kernel convolutions of sizes 5 and 7 for DW decomposition. This not only maintains the sensing field of the large kernel, but also avoids excessive computational complexity and increases the depth of the model, which improves the generalization ability of the model.

4.3.3. Lightweight Model

To evaluate the outstanding lightweight features of BAFormer-T, we compared it with the current state-of-the-art lightweight models on the Vahingen public dataset. By comparing the results (see Table 5), we found that BAFormer-T outperforms the lightest DANet model, improving the mIoU score by 15.3%. In addition, compared to the similarly sized state-of-the-art lightweight model UNetFormer, BAFormer-T has hardly increased memory consumption, parameter count, and computational complexity, yet achieved a satisfactory mIoU score of 84.1%, surpassing UNetFormer's mIoU by 1.4%. This fully demonstrates the perfect balance between model accuracy and complexity achieved by BAFormer-T. Furthermore, this lightweight high-accuracy result further proves the effectiveness of channel fusion-based mixed-frequency feature extraction and deep large-kernel multi-layer perception methods. They can efficiently run models in resource-constrained environments and provide feasible solutions for saving computational resources and deployment costs. By incorporating these excellent lightweight techniques into model design, we can achieve more efficient, flexible, and cost-effective model deployments, bringing more opportunities and challenges to various industries.

4.3.4. The Stability of Model

In practical applications, the stability and adaptability of the model become particularly important in the face of different input sizes. To more comprehensively evaluate the stability of the model's performance under different input sizes, we train the lightweight BAFormer-T by using square image inputs with different scales, including common 512x512, 768x768, 1024x1024 scales as well as the

 doi:10.20944/preprints202406.0053.v1

super-large 2048x2048 input scale, and the results are shown in TABLE 6. From the results, it can be seen that the lightweight BAFormer-T shows excellent stability and adaptability when dealing with different input scales. The overall mIoU deviation is no more than 0.6%, MeanF1 deviation is no more than 0.4%, and OA deviation is no more than 0.2%. This means that the model can maintain good performance when dealing with input data of different sizes, with excellent robustness and generalization ability, and is suitable for diverse practical application scenarios.

Table 8. Ablation study of BAFormer for each component on the Vaihingen dataset.

Dateset	BASE	METHOD				mIoU(%)
		AFM	RAF	DWLK-MLP	EdgeConstraint	
Vaihingen	✓					82.50
	✓	✓				83.61
	✓		✓			83.09
	✓			✓		83.44
	✓				✓	83.26
	✓	✓	✓			83.89
	✓	✓	✓	✓		84.22
	✓	✓	✓	✓	✓	84.48

4.3.5. Choice of Number Of Multi-heads and Window Wize

In BAFormer, abstract semantic features are mainly extracted by Transformer blocks. The extraction of these features is influenced by two important hyperparameters, the number of heads and the partition window size, which directly affect the attention performance of the model. To further investigate the setting of the number of multi-heads and window size, we conducted a series of experiments. The quantitative experimental results regarding the number of multi-heads and window size are shown in Figure 11. In the ablation experiment with the number of multiple heads, we found that the parameter setting of multiple heads should conform to the law of the number of feature channels as much as possible, instead of learning stronger with more numbers. For the lightweight BAFormer-T model with 64 Decoder channels, the best effect is achieved when num_heads is set to 8. Setting too large or too small will inhibit the feature extraction performance of the model. In the ablation experiments with window size, we found that the window size of 8 had the best overall performance. However, if the window becomes larger again, both mIoU and F1-Score decrease.

Table 9. Ablation study with different encoders on the Vaihingen dataset.

