ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

SAM2-MAFLNet: Knowledge distillation of multi-attn fusion and laplacian regularization in remote sensing segmentation

Xiaoliang Tang[1,*]

1 School of Computer Science and Technology, Zhejiang University of Science and Technology, Hangzhou, Zhejiang, 310023

SUMMARY: *Due to the significant intra-class variation, small target detection problem under low-contrast clutter background, accurate semantic segmentation of very high resolution remote sensing images still exists many difficulties at present. For the above problems, we introduce SAM2-MAFLNet as an improved teacher-student knowledge-distillation approach to build a low-resource model that achieves high-density segmentation of remote sensing images. The student develops a Multi-attention and cross-fusion architecture (MACA) that jointly learns long-range spatial relationships and channels dependency, and an enhanced Laplacian high-frequency enhancement module (LHFEM) introduces edge-aware information via laplace pyramid decomposition and foreground-background segmentation. Experiments on the ISPRS Vaihingen and Potsdam benchmarks show that the teacher network achieves 91.92% mF1/85.06% mIoU and 93.64% mF1/88.21% mIoU, respectively, and the student retains competitive accuracy with only 12.52M parameters. The above experiments show that, in terms of balancing segmentation accuracy and real-time requirements for application on remote-sensing data.*

KEYWORDS: *Remote sensing semantic segmentation; Knowledge distillation; SAM2; Laplacian enhancement; Multi-attention fusion*

1 Introduction

Recent developments in remote sensing platforms have allowed high-resolution images to be captured; thus, there are now many possibilities for land-use classification, urban planning applications, and environmental studies. [1, 2] In this case, as a way of assigning classes to every point in space; therefore, providing a strong ability to depict complex situations at high fidelity levels. It has obvious application value in the Tasks of change detection, Urban 3D modelling, Infrastructure Inventory etc., that need to keep objects' boundaries clear while ensuring uniformity of classes.

With the continuous advancement of deep learning, semantic segmentation's research has expanded from natural imagery processing; It is now also widely used in geology, urban-environmental analysis, etc., for analysing large areas or entire landscapes with high accuracy. [3-11] Recently, the improvement of convolutional and transformation-based models has been confirmed; However, very high resolution remote sensing images are still under great challenge as they contain significant changes in scale, dense objects' arrangement and obvious spectrum-texture differences.

While there have been significant advancements in precise division of very-high-definition

*121110@zust.edu.cn
https://doi.org/10.65102/is2026851
Published: 30 April, 2026

remote-sensing imagery; [12, 13]. In terms of capturing subtle local structures, such as narrow streets, vehicles and boundary transitional areas; At the same time, also need to construct distance relationship with other surroundings when modeling far-away City-scales image. If either side does not meet requirements, the prediction will have boundaries that are fragmented, classes confused, or some small objects missing.

Convolutional Neural Networks (CNN) segmentation networks have learned local textures and geometry; However, their small size of information cannot fully recognise long-distance semantic connections in extended areas with diverse properties. Transformer-based models overcome the problem of lacking global dependency modelling by means of self-attention [14-16]; however, their high computational demand and potentially poor performance with respect to preserving fine-boundary structure remain challenges. There is thus no way of achieving both high precision and feasible application simultaneously through one system.

Transfer learning can offer a direct approach to solve this problem by adjusting these huge pre-training models to the downstream segmentation task, thus saving computational resources and accelerating model convergence in remote-sensing-based training [18]. In addition, the foundation models SAM and SAM2 have received more attention due to their high representational ability and generalizability [19, 20]. These features make them good teachers' backs for remote sensing segmentation, as long as the extent of their field gap and computing load are manageable.

As SAM2 cannot be directly applied to remote sensing images [21-23]. Based on prompt-driven segmentation and developed for pre-training in a natural image distribution environment; Its prediction performance may deteriorate if no sufficient information or different scene statistics are provided during inference. Lightweight adapter strategies partially solve this problem by reducing the number of learnable parameters and retaining the frozen backbone; however, another mechanism is still required to transfer rich teacher representations to a relatively simple student model.

