remote sensing

MDPI

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 and Kunhong Li *

College of Information Engineering, Sichuan Agricultural University, Ya'an 625014, China; lzy@sicau.edu.cn (Z.L.); 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

† 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 is highly heterogeneous and fragmented, and existing methods often suffer from inaccurate boundary 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 Depthwise Large Kernel Multi-Layer Perceptron model (DWLK-MLP) to enrich the global and local cropland boundaries features separately. Specifically, FAM enhances the boundary-aware method by adaptively acquiring high-frequency features through convolution and self-attention advantages, while DWLK-MLP further supplements boundary position information using a large receptive field. The efficacy of BAFormer has been evaluated on datasets including Vaihingen, Potsdam, LoveDA, and Mapcup. It demonstrates high performance, achieving mIoU scores of 84.5%, 87.3%, 53.5%, and 83.1% on these datasets, respectively. Notably, BAFormer-T (lightweight model) surpasses other lightweight models on the Vaihingen dataset with scores of 91.3% F1 and 84.1% mIoU.

Keywords: high-resolution remote sensing; convolutional neural network; vision transformer

Citation: Li, Z.; Wang, Y.; Tian, F.; Zhang, J.; Chen, Y.; Li, K. BAFormer: A Novel Boundary-Aware Compensation UNet-like Transformer for High-Resolution Cropland Extraction. *Remote Sens.* **2024**, *16*, 2526. https://doi.org/10.3390/rs16142526

Academic Editors: Guangliang Cheng, Qi Zhao, Paolo Tripicchio and Hossein M. Rizeei

Received: 20 May 2024
Revised: 5 July 2024
Accepted: 6 July 2024
Published: 10 July 2024

1. Introduction

With the rapid development of remote sensing technology, finer and higher-resolution optical remote sensing images can now be obtained [1]. Extracting cropland information from these images is crucial for assessing food security and formulating agricultural policies [2]. The mainstream approach involves using deep learning models for cropland identification [3]. Although deep learning has achieved some results in cropland data segmentation, the segmentation of cropland boundaries is still problematic due to the highly heterogeneous and fragmented nature of cropland [4]. Specifically, the inaccuracy of boundary segmentation is often caused by the misidentification of complex boundary shapes and features, as shown in Figure 1. Addressing these inaccuracies requires urgently enhancing the model's capability to perceive edge features, thereby improving the accuracy and reliability of segmentation results [5].

In recent years, many studies have proposed integrating semantic segmentation with edge detection to better guide models in perceiving agricultural land information, thereby enhancing local segmentation accuracy and preserving global morphological continuity. Existing methods can be broadly categorized into three types: (1) Network-based approaches [6–15]: These methods design specific network architectures based on agricultural features to direct the model's attention to key characteristics. However, they often emphasize specific geographical regions or single types of agricultural land, neglecting regional variations between plots and failing to achieve universal applicability. (2) Feature-based

approaches [16–24]: By augmenting the model with additional features, these approaches enhance the representation and understanding of agricultural information. However, some redundant feature representations not only increase computational burden but also do not yield positive effects on the model. (3) Loss-based approaches [5,25–27]: These methods introduce additional supervised training during the training process to impose strong constraints on boundaries. They strengthen segmentation constraints and optimize boundary continuity. However, these methods often classify boundary pixels and internal pixels into different categories, thereby to some extent compromising the consistency of identical boundary pixels and the inter-class differences of different boundary pixels.

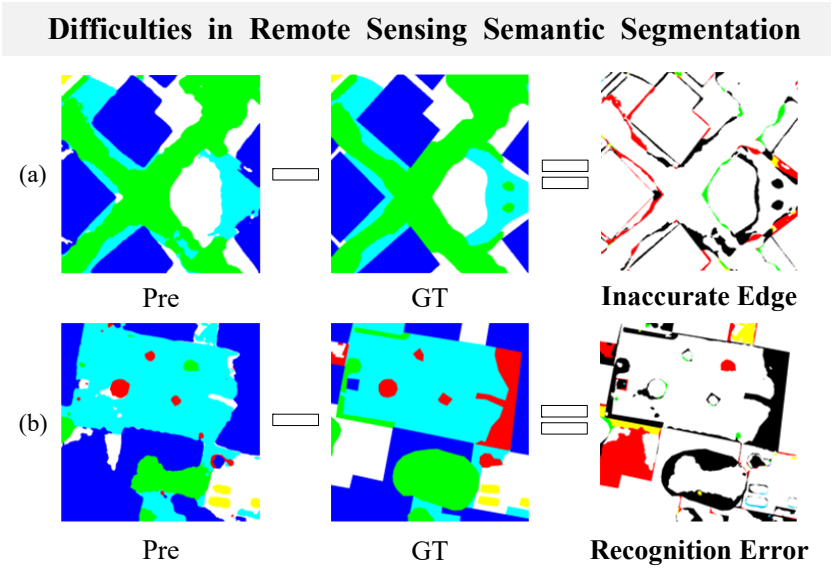


Figure 1. (**a**) Inaccurate edge issues. (**b**) Feature recognition error problems. Illustrates the differences between Ground Truth (GT) and model Predictions (Pre) obtained from the Vaihingen and Potsdam datasets, using prediction maps generated by UNetFormer.

To alleviate the issue of inaccurate boundary segmentation, we propose a UNet-like boundary-aware compensation model called BAFormer. Unlike explicit boundary detection methods, we introduce an implicit boundary-aware approach that enhances semantic contextual information while perceiving boundary features, comprehensively compensating in aspects of feature extraction, fusion, and constraint. (1) In feature extraction, we introduce the Feature Adaptive Mixer (FAM) and Depthwise Large Kernel Multi-Layer Perceptron (DWLK-MLP), significantly enhancing model information flow and expressive capability. FAM leverages the advantages of convolution and self-attention to separate high-frequency and low-frequency features of images, effectively extracting image details and global information while adaptively integrating frequency-based contributions. DWLK-MLP enlarges the convolutional receptive field through depth-separable large kernel convolutions, extracting more complete boundary features with minimal computational cost. (2) In feature fusion, we propose a Relational Adaptive Feature (RAF) fusion strategy based on spatial and channel semantic relationship perception. Unlike other static feature fusion methods, this approach dynamically learns weights by modeling spatial and channel relationships between shallow and deep feature maps. (3) In boundary constraint, we propose an edge constraint strategy implemented in deep layers of the network. This strategy guides the model to optimize boundaries from the bottom-up by extracting high-level semantic information from images, without requiring additional auxiliary task overhead.

Overall, our main contributions are summarised as follows:

1. We propose the BAFormer framework for edge optimization. The framework comprehensively improves the quality of edge segmentation of the model in terms of feature extraction, fusion, and constraints.

2. We propose a Feature Adaptive Mixer (FAM), which adaptively extracts high-frequency information represented by edges through the advantages of convolution and self-attention, thereby enhancing the model's information flow and expressive capability.
3. We propose a Depthwise Large Kernel Multi-Layer Perceptron (DWLK-MLP). The boundary features are enriched with negligible computational overhead by deeply decomposing the large kernel convolution.
4. We propose a Relational Adaptive Fusion (RAF) strategy, which optimizes feature granularity by dynamically sensing the relationships between features from both spatial and channel semantic perspectives.
5. We propose a deeply supervised edge constraint strategy. The boundary continuity is strengthened by making the model automatically focus on the boundary through deep semantic guidance.

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 [28] and deformable convolution [6], which show excellent performance in complex monitoring and segmentation tasks. These methods [10,15,29] can dynamically sense the geometric features of objects to adapt to morphologically changing structures. For example, the MDANet proposed in [6] was used to design a 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 [7–9,14,30,31], designed to adequately learn diverse target features are proposed. For example, the ASNet network proposed in ref. [7] innovatively integrates Transformer and CNN techniques in a two-branch encoder to capture global dependencies while capturing local fine-grained image features. In ref. [8], 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) Methods that combine edge detection and semantic segmentation tasks [4,10–12] are proposed to guide the model to strengthen the supervision constraint on the boundary. For example, the authors of [12] designed frequency attention to topically emphasize key high-frequency components in the feature map to improve the accuracy of boundary detection. The authors of [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 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 Feature Fusion

Feature fusion-based approaches [16–19,21–24,33] 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, ref. [16] 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 the loss of feature details due to image downsampling and the interference caused by image noise, Ref. [18] proposed to compensate for local image features and minimize noise by bootstrapping the feature extraction module. Ref. [20] 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. Ref. [21] proposed a boundary-enhanced segmentation network, HBRNet, with Swin-Transformer as the backbone of the pyramid hierarchy to obtain contextual information while enhancing boundary details. Ref. [23] proposed a pyramid scene parsing learning framework that combines high-level semantic feature extraction with low-level texture feature deep mining. Ref. [33] proposed to encode parcel features by a 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-mentioned methods, this paper proposes an adaptive fusion strategy based on the perception of spatial and channel semantic relations to dynamically adjust the fusion of shallow and deep features from spatial and channel perspectives, aiming to obtain finer-grained features.

2.3. Methods Based on Loss Function

Loss function-based methods [5,25–27,34], introduce a metric approach to complement strong constraints on morphological boundaries in the training process. SEANet [5] 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 [25] 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. Ref. [27] 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 [26] 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. Ref. [34] 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 dissimilar 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, leveraging rich high-level semantic information in the deep layers of the network to autonomously reinforce boundary constraints, without introducing additional overhead from auxiliary tasks.

3. Methodology

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 Depthwise Large Kernel Multi-Layer Perceptron (DWLK-MLP), the Relational Adaptive Fusion (RAF) strategy, and the deeply supervised edge constraint strategy. The model development is built upon the UNetFormer network [35], utilizing high-performance auxiliary branches and referring to the structural design of the encoder and decoder.

3.1. CNN-Based Encoder

In the BAFormer model (as shown in Figure 2), we adopt the encoder design of UNetFormer, using ResNet-18 at the encoder side as a shallow semantic extractor to effectively

capture shallow features and reduce computational costs. ResNet-18 consists of four residual blocks capable of extracting shallow (high-frequency) semantic features. During feature compression, the spatial resolution of each block is halved through downsampling, and the number of channels used for extracting deep semantic information is doubled. Within each block, skip connections are employed at the decoder side to link shallow semantic features with their corresponding semantic levels.