Model	Encoder	Params(M)	Complexity(G)	Flops(G)	mIoU(%)
BAFormer-T	ResNet-18	12.78	51.33	18.62	84.15
	ResNet50	25.30	191.31	31.06	83.35
	ResNet101	44.30	177.36	50.56	83.30
	EfficientNet	63.80	255.21	35.41	83.61
	Swin_Base	112.47	452.60	230.87	84.27
BAFormer	ResNet-18	34.88	141.97	147.14	84.47

 doi:10.20944/preprints202406.0053.v1

5. Disscuss

5.1. FAM Feature Visualization

The proposed FAM has the property of dynamic learning. To validate the effectiveness of this method, we visually compared it with classical attention mechanisms such as Shift Window Attention, Shuffle Attention, Criss-Cross Attention, and Global-Local Attention. To demonstrate our viewpoint from multiple perspectives, we visualized the features after attention in both spatial and frequency domains in the decoder, as shown in Figure 12. From the spatial environment (a), it can be seen that our proposed FAM Attention maximally preserves the global and local features of the image. In the D1 deep layer with only semantics and the fusion layer D2, D3, D4, Output fused with spatial fine-grained features, the network further enhances boundary perception after FAM Attention, better preserving smooth spatial context features and clear edges, textures, and other fine-grained features compared to other methods. In the frequency domain environment (b), it can be observed that the image's white halo (low-frequency features) is more round and extensive in range compared to other methods through the F-D1, F-D2, and F-D3 layers. This is due to the adaptive selection mechanism based on channel fusion contribution in FAM, which can mix high-frequency and low-frequency features in different proportions adaptively according to different network depths, thereby achieving more delicate and rich feature representation.

5.2. About The Choice of Encoder

Considering the important impact of Encoders with different feature extraction capabilities on the decoding fusion of deep Decoders and the effectiveness of the overall model, we further explored the influence of different Encoder models on the overall segmentation of the model, as shown in Table 9. The research results indicate that ResNet-18 achieves a better overall fusion effect due to its efficient extraction of shallow features, thus achieving the best segmentation results. In addition, the lightweight BAFormer-T model achieved a satisfactory mIoU score of 84.15 on the Vaihingen dataset, while BAFormer even achieved an excellent mIoU score of 84.47. It is worth noting that the BAFormer-T lightweight model reduced the model parameters, complexity, and computational cost by 50% compared to other CNN encoders and Transformer encoders, without reducing the accuracy level. At the same time, BAFormer achieved significant advanced performance compared to large and complex encoders such as EfficientNet and Swin_Base in terms of parameter quantity, complexity, and computational cost.

From the experimental results, it is generally found that using a Transformer-based encoder can achieve better accuracy compared to a CNN-based encoder. However, BAFormer enhances the information flow and expressive power of the network in the Decoder stage by selectively adapting different feature frequency bands based on channel transmission. This experimental result further proves that in the image reconstruction stage of the Decoder, apart from needing strong support from advanced semantic features, restoring clearer and more accurate image information also requires the fusion of more shallow fine-grained features. We can decode advanced semantic features through a simple encoder to achieve better feature fusion and denser feature representation.

5.3. Further Exploration on Edge Constraints

The deeply supervised edge constraint strategy significantly improves the effectiveness of boundary constraints on feature targets and enhances edge accuracy. Taking the Vaihingen dataset as an example, the mIoU accuracy is improved by 0.5% overall. However, for those image datasets where the feature classes are more difficult to recognize, there will be a gap in the improvement effect of this method. Further exploration and analysis of the effectiveness of the Vaihingen dataset reveal that the dataset is characterized by an overall region of regularity and flatness, with prominent feature class characteristics. In addition, a unique data preprocessing method was used during the imaging to color the lower vegetation areas, which were originally difficult to identify, red, thus making them

 doi:10.20944/preprints202406.0053.v1

easy to identify. This resulted in very high overall recognition accuracy for the Vaihingen dataset, with only slight boundary blurring between the predicted and labeled maps. With the use of the deep edge error correction method, the boundary information can be effectively constrained and the error signal is back-propagated through the supervising Encoder to radically improve the model's focus on edges. However, the relative Potsdam dataset has a more complex urban context and lacks the same pre-processing coloring operations as the Vahingen dataset, leading to increased difficulty in feature recognition by the model, which in turn makes the use of the deep edge error correction method less effective. Therefore, the method has some limitations. When the model's recognition accuracy of the image is high and there is only slight boundary-blurring, the method can effectively improve the edge quality. However, if the model's ability to characterize the image is insufficient, the boundary features cannot be significantly optimized by additional constraints. This result further demonstrates that boundary constraints and feature recognition ability interact with each other, and that boundary optimization alone cannot be performed without enhancing the model's recognition ability.