To solve the above problems, We build the knowledge-distillation framework of SAM2-MAFLNet combining a SAM2-based teacher and a lightweight student model. Introduce a Multi-attention and cross-fused architecture (MACA) to enhance the spatial-temporal contextual relationship model; Introduce a Laplacian high-frequency enhancement module (LHFEM), which improves Boundary segmentation and has strong small object recognition capability. Therefore, in this way, the proposed model can maintain its recognitive ability at a higher resolution than SAM-2 without increasing computing power or reducing adaptability significantly.

In short, the main results of this study are as follows.

•SAM2-guided teacher-student Framework. A knowledge-distillation Pipeline is designed for transfer learning that brings SAM 2.0 as the teacher and introduces it into lightweight representation, thereby boosting accuracy while maintaining deployability of system.

•Multi-attention and cross-fused architecture (MACA). Propose a dual-branch attention structure comprising Linear-Angular Attention and Cross-Covariance Attention; then add cross-fusion for joint learning of long-distance spatial relationships and channel interactions.

Laplacian high-frequency enhancement module (LHFEM). Introduce an edge-aware enhancement strategy based on laplacian-pyramid decomposition and reversed attention to strengthen boundary localisation and improve the recognition performance of small, visually vague objects.

2 Related Work

2.1 Remote Sensing Semantic Segmentation Based on CNN and Transformer

For many years, Convolutional Neural Networks (CNNs) have been used to solve semantic segmentation problems of remote sensing images. Pioneered by fully convolutional networks (FCNs), the first end-to-end architecture was proposed to solve dense-pixel prediction problems; However, this simple design lost some information in detail during decoding. Overcoming these limitations, U-Net [24-27] introduced a symmetrical architecture of encoder-decoder with skip connection to fuse low-level spatial information and high-level semantic feature maps, thereby improving segmentation accuracy and robustness in terms of structure.

After U-Net, there have been many attempts to expand the receptive field and enhance discriminability. Some representative developments include atrous Convolutional neural networks (D-Net) [28] in the DeepLab family, pyramid pooling of PSPNet [29], attention-enhanced methods like SENets [30] and non-local self-attention models. However, although these methods effectively enhance context-aware feature learning; Still have problems such as checkerboard noise in atrous convolution, weak multi-scale combination of multi-scale aggregation function, high computational complexity for attention mechanisms.

To enhance the global recognition capability of CNNs, some additional approaches, such as boundary refinement modules [31], global-deconvolution layers [32], and joint modelling of spatial-channels dependencies [33], have also been added. Although these approaches help enhance the global semantic consistency globally; However, problems such as an unevenly distributed global receptive field distribution and poor multi-scale combination in complex high-precision remote-sensing images remain unresolved.

Based on transformer-based segmentation frameworks that employ self-attention mechanisms to model large-scale spatial structures effectively and achieve good performance in very high-resolution remote-sensing image scene segmentation [34]. Visualisation transformers and subsequent models have demonstrated excellent scalability when dealing with vast amounts of images to extract global semantics well. But lacking in-built localisation priors, they are not conducive to learning subtle textures and boundary-level detail effectively. Balancing the scales of both Global and Local representation, several ways exist currently. UNetFormer [35], which introduces a simple global-local attention mechanism at the decoder stage of UNet to enhance feature extraction; WiCoNet [36] combines scene-level semantic information with local features through an embedding layer based on context-transformers; And Swin-Transformers [37] integrate two novel strategies: Hierarchical Window Shifting, increasing spatial resolution while reducing computational costs. While these models address some issues to varying degrees, effectively fusing detailed spatial information and comprehensive context understanding from ultra-high-resolution remote sensing images remains an unsolved problem.

It can be seen from these results that raising the strength of the primary module is one way; however, how to resolve multiple problems simultaneously remains difficult in actual implementation. Our solution takes this path of development, building on the SAM2-based teacher that provides powerful representation and developing a light-weighted model for learning via task-oriented distillation processes focused on remote-sensing-segmentation problems.