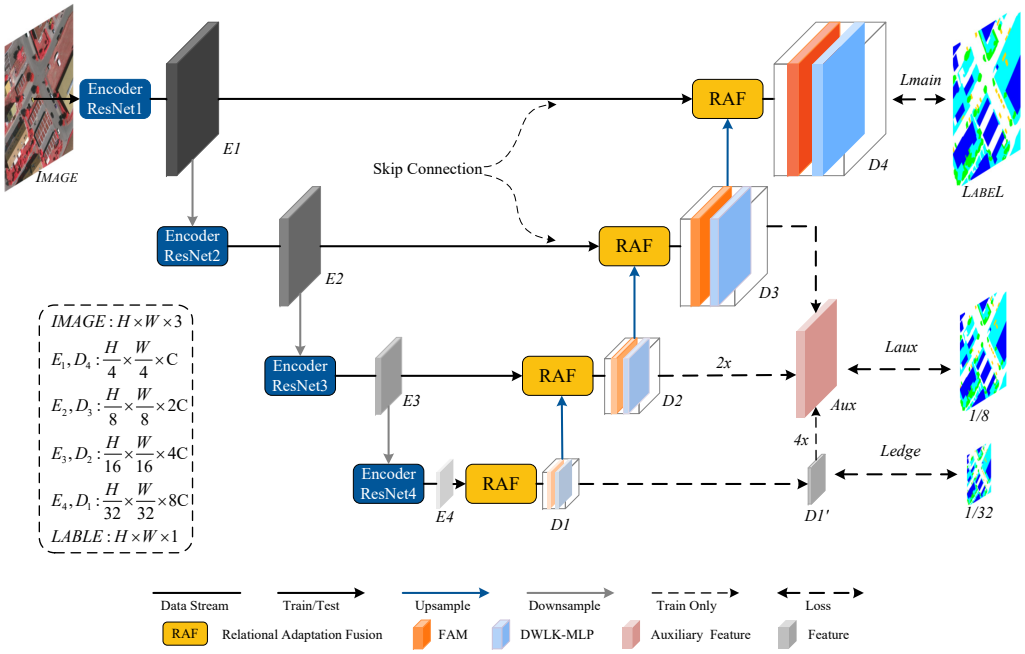


Figure 2. An overview of the BAFormer model.

3.2. Transformer-Based Decoder

The decoder adopts the same block design as UNetFormer, achieving abstract feature extraction and detail reconstruction of images by stacking four BABlock modules (as shown in Figure 3) from bottom to top. 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)

A Convolutional Neural Network (CNN) acts as a high-pass filter that can extract locally salient high-frequency information such as texture and detail [36]. The 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, determining 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 obtain 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].

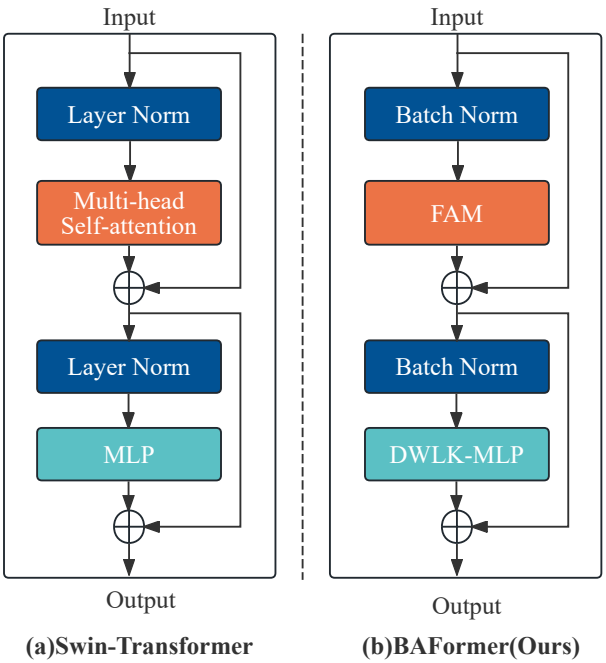


Figure 3. The structure of BABlock. (**a**) represents the block structure in Swin-Transformer, and (**b**) represents the BABlock structure in BAFormer.

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 1×1 and 3×3 regular convolutions [40]. To enhance the learning and generalization ability of self-attention, we designed it 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 $F_i \in \mathbb{R}^{C \times H \times W}$ denote the input feature map, with $H = W$ by default. After confirming the 2D feature map through identity, the size remains unchanged following standard convolutions of kernel sizes 1 and 3. The formulas for generating the high-frequency feature F_h and the high-frequency affinity matrix F_{hm} are as follows:

$$F_{c1} = \mathbb{C}_1(F_i),\ F_h = F_{c2} = \mathbb{C}_2(F_{c1}) \tag{1}$$

$$F_{hm} = \Phi(F_{c1} \otimes {F_{c2}}^T) \tag{2}$$

where $\otimes$ represents matrix multiplication, Φ represents the operation of partitioning according to a predefined window size N, T denotes matrix transpose, $\mathbb{C}_1$ represents a 1×1 convolutional operator, $\mathbb{C}_2$ represents a 3×3 convolutional operator, $F_{c1}, F_{c2}, F_h \in \mathbb{R}^{C \times H \times W}$, and $F_{hm} \in \mathbb{R}^{\left(\frac{H}{N} \times \frac{W}{N} \times C\right) \times N \times N}$.

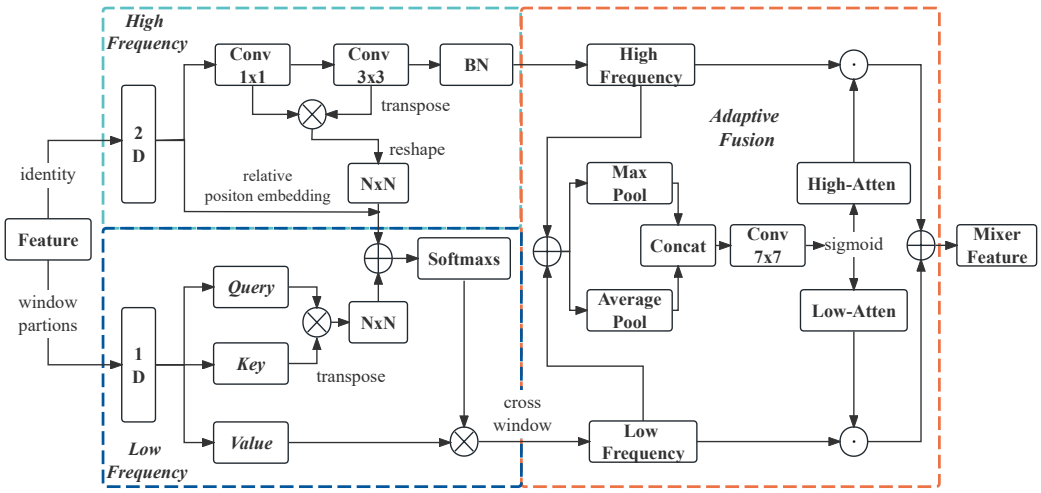


Figure 4. The structure of the Feature Adaptive Mixer (FAM) is as follows: 2D refers to a two-dimensional image, and 1D denotes a sequence stretched to one dimension. BN stands for Batch Normalization. High-Attn represents the attention weight score attributed to high-frequency features in the mixed information flow, while Low-Attn represents the attention weight score attributed to low-frequency features in the mixed information flow. Both dimensions are denoted as HxW.

The low-frequency branch plays a pivotal role in capturing global contextual relationships, primarily through a multi-head self-attention mechanism [41]. Initially, the method expands the input feature map $F_i \in \mathbb{R}^{C \times H \times W}$ by a factor of three along the channel dimension using standard 1×1 convolution. Subsequently, the 2D feature map is partitioned into windows of size $N \times N$ and flattened into a 1D sequence $\in \mathbb{R}^{3 \times \left(\frac{H}{N} \times \frac{W}{N} \times h\right) \times (N \times N) \times \frac{C}{h}}$ with adjusted dimensions considering the number of heads and channels, where N denotes the window size and h represents the number of heads. This sequence is then decomposed into Query (Q), Key (K), and Value (V) feature vectors $\in \mathbb{R}^{\left(\frac{H}{N} \times \frac{W}{N} \times h\right) \times (N \times N) \times \frac{C}{h}}$. During the self-attention computation, a learnable positional encoding (PE) is introduced to encode positional information of the image sequence. The resultant low-frequency affinity matrix F_{lm}, which is derived from multi-head self-attention, is then combined with the high-frequency affinity matrix F_{hm} to produce a blended affinity matrix F_{mm}. After applying softmax normalization to F_{mm}, a matrix multiplication with V produces the low-frequency feature map F_l. The formula is described as follows:

$$F_{lm} = Q \otimes K^T \tag{3}$$

$$F_{mm} = F_{hm} \oplus F_{lm} \tag{4}$$

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

where $\oplus$ denotes element-wise addition, $Softmax$ represents the normalization activation function, $F_{lm}, F_{mm} \in \mathbb{R}^{\left(\frac{H}{N} \times \frac{W}{N} \times C\right) \times N \times N}$, where N is the window size, PE is the learnable positional encoding of window size, d is a constant, and $F_l \in \mathbb{R}^{C \times H \times W}$.

High-low frequency adaptive fusion is a fusion mechanism built on spatial feature mapping. Inspired by the feature rescaling of SK-Net [42], 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

select 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 fused together to 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 $A_h \in \mathbb{R}^{H\times W}$ and low-frequency attention feature map $A_l \in \mathbb{R}^{H\times W}$, 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 F_h and 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:

$$F_m = F_h \oplus F_l \tag{6}$$

$$A_h = MaxPool(F_m),\ A_l = AvgPool(F_m) \tag{7}$$

$$A = F_{Conv}^{7\times 7}(Concat(A_h, A_l)) \tag{8}$$

$$\hat{A}_h, \hat{A}_l = Sigmoid(A, dim = 0) \tag{9}$$

$$F_o = \left(F_h \odot \hat{A}_h\right) \oplus \left(F_l \odot \hat{A}_l\right) \tag{10}$$

where $\odot$ represents matrix element-wise multiplication, *MaxPool* denotes global maximum pooling, *AvgPool* denotes global average pooling, *Concat* denotes channel-level splicing, *Sigmoid* denotes the activation function, and $F_{Conv}^{7\times 7}$ denotes convolution with a kernel size of 7×7.

3.2.2. RAF (Relational Adaptive Fusion)

To obtain richer boundary features, fusing feature maps of different scales is considered to be an effective method to improve image effects [43]. 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. Determining how to improve the effect of feature fusion has become a new thinking direction to optimize network performance. Inspired by the perceptual fusion of shallow and deep branches in ISDNet [44], 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, this method first models the spatial numerical differences between shallow-layer features and deep-layer features through global average pooling to learn spatial weighting factors. Then, matrix multiplication is performed under the feature mapping of spatial modeling to obtain the channel relationship matrix. By flattening the relationship matrix and compressing the features, channel weighting factors are obtained. Finally, the spatial weighting factors and channel weighting factors obtained are separately weighted and fused.