5.4. Research Difficulties and Next Steps

The extraction of generic cropland from high-resolution remote sensing images is a challenging task, facing difficulties in many aspects. Firstly, although the recognition of cropland extent and boundary awareness can be improved to a certain extent by enhancing edge awareness, facing the complex scene of high-resolution imagery makes the information of a single image insufficient to accurately extract the generic cropland parcels, because a large number of features similar to cropland, such as buildings, trees, etc., may be present in the imagery, which increases the difficulty of recognizing cropland. Second, the lack of spectral information makes it difficult to distinguish cropland from other features, especially when they are spectrally similar. Data inconsistency also limits the model's ability to generalize across time points and regions, making it challenging to extract cropland information accurately. To solve these problems, the next step is to concentrate efforts toward multimodal fusion. Multi-source data fusion can provide richer information, and improved deep learning methods can also effectively improve the accurate extraction of cropland. The combined use of these methods and techniques can effectively address the challenges of extracting cropland information from high-resolution imagery and provide reliable support for precision agriculture and land management.

6. Conclusions

This paper proposes a UNet-like generic extraction model named BAFormer for the boundary inaccuracy problem of cropland extraction from high-resolution remote sensing images. To optimize the boundary of complex forms, the BAFormer model compensation edge features in three stages: feature extraction, feature fusion, and loss constraint. Specifically, the model is designed with a channel fusion-based Feature Adaptive Mixer (FAM) and a large kernel sensory field-based DWLK-MLP module, which are used to enhance the model's information flow and expressive capability, enabling the model to recognize more complete and accurate boundaries. The Relational Adaptive Fusion (RAF) strategy and the Boundary Constraints that guide the model to pay attention to the boundaries automatically are also introduced to enhance the strong constraints of the model on the boundaries. To validate the generalized extraction performance of the proposed method, we conduct extensive experimental demonstrations on the Vaihingen, Potsdam, LoveDA, and Mapcup datasets. The results show that the method achieves remarkable results in several aspects such as model size, generalization ability, and edge quality, surpassing other current excellent networks and fully proving the effectiveness.

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

 doi:10.20944/preprints202406.0053.v1

Funding: This research was funded by Research on intelligent monitoring and early warning technology for rice pests and diseases of the Sichuan Provincial Department of Science and Technology, grant number 2022NSFSC0172; Sichuan Agricultural University Innovation Training Programme Project Funding, grant number 202210626054.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Acknowledgments: The authors thank the anonymous reviewers for the helpful comments that improved this manuscript.

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

Sample Availability: Code can be obtained at: https://github.com/WangYouM1999/BAFormer (accessed on 6 June 2024).