2.2 Segment Anything Model 2

SAM established a new paradigm of generalised Segmentation by coupling the initial size and pre-trained model. However, when applied directly to remote sensing images, the problem of domain mismatch and high computation makes its actual application restricted; In terms of Dense Urban Scenes and Resource-Constrained Deployment Settings.

The SAM2 model increases memory use and inference speed specifically in Image-Video Segmentation Workflows. However, from an absolute perspective, it is still initially dependent, and thus cannot be considered standalone remote sensing segmenters in situations without reliable prompts. Recently, some adapter-based fine-tuning approaches have alleviated this issue to some extent by adding light-weight modules to the frozen backbones and retaining strong pre-trained feature information.

Based on this line of research, we build an energy-efficient framework that employs a low-complexity teacher based on SAM2 image encoder to collaboratively learn weights in conjunction with lightweight adaptation. Distillation is used next to transmit teachers' structural and semantic knowledge into an abbreviated student network. Integrate this technology with MACA-based context model construction, LHFEM-boundary refinement techniques to boost the precision of object recognition; At the same time improve deployment efficiency.

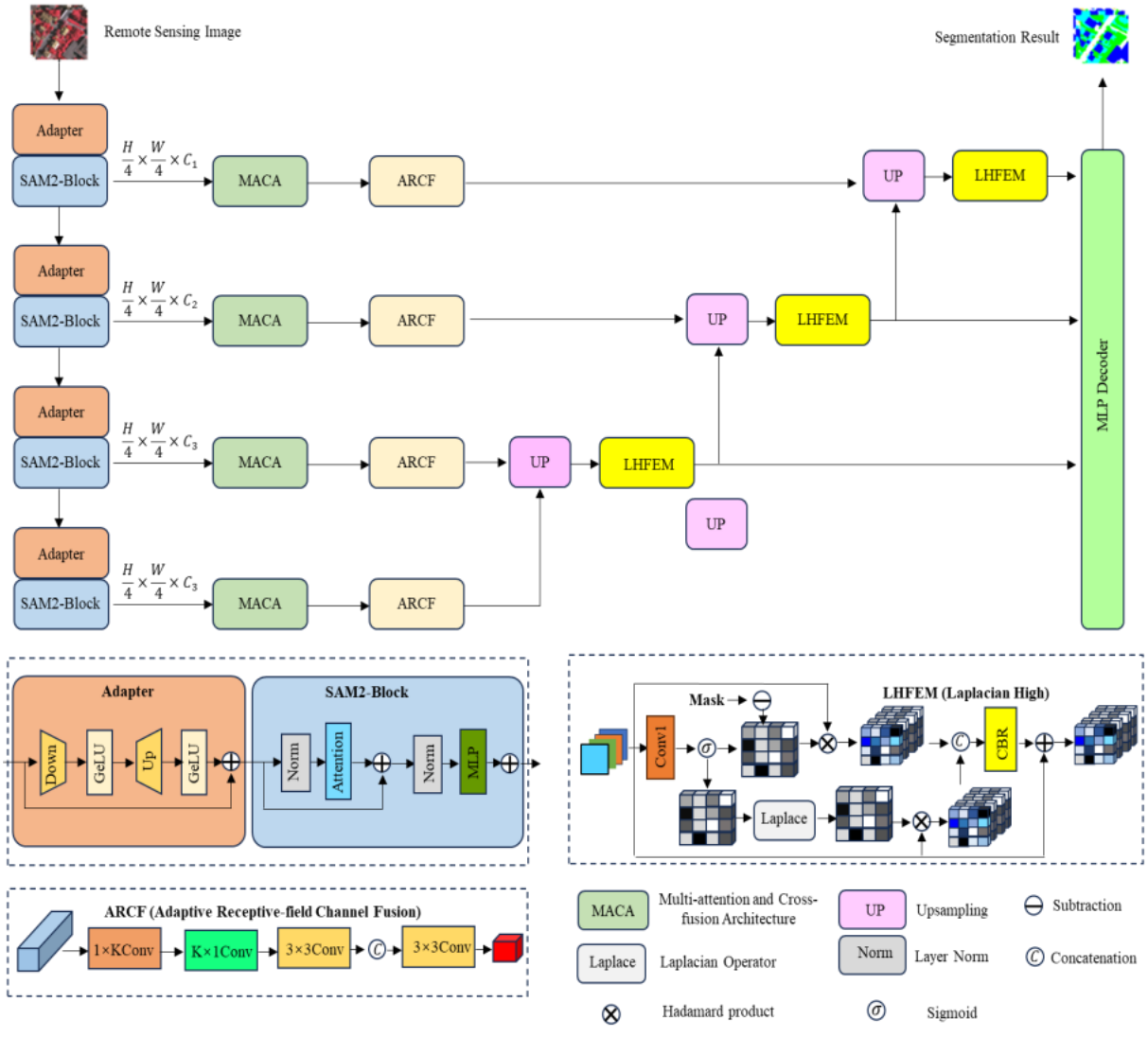

Figure 1: Network Structure.

2.3 Frequency-Domain Information in Semantic Segmentation

Several works have explored using Laplacian Pyramid to enhance the representation of structures and boundaries to improve segmentation accuracy and visual recognizability. Ghiasi and Fowlkes introduced a multi-scale refine method based on the Laplace pyramid [38]; First, it combines features at different scales from high to low order in turn; And Then, use multiplication-summation gate formulas to improve boundaries non-parametrically without introducing complicated random field models or other types of detections. Wang et al. [39] developed SWDL-Net that integrates Laplacian Pyramid decomposition with differential learning in deep CNNs; it uses the high-frequency part for edge localisation, and the low-frequency part to preserve texture continuity to obtain satisfactory results under relatively small numbers of labelled data (only 2%). Srivastava et al. [40] created LAqua to segment mariculture scenes; it decomposed the input image using three stages of Laplacians explicitly to separate high-frequency components, improving boundary detection in low-contrast and noisy images.