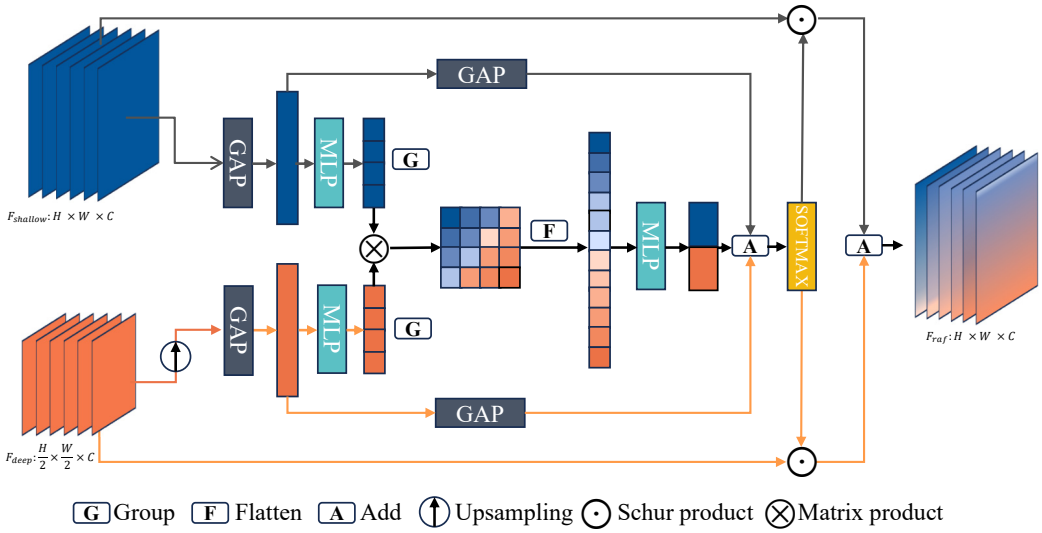


Figure 5. Illustration of the Relational Adaptive Fusion (RAF) module. GAP stands for Global Average Pooling and MLP stands for Multi-Layer Perceptron variation. Blue and orange represent the feature maps of the shallow and deep layers of the network, respectively.

Given the shallow feature map $F_s \in \mathbb{R}^{C \times H_s \times W_s}$ and the deep feature map $F_d \in \mathbb{R}^{C \times H_d \times W_d}$ where $H_s \neq H_d$ and $W_s \neq W_d$. In the first step, RAF aligns the height and width of the deep feature map with those of the shallow feature map. By explicitly extracting feature information, two one-dimensional attention vectors P_s and $P_d \in \mathbb{R}^C$ containing their respective channel information are obtained. The following formulas can represent this:

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

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

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

When modeling channel dependencies, considering the semantic differences between channels, the two one-dimensional attention vectors P_s and P_d are compressed to a length of r using perceptrons to reduce semantic errors, resulting in two contraction vectors P_{sr} and $P_{dr} \in \mathbb{R}^r$, where r is typically much smaller than C. Subsequently, based on these two contraction vectors, a channel correlation matrix $R \in \mathbb{R}^{r \times r}$ is obtained through matrix multiplication. This correlation matrix is then flattened and mapped through multiple perceptron layers to generate a channel weight factor consisting of only two numerical values, C_{ws} and C_{wd}, as shown in the following formula:

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

where $\otimes$ denotes matrix multiplication, R represents the channel relationship matrix, $FLATTEN$ denotes the flattening operation, and \$ denotes a Multi-Layer Perceptron that maps a one-dimensional vector to two channel weighting factors. In the third step, the obtained weight values are separately weighted and fused. The spatial weight factors S_{ws} and S_{wd} from the shallow feature map F_s and deep feature map F_d, along with the channel weight factors C_{ws} and C_{wd}, are summed individually. After applying a softmax operation, they yield weighted values W_s and W_d. These are then dot-multiplied with F_s and F_d,

respectively, and added together to form the final fused feature map $F_{raf} \in \mathbb{R}^{C \times H_s \times W_s}$. The formula is as follows:

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

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

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

Enhancing the convolutional perceptual field is an effective means to improve semantic segmentation [45]. Recent studies have shown that the introduction of DW convolution into MLP (Multi-Layer Perceptron) can effectively integrate the properties of self-attention and convolution, thus enhancing the generalization ability of the model [46]. Compared to ordinary MLP [41], DW-MLP [47] with a residual structure introduces a 3 × 3-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 multi-scale feedforward neural network MS-MLP [48] has been proposed. It 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 transform the multi-scale features further 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 depthwise large kernel convolution of 23 × 23 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 5 × 5 and 7 × 7 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 depthwise large kernel convolution before the activation function improves the accuracy and robustness of image recognition more than after the activation function.

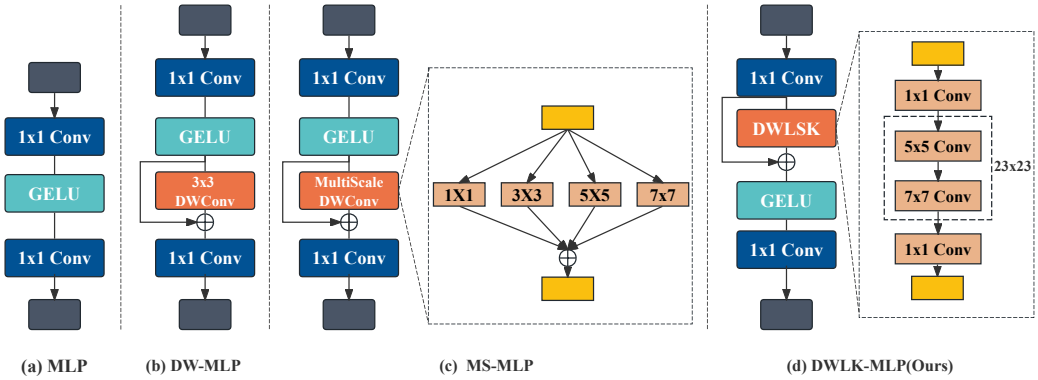


Figure 6. (**a**) Plain MLP that processes only cross-channel information. (**b**) Depthwise residuals for aggregating local tokens, DW-MLP. (**c**) Depthwise residuals for aggregating multi-scale tokens, MS-MLP. (**d**) Our proposed depthwise large kernel, DWLK-MLP.

3.2.4. Edge Constraint with Deep Supervision

The deeper the network layers, the richer the high-level semantic information [49]. 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 perform the supervision. The method significantly improves 0.5% mIoU on the Vaihingen dataset, enhancing edge supervision effectively with negligible computational overhead and no increase in model parameters. The deep edge constraint can be formulated as follows:

$$y = \delta(F_{\mathrm{df}}) \tag{16}$$

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

where F_{df} is the deepest feature map of the network, δ denotes the number of channels corresponding to the number of classes of this feature map, and $\mathcal{L}_{\mathrm{edge}}$ refers to the cross-entropy constraint with the labels.

3.3. Loss Function

The model primarily utilizes Cross-Entropy Loss ($\mathcal{L}_{\mathrm{ce}}$) and Dice Loss ($\mathcal{L}_{\mathrm{dice}}$). The main segmentation loss $\mathcal{L}_{\mathrm{main}}$ integrates these two in a joint form. To enhance segmentation performance for multi-scale targets, we introduce Cross-Entropy Loss as an auxiliary constraint $\mathcal{L}_{\mathrm{aux}}$ in intermediate layers. Innovatively, we introduce deep supervision edge constraint $\mathcal{L}_{\mathrm{edge}}$ to guide the model's focus on boundaries from deep layers. Importantly, $\mathcal{L}_{\mathrm{aux}}$ and $\mathcal{L}_{\mathrm{edge}}$ only operate during training to avoid inference speed impact. In summary, the total loss $\mathcal{L}$ of the model is the sum of $\mathcal{L}_{\mathrm{main}}$, $\mathcal{L}_{\mathrm{aux}}$, and $\mathcal{L}_{\mathrm{edge}}$, formulated as follows:

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

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

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

$$\mathcal{L} = \mathcal{L}_{\mathrm{main}} + \mathcal{L}_{\mathrm{aux}} + \mathcal{L}_{\mathrm{edge}} \tag{21}$$

where 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 extensive experiments to compare the performance of BAFormer with other advanced segmentation models, validating our model's effectiveness from multiple perspectives. Specifically, we evaluated the model on public datasets Vaihingen, Potsdam, and LoveDA, along with our custom Mapcup dataset, to verify its capability in general farmland extraction. Additionally, we performed numerous ablation studies to rigorously demonstrate the scientific foundation of our module's components and parameter settings. In comparison with other methods, accuracy indicators are typically obtained from cited

literature by default, and results obtained by ourselves are denoted with a # symbol, as specified in each table.

4.1. Experimental Setup

4.1.1. Dataset

Vaihingen: The dataset consists of 33 top-view image patches with a very fine spatial resolution, with an average size of 2494 × 2064 pixels. Each image patch contains three multi-spectral 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 utilized ID: 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, 38 for testing, ID: 30 for validation, and the remaining 15 images for training.

Potsdam: The dataset comprises 38 top-view image blocks with an 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 multi-spectral 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 ID: 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 image block 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.

LoveDA: The dataset [50] 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 m. The dataset covers seven 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 dataset for testing. These images were captured from urban and rural scenes in three cities in China (Nanjing, Changzhou, and Wuhan).

Mapcup: This dataset was annotated and provided by the Key Laboratory of Farmland Resources Monitoring and Protection of Sichuan Agricultural University. It includes a total of 507 high-resolution cropland images, each with a resolution of 1024 × 1024 pixels. The ground sampling distance of the images is 0.6 meters. The dataset covers two classes: cropland and non-cropland, accounting for 60.4% and 39.6% of the dataset, respectively. The dataset is divided into three parts: a training set, a validation set, and a test set. There are 373 images for training, and the remaining 134 are for validation and testing. Additionally, there is an unlabeled area of 21,433 × 27,976 pixels which, after cropping, resulted in 1160 images used for mapping inference. These images are from the Northern Plains region and are finely annotated. They exhibit significant characteristics such as complex scenes, multi-scale objects, and class imbalance, posing considerable challenges for tasks involving high-resolution remote sensing image farmland extraction.

4.1.2. Implementation Details

The experiment referenced the dataset processing and parameter settings of UNetFormer, conducted using the PyTorch framework on a single Nvidia GTX 3090 GPU. The encoder utilized pre-trained weights from the timm library's ResNet-18 to accelerate training through transfer learning. An AdamW optimizer with a base learning rate of 6×10^{-4} and weight decay of 0.01 was employed, complemented by a cosine learning rate scheduler for improving convergence speed.