References

1. Zhong, H.F.; Sun, Q.; Sun, H.M.; Jia, R.S. NT-Net: A Semantic Segmentation Network for Extracting Lake Water Bodies From Optical Remote Sensing Images Based on Transformer. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–13.
2. Dong, X.; Xie, J.; Tu, K.; Qi, K.; Yang, C.; Zhai, H. DSFNet: Dual-Stream-Fusion Network for Farmland Parcel Mapping in High-Resolution Satellite Images. 2023 11th International Conference on Agro-Geoinformatics (Agro-Geoinformatics); IEEE: Wuhan, China, 2023; pp. 1–6.
3. Shunying, W.; Ya'nan, Z.; Xianzeng, Y.; Li, F.; Tianjun, W.; Jiancheng, L. BSNet: Boundary-semantic-fusion Network for Farmland Parcel Mapping in High-Resolution Satellite Images. *Computers and Electronics in Agriculture* **2023**, *206*, 107683.
4. Xie, D.; Xu, H.; Xiong, X.; Liu, M.; Hu, H.; Xiong, M.; Liu, L. Cropland Extraction in Southern China from Very High-Resolution Images Based on Deep Learning. *Remote Sensing* **2023**, *15*, 2231.
5. Xu, Y.; Zhu, Z.; Guo, M.; Huang, Y. Multiscale Edge-Guided Network for Accurate Cultivated Land Parcel Boundary Extraction From Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *62*, 1–20.
6. Li, M.; Long, J.; Stein, A.; Wang, X. Using a Semantic Edge-Aware Multi-Task Neural Network to Delineate Agricultural Parcels from Remote Sensing Images. *ISPRS Journal of Photogrammetry and Remote Sensing* **2023**, *200*, 24–40.
7. Zuo, R.; Zhang, G.; Zhang, R.; Jia, X. A Deformable Attention Network for High-Resolution Remote Sensing Images Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–14.
8. Yan, R.; Yan, L.; Geng, G.; Cao, Y.; Zhou, P.; Meng, Y. ASNet: Adaptive Semantic Network Based on Transformer–CNN for Salient Object Detection in Optical Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *62*, 1–16.
9. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–15.
10. 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 Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–20.
11. Long, J.; Li, M.; Wang, X.; Stein, A. Delineation of Agricultural Fields Using Multi-Task BsiNet from High-Resolution Satellite Images. *International Journal of Applied Earth Observation and Geoinformation* **2022**, *112*, 102871.
12. Xia, L.; Luo, J.; Sun, Y.; Yang, H. Deep Extraction of Cropland Parcels from Very High-Resolution Remotely Sensed Imagery. 2018 7th International Conference on Agro-geoinformatics (Agro-geoinformatics); IEEE: Hangzhou, 2018; pp. 1–5.
13. Xie, Y.; Zheng, S.; Wang, H.; Qiu, Y.; Lin, X.; Shi, Q. Edge Detection With Direction Guided Postprocessing for Farmland Parcel Extraction. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2023**, *16*, 3760–3770.