Based on this study, a Design is introduced for High-Frequency Boundary Information within The Distillation Framework of LHFEM. LHFEM does not perform a significant frequency domain transformation but instead combines a Spatial-Domain Laplacian Enhancement function with an Edge-Attention Module and a Reverse Attention Module for boundary reconstruction and supplementing Region-Level semantics.

2.4 Knowledge Distillation in Semantic Segmentation

Knowledge distillation (KD) has emerged as a popular way of making small-weighted segmentation models approach that of the much larger-sized network through knowledge transfer from a teacher model to a student one. The basic idea is to extract rich semantic features in a large-scale network to train an appropriate small compact teacher model better for generalisation, thus reducing the need for resources. Hinton, et al. [41] presented a formulation of KD through softened teacher's output distribution to enhance its capacity for capturing the inter-class relationship model; Adriana et al. [42] Also mentioned that it is necessary to transfer the middle results; An experiment shows that structured information included in feature maps can be learned from distilled models effectively.

Subsequently, many kinds of KD are put forward; Generally speaking, there are mainly two types: Response-oriented approach [43, 44]; The other is Feature-based Approach [45]; Relation-based Distillation Methods [46, 47]. Chen et al. added a "learning from the past" mechanism to gradually add historical cue features in the distillation process; Feng et al. used residual attention to combine both pixel-level and class-level correlation characteristics of teachers and students. Liu, etc., have applied adversarial learning to enhance the consistency of semantics; And Shu et al. add selective transmission of discriminative knowledge via a channel attention mechanism. Among them are other well-known relations-of transfer methods, including Yang et al.'s [48]' Global Semantic Constraint Exploitation Relation Transfer Method and Xia et al.'s relation transfer emphasizing the need to accurately transfer Inter-class Relationships For Robust Segmentation.

Although knowledge distillation has obtained excellent performance in semantic segmentation; many existing methods focus only on output supervision or local feature correspondence and fail to meet the dual needs of precise boundaries under long-distance dependence models for remote sensing images comprehensively. Sam2-MAFLNet fills this gap by integrating a strong capacity SAM2 teacher, multi-aggregation of features, and frequent Boundary redefinition; The student model thus acquires structural fidelity and contextual discriminability simultaneously.