During the data preprocessing stage, we handled each dataset separately. For the Vaihingen dataset, due to insufficient samples, we performed sliding window cropping of images into patches of 1024 × 1024 pixels with a sliding step of 512 pixels. For the Potsdam dataset, we also used 1024 × 1024 pixel patches with a sliding step of 1024 pixels. For the LoveDA dataset, we merged the training and validation datasets and treated them together as the training set. The Mapcup dataset, having undergone meticulous refinement, required no further processing. Additionally, to expedite model training, RGB labels were converted into the one-hot encoded format.

During model training, multiple data augmentation techniques were applied to the Vaihingen, Potsdam, LoveDA, and Mapcup datasets. These included Gaussian blur, random scaling factors [0.5, 0.75, 1.0, 1.25, 1.5], random horizontal and vertical flips, as well as random adjustments to brightness and contrast. Specific parameter settings referenced the experimental setup of UNetFormer. Vaihingen training was set to 105 epochs, and the batch size for training and validation was set to 4; Potsdam training was set to 45 epochs, and the batch size for training and validation was set to 4; LoveDA training was set to 35 epochs, and the batch size for training and validation process was set to 16; Mapcup was trained for 35 epochs, and the batch size of the training and validation process was set to 4. Moreover, all ablation and discussion experiments used identical configuration files and training strategies on the datasets to ensure fair and effective results. When comparing with other methods, default parameters of the original methods were adopted, with no changes to the data augmentation techniques.

4.1.3. Evaluation Indicators

The overall evaluation of the model follows the ISPRS benchmark. To assess the model's effectiveness, we employed IoU, OA, F1-score, and Boundary IoU [51]. Among these, IoU is the most common and important evaluation metric, providing an intuitive reflection of the model's ability to separate foreground and background. A higher IoU value indicates more accurate segmentation results. Here, we define TP as true positives, TN as true negatives, FP as false positives, and FN as false negatives. According to these definitions, the relevant formulas are expressed as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \ \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \ \text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{22}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \ \text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{23}$$

To measure the effectiveness of boundary segmentation, we introduced the Boundary IoU metric, which evaluates the quality of edge segmentation by measuring the intersection over union (IoU) between predicted and ground truth boundaries. Here, we define G as the ground truth binary mask, P as the predicted binary mask, G_d, and P_d as sets of pixels on the boundary of these masks. This is expressed with the following formula:

$$\text{Boundary IoU} = \frac{|(G_d \cap G) \cap (P_d \cap P)|}{|(G_d \cap G) \cup (P_d \cap P)|}. \tag{24}$$

4.2. Experimental Results

4.2.1. Results on Vaihingen Dataset

To validate the effectiveness of the proposed method, extensive comparative experiments were conducted. Quantitative comparisons were made on the ISPRS Vaihingen dataset against state-of-the-art models, as shown in Table 1. We compared classic semantic segmentation algorithms such as FCN [52] and DeepLabV3+ [53], as well as advanced CNN-encoder-based algorithms including MAResU-Net [54], ABCNet [55], BANet [56], UNetFormor [35], and MANet [57]. Additionally, we compared advanced large models based on Transformer-encoder, namely DC-Swin [58], Mask2Former [59], and FT-UNetformer [35].

The experimental results demonstrate that BAFormer achieved outstanding segmentation performance on the Vaihingen dataset, achieving 91.5% MeanF1, 91.8% OA, and 84.5% mIoU. Compared to advanced models using ResNet-18 blocks as encoders, BAFormer outperformed UNetFormer by 1.8% mIoU. Even when compared to large models based on Swin encoders, the proposed lightweight model BAFormer-T surpassed the performance of FT-UNetformer, yielding satisfactory results. Notably, the method achieved a score of 91.2% in the "car" class, surpassing most advanced models by approximately 1–2%. This improvement can be attributed to DWLK-MLP, which facilitates the fusion of convolution and self-attention, enhancing the capture of boundary information through a large receptive field, thereby improving the accuracy of small object recognition.

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

Method	Backbone	Per-Class F1 (%)					MeanF1 (%)	OA (%)	mIoU (%)
		Imp.surf	Building	Low.veg	Tree	Car			
FCN [52]	VGG-16	88.2	93.0	81.5	83.6	75.4	84.3	87.0	74.2
DeepLabv3+ [53]	ResNet-50	88.0	94.2	81.3	87.8	78.1	85.9	88.9	76.3
MAREsU-Net [54]	ResNet-18	92.0	95.0	83.7	89.3	78.3	87.7	89.7	78.6
ABCNet [55]	ResNet-18	92.7	95.2	84.5	89.7	85.3	89.5	90.7	81.3
BANet [56]	ResT-Lite	92.2	95.2	83.8	89.9	86.8	89.6	90.5	81.4
UNetFormer [35]	ResNet-18	92.7	95.3	84.9	90.6	88.5	90.4	91.0	82.7
MANet [57]	ResNet-50	93.0	95.5	84.6	90.0	89.0	90.4	91.0	82.7
Mask2Former [59]	Swin-B	92.9	94.5	85.3	90.4	88.5	90.3	90.8	83.0
DC-Swin [58]	Swin-S	93.6	**96.2**	**85.8**	90.4	87.6	90.7	91.6	83.2
FT-UNetFormer [35]	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**

The qualitative comparisons of the ISPRS Vaihingen test set are shown in Figure 7. We selected three representative instances to evaluate the effectiveness of our model, highlighting specific areas of interest with black dashed boxes. In the first row of results, three buildings with distinct shadows, similar colors, and textures to low-level vegetation exhibit significant inter-class similarity. Convolution-based semantic segmentation methods like BANet and UNetFormer misclassified this building area as low-level vegetation, whereas our model made accurate predictions. This demonstrates that our BAFormer method effectively learns global spatial context relationships, capturing inter-class differences beyond local color and texture features. In the second row of results, debris on an opaque foreground road shows notable texture and shape differences compared to surrounding roads, indicating substantial intra-class variability. Other models misclassified most of this debris area as "Imp.surf", whereas our BAFormer accurately extracted interactive representations between foreground and background for correct classification. This interaction highlights the importance of foreground-background class balance in remote sensing semantic segmentation, a critical issue previous methods have not adequately addressed. In the third row of results, within a complex area surrounded by low-level vegetation, sensor imaging, and image processing led to significant category distortions beyond human visual recognition. Our proposed method achieved results closest to ground truth labels, capturing more precise boundary details.

Overall, our BAFormer method effectively learns global contextual correlations, comprehensively extracts inter-class differences and intra-class variabilities, models dependencies between foreground and background, and generates segmentation maps with higher accuracy and richer details.

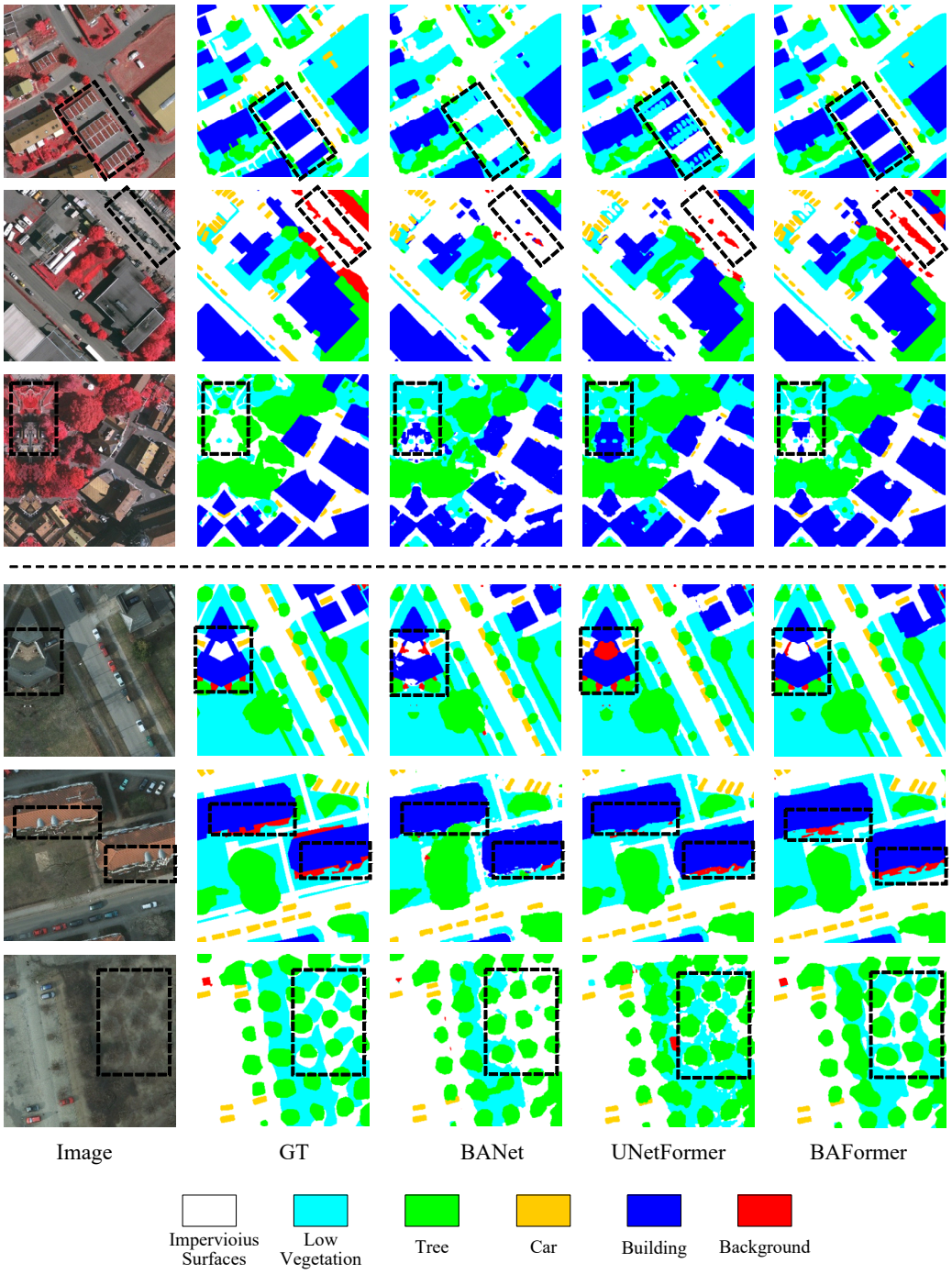


Figure 7. Qualitative comparisons under ISPRS Vaihaigen (**top**) and ISPRS Postdam (**bottom**) test sets. We add some black dotted boxes to highlight the differences to facilitate model comparisons.