 doi:10.20944/preprints202406.0053.v1

14. Awad, B.; Erer, I. FAUNet: Frequency Attention U-Net for Parcel Boundary Delineation in Satellite Images. *Remote Sensing* **2023**, *15*, 5123.
15. Hao, S.; Wang, W.; Ye, Y.; Nie, T.; Bruzzone, L. Two-Stream Deep Architecture for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* **2018**, *56*, 2349–2361.
16. Doersch, C.; Gupta, A.; Efros, A.A. Unsupervised Visual Representation Learning by Context Prediction. *ICCV* **2015**, p. 9.
17. Zhang, W.; Guo, S.; Zhang, P.; Xia, Z.; Zhang, X.; Lin, C.; Tang, P.; Fang, H.; Du, P. A Novel Knowledge-Driven Automated Solution for High-Resolution Cropland Extraction by Cross-Scale Sample Transfer. *IEEE Transactions on Geoscience and Remote Sensing* **2023**, *61*, 1–16.
18. Iizuka, R.; Xia, J.; Yokoya, N. Frequency-Based Optimal Style Mix for Domain Generalization in Semantic Segmentation of Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *62*, 1–14.
19. Zhang, L.; Tan, Z.; Zhang, G.; Zhang, W.; Li, Z. Learn More and Learn Usefully: Truncation Compensation Network for Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *62*, 1–14.
20. Xu, L.; Ming, D.; Zhou, W.; Bao, H.; Chen, Y.; Ling, X. Farmland Extraction from High Spatial Resolution Remote Sensing Images Based on Stratified Scale Pre-Estimation. *Remote Sensing* **2019**, *11*, 108.
21. Li, Z.; Chen, S.; Meng, X.; Zhu, R.; Lu, J.; Cao, L.; Lu, P. Full Convolution Neural Network Combined with Contextual Feature Representation for Cropland Extraction from High-Resolution Remote Sensing Images. *Remote Sensing* **2022**, *14*, 2157.
22. Sheng, J.; Sun, Y.; Huang, H.; Xu, W.; Pei, H.; Zhang, W.; Wu, X. HBRNet: Boundary Enhancement Segmentation Network for Cropland Extraction in High-Resolution Remote Sensing Images. *Agriculture* **2022**, *12*, 1284.
23. Luo, W.; Zhang, C.; Li, Y.; Yan, Y. MLGNet: Multi-Task Learning Network with Attention-Guided Mechanism for Segmenting Agricultural Fields. *Remote Sensing* **2023**, *15*, 3934.
24. Shen, Q.; Deng, H.; Wen, X.; Chen, Z.; Xu, H. Statistical Texture Learning Method for Monitoring Abandoned Suburban Cropland Based on High-Resolution Remote Sensing and Deep Learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2023**, *16*, 3060–3069.
25. Yan, S.; Yao, X.; Sun, J.; Huang, W.; Yang, L.; Zhang, C.; Gao, B.; Yang, J.; Yun, W.; Zhu, D. TSANet: A Deep Learning Framework for the Delineation of Agricultural Fields Utilizing Satellite Image Time Series. *Computers and Electronics in Agriculture* **2024**, *220*, 108902.
26. Zhang, D.; Pan, Y.; Zhang, J.; Hu, T.; Zhao, J.; Li, N.; Chen, Q. A Generalized Approach Based on Convolutional Neural Networks for Large Area Cropland Mapping at Very High Resolution. *Remote Sensing of Environment* **2020**, *247*, 111912.
27. Pan, Y.; Wang, X.; Wang, Y.; Zhong, Y. RBP-MTL: Agricultural Parcel Vectorization via Region-Boundary-Parcel Decoupled Multitask Learning. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *62*, 1–15.
28. Wang, C.; Zhang, Y.; Cui, M.; Ren, P.; Yang, Y.; Xie, X.; Hua, X.S.; Bao, H.; Xu, W. Active Boundary Loss for Semantic Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence* **2022**, *36*, 2397–2405.
29. Kervadec, H.; Bouchtiba, J.; Desrosiers, C.; Granger, E.; Dolz, J.; Ben Ayed, I. Boundary Loss for Highly Unbalanced Segmentation. *Medical Image Analysis* **2021**, *67*, 101851.
30. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* **2015**.
31. Jin, Q.; Meng, Z.; Pham, T.D.; Chen, Q.; Wei, L.; Su, R. DUNet: A Deformable Network for Retinal Vessel Segmentation. *Knowledge-Based Systems* **2019**, *178*, 149–162.
32. Wang, Y.; Gu, L.; Jiang, T.; Gao, F. MDE-UNet: A Multitask Deformable UNet Combined Enhancement Network for Farmland Boundary Segmentation. *IEEE Geoscience and Remote Sensing Letters* **2023**, *20*, 1–5.
33. Li, Z.; Zheng, Y.; Shan, D.; Yang, S.; Li, Q.; Wang, B.; Zhang, Y.; Hong, Q.; Shen, D. ScribFormer: Transformer Makes CNN Work Better for Scribble-based Medical Image Segmentation. *IEEE Transactions on Medical Imaging* **2024**, pp. 1–1.
34. Pham, T.H.; Li, X.; Nguyen, K.D. SeUNet-Trans: A Simple yet Effective UNet-Transformer Model for Medical Image Segmentation, 2023, [arxiv:cs, eess/2310.09998].
35. Wu, D.; Guo, Z.; Li, A.; Yu, C.; Gao, C.; Sang, N. Conditional Boundary Loss for Semantic Segmentation. *IEEE Transactions on Image Processing* **2023**, *32*, 3717–3731.