3 Methodology

3.1 Network Architecture

SAM2-MAFLNet includes a teacher network (MAFLNet-T) and a student network (MAFLNet-S), which use different backbones under the encoder-decoder framework; As shown in Figure 1. The teacher employs a SAM2-based encoder for its rich semantic representation; The student utilises SegFormer-B0 as it is relatively light in terms of computation. Since both branches have a shared top-tier pipeline; This time will pay more attention to the Teacher Network, then students are compact counterparts of this structure in distillation frameworks.

MAFLNet-T's encoder consists of a lightweight version of SAM2; an adapter is added in front of each hierarchical level to make it efficient and still allows the pre-trained backbone to be frozen. Finally, after being processed by the proposed MACA block and cross-fusion module simultaneously, multi-resolution feature information is refined to improve long-distance Spatial Reasoning Ability as well as Channel-Interaction capability.

The decoder also combines LHFEM explicitly at the high-frequency boundary during progressive upsampling. This Design improves contour localisation; eliminates the influence of Background noise; preserves small-scale structures which are lost in coarse-level semantic coding. In addition, after transferring its semantic knowledge using the developed distilled method for MAFLNet-T and MAFLNet-S respectively, it can improve the segmentation accuracy of the student model without increasing too many calculations.

3.2 Multi-Attention and Cross-Fusion Architecture (MACA)

Ensure the accuracy of remote sensing semantic segmentation through joint modelisation of far-reaching spatial relationships and in-channels correlation. Most commonly, the old method of attention focuses at a low level and fails to extract universal characteristics for intricate cities generally. MacA tackles the above problems through integration of linear-angular Attention (LAA), Cross-Covariance Attention (XCA) with another novel Fusion Module located at the Cellular Level that can improve Spatial-Time Resolution across all Scales simultaneously.

LAA is less sensitive than DOTA attention due to its greater emphasis on the angles of two vectors after being normalised by each other. To enhance its adaptability to illumination changes, heterogenous texture variations, and backgrounds containing other elements across remote-sensing images. LAA performs a two-stage interaction to establish a channel-level Structure firstly and then further Refine The Informative Spatial Response Secondly; thereby improving Selecting Discriminative Features.

A Depthwise Convolution is then added to introduce locality continuity in the attention output. Therefore, LAA can model both long-range interactions and the neighbourhood-level geometric continuity of road networks, as well as those of other objects such as vehicles that are thin and patchy.