4.2.2. Results on Potsdam Dataset

On the ISPRS Potsdam test set, we conducted quantitative comparisons with state-of-the-art models, as presented in Table 2. We evaluated prominent methods including FCN [52], DeepLabV3+ [53], MAResU-Net [54], ABCNet [55], BANet [56], UNetFormor [35], MANet [57], Mask2Former [59], SwinTF-FPN [60], and FT-UNetFormer [35]. Our proposed BAFormer achieved significant scores on this dataset: 87.3% mIoU, 93.2% F1, and 92.2% OA, surpassing the state-of-the-art Mask2Former and FT-UNetFormer models in large-scale semantic segmentation. Notably, it achieved the highest precision in easily discernible categories such as "Car" and " Building", surpassing other convolution-based advanced models by 1–2% in F1. These results underscore our method's superior feature recognition

capabilities for both small- and large-scale geographical features. Additionally, BAFormer also demonstrated the highest accuracy in the most challenging category "Low.veg", further validating its robust feature learning and representation abilities.

Table 2. The Potsdam dataset was quantitatively compared with existing methods. The best values in each column are shown in bold. '#' signifies results obtained by us, while other results are copied from the original paper.

Method	Backbone	Per-Class F1 (%)					MeanF1 (%)	OA (%)	mIoU (%)
		Imp.surf	Building	Low.veg	Tree	Car			
FCN [52]	VGG-16	88.5	89.9	78.3	85.4	88.8	86.2	86.6	78.5
DeepLabv3+ [53]	ResNet-50	90.4	90.7	80.2	86.8	90.4	87.7	87.9	80.6
MAREsU-Net [54]	ResNet-18	91.4	85.6	85.8	86.6	93.3	88.5	89.0	83.9
BANet [56]	ResT-Lite	93.3	95.7	87.4	89.1	96.0	92.3	91.0	85.3
ABCNet [55]	ResNet-18	93.5	95.9	87.9	89.1	95.8	92.4	91.3	85.5
SwinTF-FPN [60]	Swin-S	93.3	96.8	87.8	88.8	95.0	92.3	91.1	85.9
UNetFormer # [35]	ResNet-18	93.6	96.8	87.7	88.9	95.8	92.6	91.3	86.0
MANet [57]	ResNet-50	93.4	96.7	88.3	89.3	96.5	92.8	91.3	86.4
Mask2Former [59]	Swin-B	**98.0**	96.9	88.4	**90.7**	84.6	91.7	**92.5**	86.6
FT-UNetFormer # [35]	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**

Similarly, qualitative experimental comparisons were conducted on the ISPRS Potsdam test set, and localized visualization results are depicted in Figure 7. We selected three representative samples for qualitative analysis. As shown in the results of the fourth row, the central area of a building complex includes an opaque water surface (plaza), posing challenges in visually distinguishing between building structures, opaque water, or background. Other models either misclassify it as background or predict rough, blurry boundaries. In contrast, our proposed BAFormer achieved satisfactory segmentation results, closely approaching the ground truth. This outcome highlights our method's capability to understand long-range spatial contextual dependencies. In the fifth row of results, within the dashed box surrounding buildings, there is a ring of blurry noise, indicating intra-class abrupt changes in the background category. Other models erroneously classify such intra-class changes as part of the building structure, whereas only our BAFormer effectively identifies these anomalous intra-class samples and makes accurate predictions. The sixth row of results reveals discrete low-level vegetation and trees within the dashed box, showing significant yet subtle differences. BAFormer uniquely captures inter-class disparities, resulting in more precise edge detection.

In summary, BAFormer extracts richer global-local detail features, thereby achieving more accurate visual results across various scenarios.

4.2.3. Results on LoveDA Dataset

To further assess the effectiveness of the model, extensive comparative experiments were conducted on the LoveDA dataset, involving comprehensive evaluations in both quantitative and qualitative aspects. Quantitative results are presented in Table 3, and qualitative assessments are depicted in Figure 8. In our experiments, we compared our proposed BAFormer with state-of-the-art models including FCN [52], DeepLabV3+ [53], SemanticFPN [61], FarctSeg [62], TransUNet [63], BANet [56], UNetFormer [35], SwinUperNet [41], DC-Swin [58], and MaskFormer [64]. From the experimental results, BAFormer achieved the highest mIoU of 53.5%, surpassing the advanced UNetFormer model by 1.1% mIoU and outperforming the Transformer-encoder based MaskFormer by 2.7% mIoU. It demonstrated strong segmentation performance in prominent terrain features such as "Road", "Water", and "Forest" classes. However, recognition in complex and variable non-prominent terrain classes such as "Background" and "Barren" was less satisfactory. This observation indicates that our proposed method enhances learning and discrimination capabilities for prominent features but shows some limitations compared to MaskFormer in handling complex non-prominent features using mask-based processing.

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

Method	Backbone	Per-Class IoU (%)							mIoU (%)
		Background	Building	Road	Water	Barren	Forest	Agriculture	
FCN [52]	VGG-16	42.6	49.5	48.1	73.1	11.8	43.5	58.3	46.7
DeepLabv3+ [53]	ResNet-50	43.0	50.9	52.0	74.4	10.4	44.2	58.5	47.6
SemanticFPN [61]	ResNet-50	42.9	51.5	53.4	74.7	11.2	44.6	58.7	48.2
FarctSeg [62]	ResNet-50	42.6	53.6	52.8	76.9	16.2	42.9	57.5	48.9
TransUNet [63]	Vit-R50	43.3	56.1	53.7	78.0	9.3	44.9	56.9	48.9
BANet [56]	ResT-Lite	43.7	51.5	51.1	76.9	16.6	44.9	**62.5**	49.6
SwinUperNet [41]	Swin-Tiny	43.3	54.3	54.3	78.7	14.9	45.3	59.6	50.1
DC-Swin [58]	Swin-Tiny	41.3	54.5	56.2	78.1	14.5	47.2	62.4	50.6
MaskFormer [64]	Swin-Base	**52.5**	60.4	56.0	65.9	**27.7**	38.8	54.3	50.8
UNetFormer [35]	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**

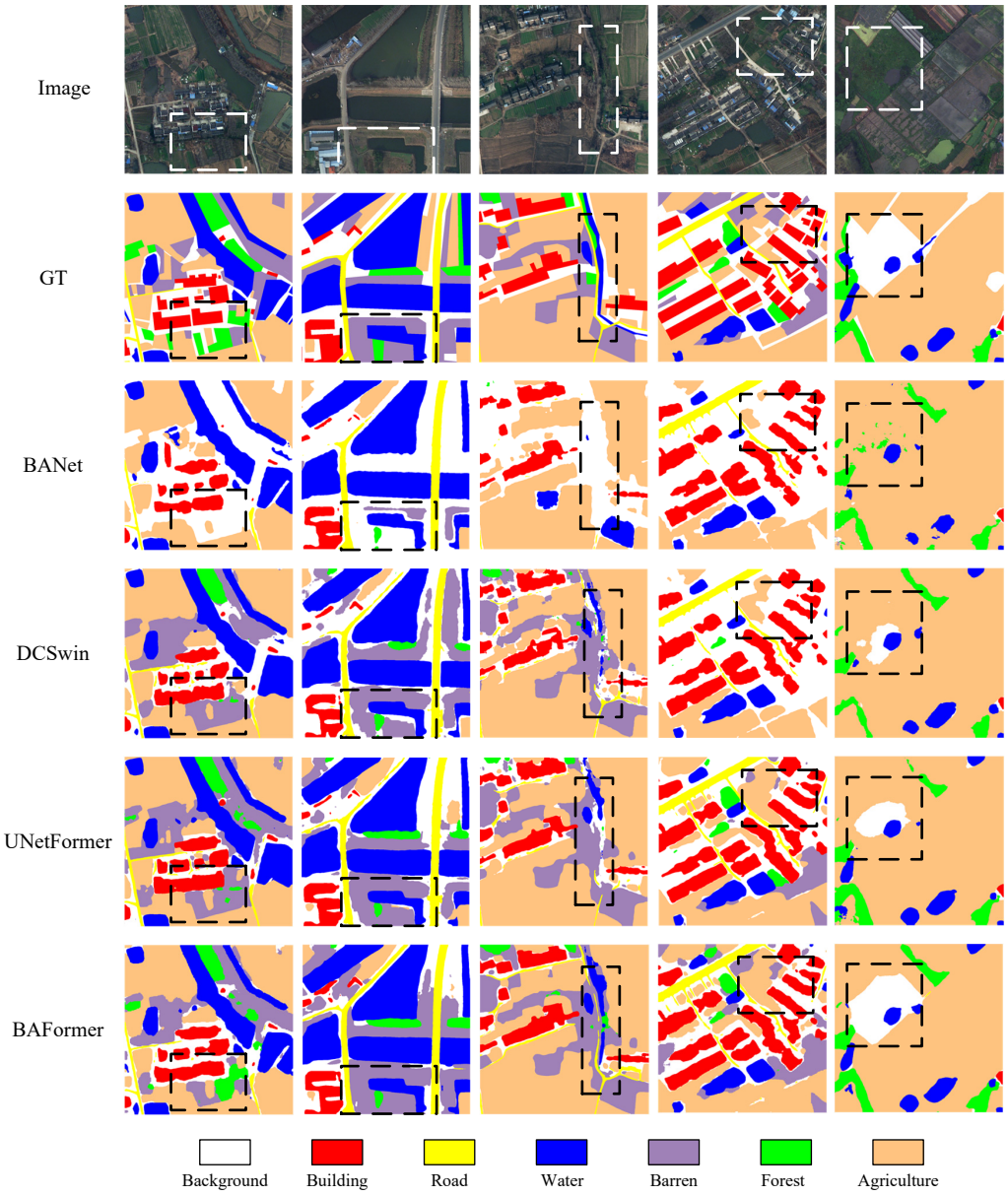


Figure 8. Qualitative comparisons with different methods on the LoveDA validation set. We add some dotted boxes to highlight the differences to facilitate model comparisons.

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 [52], DeepLabV3+ [53], A2FPN [65], ABCNet [55], MANet [57], BANet [56], DC-Swin [58], UNetFormer [35], and FT-UNetFormer [35]. From the experimental results, BAFormer achieves satisfactory results, obtaining 90.7% F1, 90.8% OA, and 83.1% mIoU, which are 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, BAFormer demonstrated significantly improved representation capability. Compared to the state-of-the-art UNetFormer model based on convolution, BAFormer outperformed by 2.5% mIoU. Furthermore, compared to the state-of-the-art FT-UNetFormer model based on Transformer, BAFormer exceeded 1.5% mIoU. 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.

Table 4. Comparison of different methods on the Mapcup dataset. The best values in each column are shown in bold. '#' signifies results obtained by ourselves.