 doi:10.20944/preprints202406.0053.v1

36. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM* **2017**, *60*, 84–90.
37. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; ukasz Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017, Vol. 30.
38. Li, J.; Xia, X.; Li, W.; Li, H.; Wang, X.; Xiao, X.; Wang, R.; Zheng, M.; Pan, X. Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios. *arXiv preprint arXiv:2207.05501* **2022**.
39. Tan, W.; Geng, Y.; Xie, X. FMViT: A multiple-frequency mixing Vision Transformer. *arXiv preprint arXiv:2311.05707* **2023**.
40. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
41. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* **2017**, *40*, 834–848.
42. Li, R.; Zheng, S.; Duan, C.; Su, J.; Zhang, C. Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters* **2021**, *19*, 1–5.
43. 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 journal of photogrammetry and remote sensing* **2021**, *181*, 84–98.
44. 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 Sensing* **2021**, *13*, 3065.
45. 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 Journal of Photogrammetry and Remote Sensing* **2022**, *190*, 196–214.
46. 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 Transactions on Geoscience and Remote Sensing* **2021**, *60*, 1–13.
47. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 1290–1299.
48. 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 Geoscience and Remote Sensing Letters* **2022**, *19*, 1–5.
49. Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Transformer-based decoder designs for semantic segmentation on remotely sensed images. *Remote Sensing* **2021**, *13*, 5100.
50. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition, 2015, [arxiv:cs/1512.03385].
51. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
52. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 510–519.
53. Kirillov, A.; Girshick, R.; He, K.; Dollár, P. Panoptic feature pyramid networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6399–6408.
54. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 4096–4105.
55. Chen, K.; Zou, Z.; Shi, Z. Building extraction from remote sensing images with sparse token transformers. *Remote Sensing* **2021**, *13*, 4441.
56. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
57. Cheng, B.; Schwing, A.; Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems* **2021**, *34*, 17864–17875.

 doi:10.20944/preprints202406.0053.v1

58. Zhang, X.; Gong, Y.; Li, Z.; Gao, X.; Jin, D.; Li, J.; Liu, H. SkipcrossNets: Adaptive Skip-cross Fusion for Road Detection. *arXiv preprint arXiv:2308.12863* **2023**.
59. Guo, S.; Liu, L.; Gan, Z.; Wang, Y.; Zhang, W.; Wang, C.; Jiang, G.; Zhang, W.; Yi, R.; Ma, L.; others. Isdnet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4361–4370.
60. Guo, M.H.; Lu, C.Z.; Liu, Z.N.; Cheng, M.M.; Hu, S.M. Visual attention network. *Computational Visual Media* **2023**, *9*, 733–752.
61. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems* **2021**, *34*, 3965–3977.
62. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. Cmt: Convolutional neural networks meet vision transformers. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 12175–12185.
63. Shi, D. TransNeXt: Robust Foveal Visual Perception for Vision Transformers, 2023, [arxiv:cs/2311.17132].
64. Li, R.; Wang, L.; Zhang, C.; Duan, C.; Zheng, S. A2-FPN for semantic segmentation of fine-resolution remotely sensed images. *International journal of remote sensing* **2022**, *43*, 1131–1155.
65. He, W.; Li, J.; Cao, W.; Zhang, L.; Zhang, H. Building extraction from remote sensing images via an uncertainty-aware network. *arXiv preprint arXiv:2307.12309* **2023**.
66. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation, 2022, [2110.08733].
67. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. Proceedings of the European conference on computer vision (ECCV), 2018, pp. 325–341.
68. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 7262–7272.
69. Srinivas, A.; Lin, T.Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck transformers for visual recognition. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 16519–16529.
70. Hu, P.; Perazzi, F.; Heilbron, F.C.; Wang, O.; Lin, Z.; Saenko, K.; Sclaroff, S. Real-time semantic segmentation with fast attention. *IEEE Robotics and Automation Letters* **2020**, *6*, 263–270.
71. Zhuang, J.; Yang, J.; Gu, L.; Dvornek, N. Shelfnet for fast semantic segmentation. Proceedings of the IEEE/CVF international conference on computer vision workshops, 2019, pp. 0–0.
72. Oršić, M.; Šegvić, S. Efficient semantic segmentation with pyramidal fusion. *Pattern Recognition* **2021**, *110*, 107611.