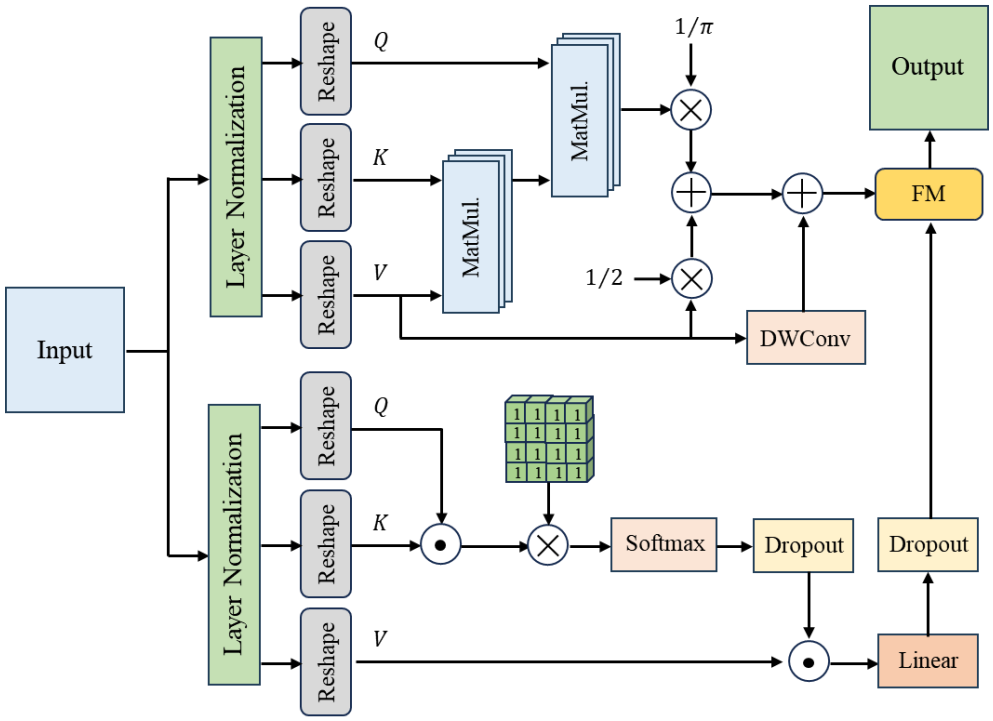

Figure 2: Multi-attention and cross-fused architecture.

Formally, given the input $X \in \mathbb{R}^{N\times L\times C}$, the query, key, and value matrices are obtained through:

$$[Q,K,V] = Linear(X) \tag{1}$$

where $Q,K,V \in \mathbb{R}^{N\times h\times L\times d}$ with h denoting the number of heads and $d = C/h$ the head dimension. To remove magnitude dependence, Q and K are normalized as:

$$\hat{Q} = Q/|Q|_2, \hat{K} = K/|K|_2 \tag{2}$$

Instead of directly computing $QK^{\mathrm{T}}V$, LAA first derives an intermediate representation through channel interaction:

$$A = \hat{K}^{\mathrm{T}}V \tag{3}$$

which captures the distribution of semantic information across channels. Then, the refined output is obtained by

$$Z = \alpha V + \beta\hat{Q}A \tag{4}$$

where $\alpha = 0.5$ and $\beta = 1/\pi$ are balancing factors following angular formulation. To further incorporate local structural cues, a depthwise convolution is applied on V, and the final output is expressed as

$$Y = \mathrm{Proj}\left(Z + \mathrm{DConv}(V)\right) \tag{5}$$

where $\mathrm{Proj}(\cdot)$ denotes a linear projection layer, and $\mathrm{DConv}(\cdot)$ denotes the depthwise convolution.

In summary, LAA is more suitable for modelling structures in very high resolution remote sensing image data compared with only considering magnitude when using magnitude-dominated attention.

XC-A uses a self-attention mechanism that is extended beyond its original scope and can address cross-channel dependence issues in traditional spatial-temporal Attention Models specifically. Traditionally, attention Mechanisms mainly pay attention to Spatial Relationships between Words; but it Is difficult To capture Semantics Over Multiple Feature Intervals. XC-A introduces a new cross-correlation definition to address the issue; its performance is good at improving the generalization capability of semantic information in the network. As an object that better solves issues related to channels' required semantic representations in the form of spectral decomposition and other means.

In XCA, given an input feature sequence $U \in \mathbb{R}^{N\times L\times C}$, the module computes its query Q_U, key K_U, and value V_U using (1): $\left[Q_U, K_U, V_U\right] = Linear\left(U\right)$.

Each of them is reshaped into multi-head form and normalized along the last dimension to stabilize the similarity measurement by employing (2): $\hat{Q}_U = Q_U / Q_{U2}, \hat{K} = K_U / K_{U2}$.

The cross-covariance matrix is then constructed as:

$$A = Softmax\left(\frac{\hat{Q}_U K_U}{\tau}\right) \tag{6}$$

where τ denotes a learnable temperature parameter that regulates the sharpness of the attention distribution. This matrix represents inter-channel feature correlations, and the resulting attention weights are subsequently applied to the value matrix to produce refined feature representations:

$$X' = AV \tag{7}$$

Finally, the refined representation is reshaped back to the original form and passed through a linear projection to restore the dimensionality:

$$XCA\left(U\right) = \mathrm{Proj}\left(X'\right) \tag{8}$$

where $\mathrm{Proj}(\cdot)$ denotes a linear projection layer.

XCA adds explicit covariance models to LAA through supplement. To strengthen the channel-based semantics, reduce repetition in response generation, and ensure optimisation stability via normalisation and temperature-scaling techniques. XCA is located inside MACA to achieve a level-of-abstraction across different channels and support spatial attention LAA.