Method	Backbone	Per-Class F1 (%)		MeanF1 (%)	OA (%)	Training Time (h)	Boundary IoU (%)	mIoU (%)
		Cropland	Non-Cropland					
FCN # [52]	VGG-16	81.6	86.8	84.2	84.7	0.9	44.8	72.5
DeepLabv3+ # [53]	ResNet-50	82.7	87.5	85.1	85.5	1.2	47.6	74.2
A2FPN # [65]	ResNet-18	83.2	87.8	85.5	85.9	0.4	48.0	74.8
ABCNet # [55]	ResNet-18	84.0	88.1	86.1	86.4	0.6	48.4	75.6
MANet # [57]	ResNet-50	86.0	89.3	87.7	87.7	0.7	51.3	78.1
BANet # [56]	ResT-Lite	86.7	89.8	88.3	88.5	1.0	53.9	79.0
DC-Swin # [58]	Swin-S	86.8	89.7	88.2	88.4	1.0	55.1	79.0
UNetFormer # [35]	ResNet-18	87.2	90.3	88.8	88.5	**0.2**	53.8	79.6
FT-UNetFormer # [35]	Swin-B	88.7	91.0	89.8	90.0	1.1	55.8	81.6
BAFormer-T	ResNet-18	88.2	90.8	89.5	89.6	0.4	54.9	81.0
BAFormer	ResNet-18	**89.8**	**91.7**	**90.7**	**90.8**	1.0	**56.4**	**83.1**

Qualitative results on the Mapcup test set are shown in Figure 9, displaying six selected visualizations. Observing the results in the first, second, and third rows, distinct types of farmland can be identified within dashed boxes, including rice fields at different growth stages, cultivated land covered by low-level vegetation, and recreational areas. These variations in farmland exhibit significant differences in color and texture, with considerable intra-class variability. In contrast, the BAFormer model effectively learns similarities among different variants, achieving the best identification of variants. This outcome demonstrates the method's robust feature learning and fitting capabilities. In the fourth row of results, there is a road between the areas of cropland in the dotted box that is barely noticeable to the naked eye, yet BAFormer can represent and distinguish less prominent category differences. This capability is further demonstrated in the results of the fifth and sixth rows. The proposed BAFormer method effectively showcases its ability to discern inter-class differences and judge intra-class variability when handling complex scenes.

To assess the optimization effects of the model on edges, we introduce the Boundary IoU measurement metric, with quantitative results presented in Table 4. Furthermore, we conduct visual comparisons of edges extracted by calculating Boundary IoU, showcasing qualitative results in Figure 10. A comprehensive evaluation from both quantitative and qualitative perspectives reveals that the proposed BAFormer achieves higher edge quality than mainstream methods.

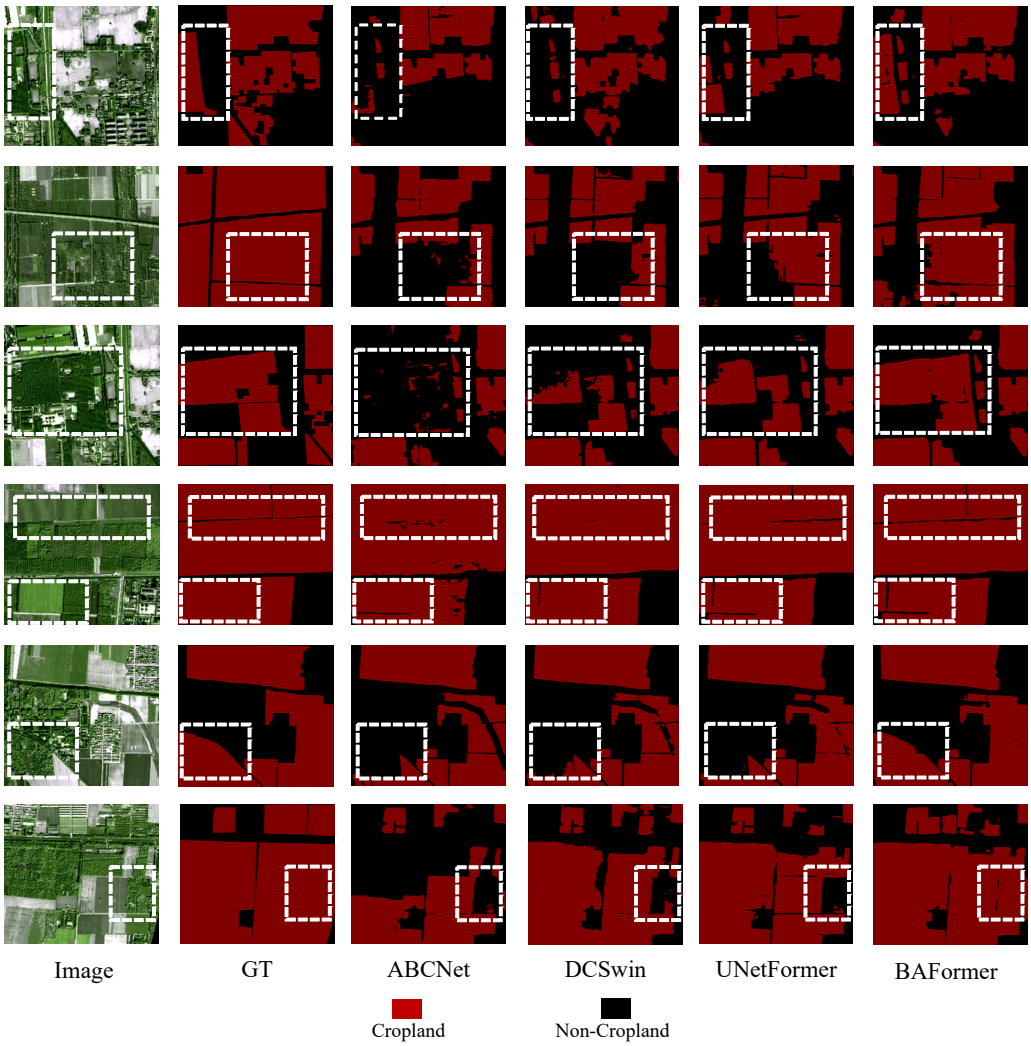


Figure 9. Qualitative comparisons with different methods on the Mapcup test set. We add some white dotted boxes to highlight the differences to facilitate model comparisons.

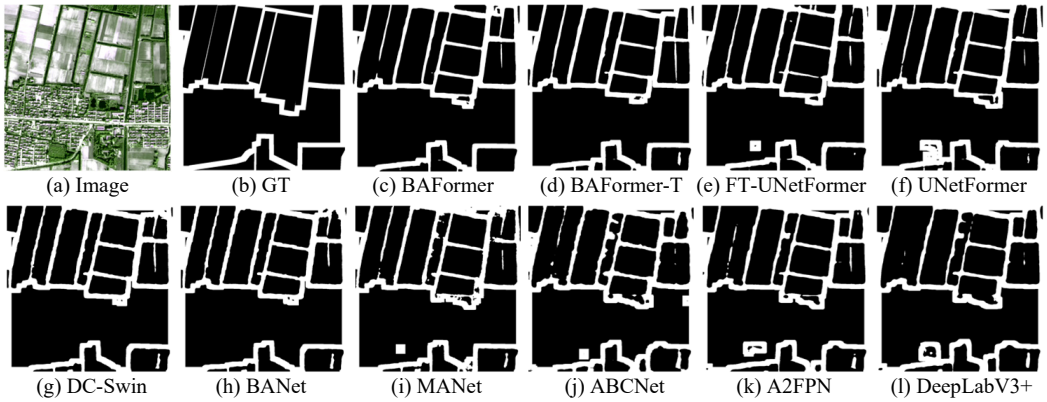
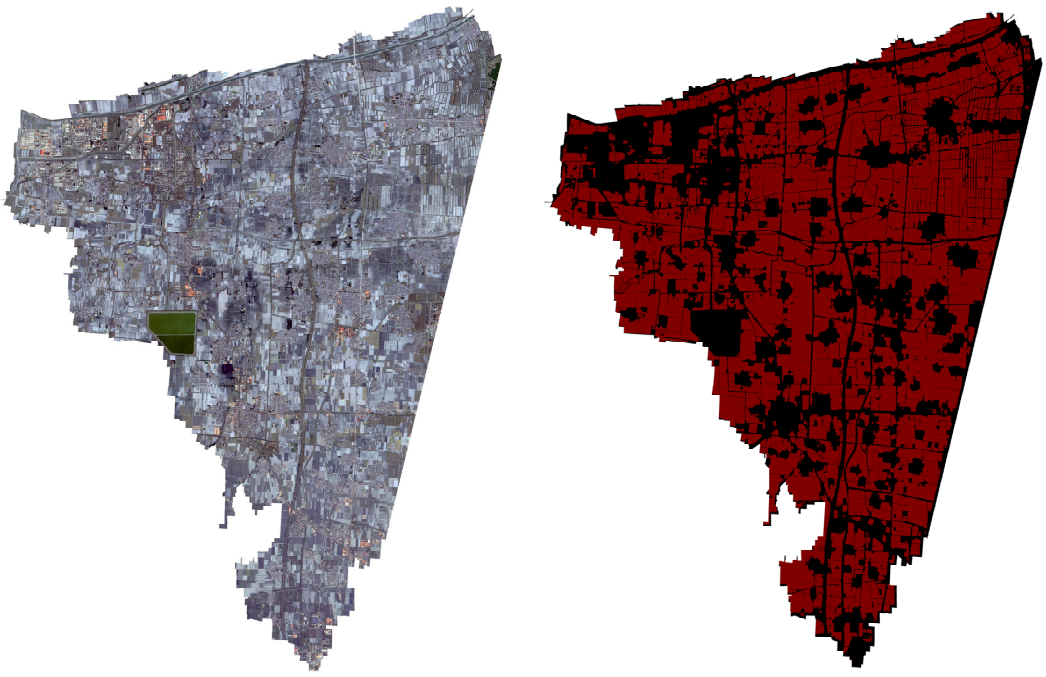


Figure 10. Visual comparisons of boundary extracted by different methods on the Mapcup test set will be conducted, and the resulting edges will be utilized for calculating Boundary IoU.

To further observe the performance of the model in the production environment, we randomly selected some areas for inference visualization, with the results shown in Figure 11. By analyzing the inference results from the visualization, we find that BAFormer can effectively eliminate the interference from the feature background, has

accuracy in boundary identification for various complex targets, and achieves satisfactory segmentation quality.



(a) Large Image (b) Inference Image

Figure 11. 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.

4.3. Ablation Experiment

4.3.1. Each Component of BAFormer

In BAFormer, we effectively enhance the accuracy of the model for boundary-aware segmentation by improving three aspects: feature extraction, feature fusion, and loss constraints. To further verify the effectiveness of each module, we conducted extensive ablation experiments on the Vaihingen dataset, and the results are shown in Table 5. It should be noted that for a fair comparison, we uniformly adopted the same stochastic enhancement strategy and test-time enhancement strategy, with relevant hyperparameters consistent with the benchmark model UNetFormer by default.

Table 5. Ablation study of BAFormer for each component on the Vaihingen dataset.

Dataset	Base	Component				mIoU (%)
		FAM	RAF	DWLK-MLP	Edge Constraint	
Vaihingen	✓					82.50
	✓	✓				83.61
	✓		✓			83.09
	✓			✓		83.44
	✓				✓	83.26
	✓	✓	✓			83.89
	✓	✓	✓	✓		84.22
	✓	✓	✓	✓	✓	84.48

4.3.2. Selection of Large Kernel Convolution

The DWLK-MLP module is designed to enhance the completeness of boundary extraction by deeply decomposing large kernel sensory fields. To further explore the effect of

using large kernel convolution in this module, we conducted correlation validation on the Mapcup and Vaihingen datasets, and the results are shown in Table 6. The experimental results demonstrate that choosing a large kernel convolution with a size of 23 yields the highest accuracy. Further increasing the kernel size reduces accuracy and increases model complexity. Therefore, we set the size of the large kernel convolution to 23 by default and use a sequence of small kernel convolutions with sizes 5 and 7 for DW decomposition. This approach not only maintains the sensing field of the large kernel but also avoids excessive computational complexity and increases model depth, thereby improving the model's generalization ability.

Table 6. Ablation of large kernel in DWLK-MLP. Here, 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

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 Vaihingen public dataset. By comparing the results (see Table 7), 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, or 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 Depthwise Large Kernel Multi-Layer Perceptron 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.

Table 7. Comparison with current state-of-the-art lightweight networks on the Vaihingen dataset and testing their complexity and model parameters.

Method	Backbone	Memory (M)	Params (M)	Complexity (G)	mIoU (%)
DANet [56]	ResNet-18	611.1	12.6	39.6	68.8
BiSeNet [66]	ResNet-18	970.6	12.9	51.8	69.1
Segmenter [67]	ViT-Tiny	933.2	13.7	63.3	73.6
BoTNet [68]	ResNet-18	710.5	17.6	49.9	74.3
FANet [69]	ResNet-18	971.9	13.6	86.8	75.6
ShelfNet [70]	ResNet-18	579.0	14.6	46.7	78.3
SwifNet [71]	ResNet-18	835.8	11.8	51.6	79.9
ABCNet [55]	ResNet-18	1105.1	14.0	62.9	81.3
MANet [57]	ResNet-50	1169.2	12.0	51.7	82.7
UNetFormer [35]	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

4.3.4. The Stability of the Model

In practical applications, the stability and adaptability of the model are particularly important when faced with different input sizes. To comprehensively evaluate the stability of the model's performance under various input sizes, we trained the lightweight BAFormer-T

using square image inputs of different scales, including common scales, e.g., 512 × 512, 768 × 768, 1024 × 1024, as well as the super-large 2048 × 2048 input scale. The results are presented in Table 8. The results demonstrate that the lightweight BAFormer-T exhibits excellent stability and adaptability when handling various input scales. The overall mIoU deviation is no more than 0.6%, the MeanF1 deviation is within 0.4%, and the OA deviation is less than 0.2%. This indicates that the model can maintain good performance with input data of different sizes, showcasing excellent robustness and generalization ability, and is suitable for diverse practical application scenarios.

Table 8. Ablation on the Vaihingen dataset investigates the effect of different input sizes on model stability. The experiments were carried out on a single NVIDIA GTX 3090 GPU using the BAFormer-T model.

Input Size	Per-Class F1 (%)					MeanF1 (%)	OA (%)	mIoU (%)
	Imp.surf	Building	Low.veg	Tree	Car			
512 × 512	93.19	95.83	85.36	90.89	89.39	90.93	91.48	83.63
768 × 768	93.56	95.93	85.42	90.72	90.14	91.15	91.60	83.94
1024 × 1024	93.67	95.89	85.33	90.79	90.87	91.30	91.63	84.16
2048 × 2048	93.40	95.79	85.60	90.73	89.88	91.08	91.54	83.80

4.3.5. Number of Multi-Heads and Window Size

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 12. 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 it too large or too small will hinder the feature extraction performance of the model. In the ablation experiments with window size, we found that a window size of 8 yielded the best overall performance. However, increasing the window size further led to decreases in mIoU and F1-Score.

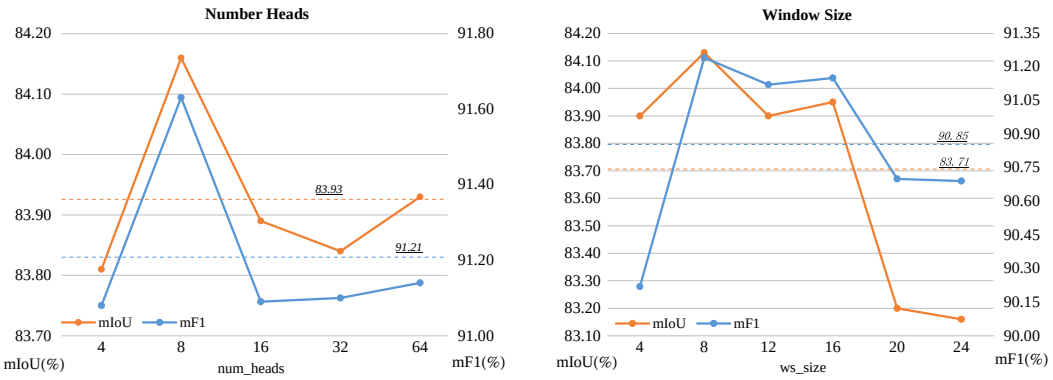


Figure 12. Ablation study on the number of model multi-heads and window size on the Vaihingen dataset.

5. Discussion

5.1. FAM Feature Visualization

The proposed FAM possesses dynamic learning capabilities. To validate its effectiveness, we conducted visual comparisons with classical attention mechanisms such as Shift Window Attention, Shuffle Attention, Criss-Cross Attention, and Global-Local Attention.

We visualize the features post-attention in both spatial and frequency domains within the decoder, as depicted in Figure 13. From the spatial perspective (a), our proposed FAM Attention prominently preserves global and local image features. In the D1 deep layer with semantics alone and the fusion layers D2, D3, and D4, the feature maps after FAM further enhance boundary perception, better preserving clear edges, textures, and other fine-grained features. In the frequency domain (b), we observe that the image's low-frequency features appear more rounded and extensive across F-D1, F-D2, and F-D3 layers. This results from FAM's adaptive selection mechanism based on channel fusion contributions, dynamically blending high and low-frequency features in varying proportions across different network depths, thereby achieving a refined and comprehensive feature representation.

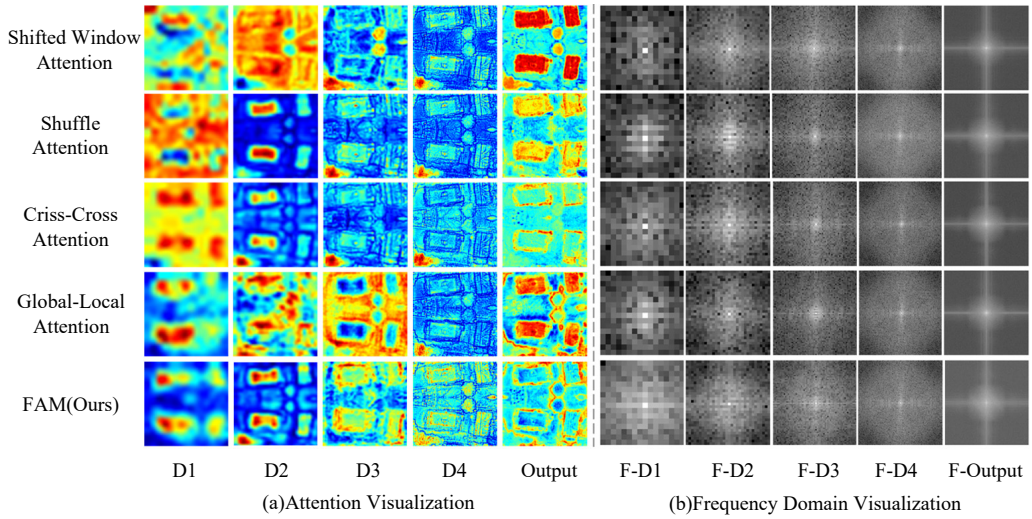


Figure 13. Feature Adaptive Mixer (FAM) feature map visualization.

5.2. About the Choice of Encoder

Considering the significant impact of encoders with different feature extraction capabilities on decoder fusion, we further explored their influence on model segmentation, as shown in Table 9. The study results indicate that ResNet-18 achieves the best overall fusion and segmentation accuracy due to its efficient extraction of shallow features. Specifically, the BAFormer model achieves a mIoU of 84.47%, while the lightweight model BAFormer-T achieves 84.15% mIoU. Notably, BAFormer-T significantly reduces model parameters, complexity, and flops compared to other CNN-based encoders, without compromising accuracy levels. Moreover, compared to the Transformer-based Swin_Base encoder, BAFormer only slightly decreases accuracy, demonstrating satisfactory performance.

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

Continuing from Table 9, Transformer-based encoders generally achieve higher accuracy compared to CNN-based encoders. However, BAFormer benefits from the specialized design of RAF, enabling dynamic feature fusion selection in the decoder stage based on task requirements, thereby enriching feature granularity. This result further confirms that in the image reconstruction phase of the decoder, besides robust high-level semantic feature support, integrating more shallow-level fine-grained features is essential.

5.3. Further Exploration on Edge Constraint

The depth-supervised edge constraint strategy guides the model to improve edge accuracy by applying strong constraints to the edges. For instance, on the Vaihingen dataset, the overall mIoU accuracy improved by approximately 0.5%. However, the effectiveness of this method may vary for image datasets with difficult-to-identify terrain categories. Further exploration of the Vaihingen dataset reveals its characteristics of regular, flat terrain with prominent features of terrain categories. Furthermore, a unique preprocessing method involves dyeing low-altitude vegetation areas red during imaging, which enhances their identification. This significantly enhances the overall recognition accuracy of the Vaihingen dataset, with only slight boundary blurring between predicted and labeled images, as shown in Figure 1a. In contrast, the Potsdam dataset features a more complex urban background and lacks the same preprocessing dyeing operations as the Vaihingen dataset, increasing the difficulty of object recognition for models. Therefore, the effectiveness of deep edge supervision is less obvious. This finding further underscores the interdependence between edge constraint and feature recognition capabilities. When the model achieves high image recognition accuracy with minimal boundary-blurring, this method effectively improves edge quality. Conversely, if the model's ability to represent images is insufficient, the effect of edge constraint may not be significant.

5.4. Research Difficulties and Next Steps

Cropland extraction is a challenging task. Firstly, enhancing edge perception can improve the recognition of farmland and its boundaries to some extent. However, images often contain objects resembling farmland, such as buildings and trees, making reliance on individual image information insufficient for accurate patch extraction. Secondly, insufficient spectral information complicates the distinction between farmland and other features, particularly when they share similar spectral characteristics. Furthermore, data inconsistency limits the model's ability to generalize across different times and regions, posing challenges for accurate farmland extraction.

To address these challenges, the next step will focus on multi-modal fusion. Multi-source data fusion can provide richer information and effectively enhance the precise extraction capability of croplands.

6. Conclusions

This paper addresses inaccurate boundary extraction in high-resolution remote sensing images for cropland by proposing BAFormer, a UNet-like universal extraction model. BAFormer enhances edge feature compensation through three stages: feature extraction, feature fusion, and loss constraints. It incorporates a FAM based on channel fusion and a DWLK-MLP module with a large receptive field to amplify high-frequency information and feature expression, enabling it to identify more complete and accurate boundaries. Additionally, a Relational Adaptive Fusion (RAF) strategy and edge constraint guide the model's attention towards edges, enhancing accuracy. The evaluation of datasets including Vaihingen, Potsdam, LoveDA, and Mapcup demonstrates BAFormer's promising performance in model size reduction, generalization ability, and edge quality, validating its effectiveness.

Author Contributions: Conceptualization, Y.W. and Z.L.; methodology, Y.W. and K.L; software, Y.W.; validation, F.T. and Y.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.

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.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author. Code can be obtained at https://github.com/WangYouM1999/BAFormer (accessed on 5 July 2024).

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

References

1. Toth, C.; Jóźków, G. Remote Sensing platforms and sensors: A survey. *ISPRS J. Photogramm. Remote Sens.* **2016**, *115*, 22–36. [CrossRef]
2. Yang, C. Remote sensing and precision agriculture technologies for crop disease detection and management with a practical application example. *Engineering* **2020**, *6*, 528–532. [CrossRef]
3. 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 Sens. Environ.* **2020**, *247*, 111912. [CrossRef]
4. 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. *Comput. Electron. Agric.* **2023**, *206*, 107683. [CrossRef]
5. 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 J. Photogramm. Remote Sens.* **2023**, *200*, 24–40. [CrossRef]
6. Zuo, R.; Zhang, G.; Zhang, R.; Jia, X. A Deformable Attention Network for High-Resolution Remote Sensing Images Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
7. 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 Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. [CrossRef]
8. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
9. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20. [CrossRef]
10. Xia, L.; Luo, J.; Sun, Y.; Yang, H. Deep Extraction of Cropland Parcels from Very High-Resolution Remotely Sensed Imagery. In Proceedings of the 2018 7th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Hangzhou, China, 6–9 August 2018; pp. 1–5.
11. Xie, Y.; Zheng, S.; Wang, H.; Qiu, Y.; Lin, X.; Shi, Q. Edge Detection With Direction Guided Postprocessing for Farmland Parcel Extraction. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 3760–3770. [CrossRef]
12. Awad, B.; Erer, I. FAUNet: Frequency Attention U-Net for Parcel Boundary Delineation in Satellite Images. *Remote Sens.* **2023**, *15*, 5123. [CrossRef]
13. Hao, S.; Wang, W.; Ye, Y.; Nie, T.; Bruzzone, L. Two-Stream Deep Architecture for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2349–2361. [CrossRef]
14. Doersch, C.; Gupta, A.; Efros, A.A. Unsupervised Visual Representation Learning by Context Prediction. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; p. 9.
15. 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. In Proceedings of the 2023 11th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Wuhan, China, 25–28 July 2023; pp. 1–6.
16. 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 Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16. [CrossRef]
17. Iizuka, R.; Xia, J.; Yokoya, N. Frequency-based Optimal Style Mix for Domain Generalization in Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 1–14. [CrossRef]
18. 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 Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
19. 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 Sens.* **2019**, *11*, 108. [CrossRef]

20. 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 Sens.* **2022**, *14*, 2157. [CrossRef]
21. 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. [CrossRef]
22. Luo, W.; Zhang, C.; Li, Y.; Yan, Y. MLGNet: Multi-Task Learning Network with Attention-Guided Mechanism for Segmenting Agricultural Fields. *Remote Sens.* **2023**, *15*, 3934. [CrossRef]
23. 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 J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 3060–3069. [CrossRef]
24. 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. *Comput. Electron. Agric.* **2024**, *220*, 108902. [CrossRef]
25. Pan, Y.; Wang, X.; Wang, Y.; Zhong, Y. RBP-MTL: Agricultural Parcel Vectorization via Region-Boundary-Parcel Decoupled Multitask Learning. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15. [CrossRef]
26. 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. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 2397–2405. [CrossRef]
27. Kervadec, H.; Bouchtiba, J.; Desrosiers, C.; Granger, E.; Dolz, J.; Ben Ayed, I. Boundary Loss for Highly Unbalanced Segmentation. *Med. Image Anal.* **2021**, *67*, 101851. [CrossRef] [PubMed]
28. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
29. Jin, Q.; Meng, Z.; Pham, T.D.; Chen, Q.; Wei, L.; Su, R. DUNet: A deformable network for retinal vessel segmentation. *Knowl. Based Syst.* **2019**, *178*, 149–162. [CrossRef]
30. 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 Trans. Med. Imaging* **2024**, *43*, 2254–2265. [CrossRef]
31. Pham, T.H.; Li, X.; Nguyen, K.D. Seunet-trans: A simple yet effective unet-transformer model for medical image segmentation. *arXiv* **2023**, arXiv:2310.09998.
32. Wang, Y.; Gu, L.; Jiang, T.; Gao, F. MDE-UNet: A Multitask Deformable UNet Combined Enhancement Network for Farmland Boundary Segmentation. *IEEE Geosci. Remote Sensing Lett.* **2023**, *20*, 1–5.
33. Xu, Y.; Zhu, Z.; Guo, M.; Huang, Y. Multiscale Edge-Guided Network for Accurate Cultivated Land Parcel Boundary Extraction From Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–20. [CrossRef]
34. Wu, D.; Guo, Z.; Li, A.; Yu, C.; Gao, C.; Sang, N. Conditional Boundary Loss for Semantic Segmentation. *IEEE Trans. Image Process.* **2023**, *32*, 3717–3731. [CrossRef] [PubMed]
35. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of Remote Sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
36. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
37. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998-6008.
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* **2022**, arXiv:2207.05501.
39. Tan, W.; Geng, Y.; Xie, X. FMViT: A multiple-frequency mixing Vision Transformer. *arXiv* **2023**, arXiv:2311.05707.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognit, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
41. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
42. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
43. Zhang, X.; Gong, Y.; Li, Z.; Gao, X.; Jin, D.; Li, J.; Liu, H. SkipcrossNets: Adaptive Skip-cross Fusion for Road Detection. *arXiv* **2023**, arXiv:2308.12863.
44. Guo, S.; Liu, L.; Gan, Z.; Wang, Y.; Zhang, W.; Wang, C.; Jiang, G.; Zhang, W.; Yi, R.; Ma, L.; et al. Isdnet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit, New Orleans, LA, USA, 18–24 June 2022; pp. 4361–4370.
45. Guo, M.H.; Lu, C.Z.; Liu, Z.N.; Cheng, M.M.; Hu, S.M. Visual attention network. *Comput. Vis. Media* **2023**, *9*, 733–752. [CrossRef]
46. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M. Coatnet: Marrying convolution and attention for all data sizes. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3965–3977.
47. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. Cmt: Convolutional neural networks meet vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit, New Orleans, LA, USA, 18–24 June 2022; pp. 12175–12185.

48. Shi, D. TransNeXt: Robust Foveal Visual Perception for Vision Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit, Seattle, DC, USA, 17–21 June 2024; pp. 17773–17783.
49. He, W.; Li, J.; Cao, W.; Zhang, L.; Zhang, H. Building extraction from Remote Sensing images via an uncertainty-aware network. *arXiv* **2023**, arXiv:2307.12309.
50. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. Loveda: A remote sensing land-cover dataset for domain adaptation semantic segmentation. *arXiv* **2021**, arXiv:2110.08733.
51. Sun, Y.; Wang, S.; Chen, C.; Xiang, T.Z. Boundary-guided camouflaged object detection. *arXiv* **2022**, arXiv:2207.00794.
52. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognit, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
53. 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 Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
54. Li, R.; Zheng, S.; Duan, C.; Su, J.; Zhang, C. Multistage attention ResU-Net for semantic segmentation of fine-resolution Remote Sensing images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
55. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Wang, L.; Atkinson, P.M. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 84–98. [CrossRef]
56. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. *Remote Sens.* **2021**, *13*, 3065. [CrossRef]
57. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution Remote Sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
58. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A novel transformer based semantic segmentation scheme for fine-resolution Remote Sensing images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
59. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit, New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299.
60. Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Transformer-based decoder designs for semantic segmentation on remotely sensed images. *Remote Sens.* **2021**, *13*, 5100. [CrossRef]
61. Kirillov, A.; Girshick, R.; He, K.; Dollár, P. Panoptic feature pyramid networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit, Long Beach, CA, USA, 15–20 June 2019; pp. 6399–6408.
62. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A. Foreground-aware relation network for geospatial object segmentation in high spatial resolution Remote Sensing imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit, Seattle, WA, USA, 13–19 June 2020; pp. 4096–4105.
63. Chen, K.; Zou, Z.; Shi, Z. Building extraction from Remote Sensing images with sparse token transformers. *Remote Sens.* **2021**, *13*, 4441. [CrossRef]
64. Cheng, B.; Schwing, A.; Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17864–17875.
65. Li, R.; Wang, L.; Zhang, C.; Duan, C.; Zheng, S. A2-FPN for semantic segmentation of fine-resolution remotely sensed images. *Int. J. Remote Sens.* **2022**, *43*, 1131–1155. [CrossRef]
66. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.
67. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 7262–7272.
68. Srinivas, A.; Lin, T.Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck transformers for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit, Nashville, TN, USA, 20–25 June 2021; pp. 16519–16529.
69. Hu, P.; Perazzi, F.; Heilbron, F.C.; Wang, O.; Lin, Z.; Saenko, K.; Sclaroff, S. Real-time semantic segmentation with fast attention. *IEEE Rob. Autom. Lett.* **2020**, *6*, 263–270. [CrossRef]
70. Zhuang, J.; Yang, J.; Gu, L.; Dvornek, N. ShelfNet for Fast Semantic Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 847–856.
71. Oršić, M.; Šegvić, S. Efficient semantic segmentation with pyramidal fusion. *Pattern Recognit.* **2021**, *110*, 107611. [CrossRef]