Although linear-angular attention and cross-covariance attention can respectively capture complementarities of dependencies, their outputs require fusion to fully utilise the effects of space and channels. Therefore, a Fusion module (FM) is designed to strengthen the mutual influence of feature representations in enhancing model robustness by adjusting their weights adaptively across space and channels jointly.

Given two feature maps $f_1, f_2 \in \mathbb{R}^{B\times C\times H\times W}$ from the previous attention modules, FM first flattens them into sequences and applies average pooling and max pooling to extract both global

and salient responses. FM computes a cross-attention matrix to measure the dependence among two feature branches.

$$M = a_1 a_2^{\mathrm{T}} \tag{9}$$

$$\tilde{f}_1 = \mathrm{Softmax}\left(M\right) f_1 \tag{10}$$

$$\tilde{f}_2 = \mathrm{Softmax}\left(M\right) f_2 \tag{11}$$

where a_1 and a_2 are two feature branches. This way, features in one branch can serve as references for improving them in another.

To further highlight the information region $\tilde{f}$, FM adds a spatial attention refiner. Each time the updated function is added for averaging and maximumPooling, then pass through two sets of convolutional kernels to obtain Spatial Masks after Softmax Normalization.

$$\alpha = \mathrm{Softmax}\left(\mathrm{Conv}_2\left(\left[\mathrm{Avg}\left(\tilde{f}\right), \mathrm{Max}\left(\tilde{f}\right)\right]\right)\right) \tag{12}$$

Finally, the re-weighted features are as follows:

$$f_1^{'} = f_1 \odot \alpha_1 + f_1 \quad \text{and} \quad f_2^{'} = f_2 \odot \alpha_2 + f_2 \tag{13}$$

where α_1 and α_2 are obtained by substituting $\tilde{f}_1$ and $\tilde{f}_2$ into (12), respectively, and $\odot$ denotes Hadmard product operator (element-wise multiplication).

This Design can achieve the fusing effect by adaptingly merging the outputs from both LAA and XCA modules. Thus, this obtained model can resist noisy data better and is also suitable for representing multi-scale heterogenous spatiotemporal structures of remote sensing images.

Although Linear Angular attention (LAA) and cross-covariance attention (XCA) are able to detect different kinds of features separately, the last one performs best with their combinations. Thus, our Design of the multi-attention and cross-fusion architecture is proposed: The two attention sub-expressions work concurrently; After that, they will be combined through a fusion module. Next, given the input feature map of, we have obtaining $F \in \mathbb{R}^{B \times C \times H \times W}$.

$$F^{LAA} = LAA\left(F\right),? \quad F^{XCA} = XCA\left(F\right) \tag{14}$$

LAA is a kind of Angular Dependence and has local geometric invariance; On the contrary, XCA tries to catch cross-channels covariance interactions and overall dependencies. The two refined feature maps are inputted into the fusion module (fm) that is realised via cross attention mechanisms, and its function will combine their respective clues.

$$F^{MACA} = \mathrm{FM}\left(F^{LAA}, F^{XCA}\right) \tag{15}$$

Therefore, MACA combines LAA, XCA and cross-fusion within the same multi-attention module. Through joint enhancement of spatial dependence, channel interactions, and features to overcome the shortcomings of singular attention design and produce a more discriminative one for downstream segmentation task.

3.3 Laplacian High-Frequency Enhancement Module (LHFEM)

Boundary information must be provided for semantic segmentation to separate close objects and keep the structure of detailed features. Edges in remote-sensing images become less pronounced due to background clutter, changes in light conditions, texture mix-up, etc.; thus, they may blur or fail to identify smaller areas properly.

Taking into account these deficiencies, the Laplacian High-Frequency Enhancement Module (LHFEM) is proposed. As shown in Figure 1, the sub-module has an attention combination of edges and reverses. The former aims to detect contour data; The latter must exclude clutter from the background of other types through processing. Through combination of them to obtain a more reliable prediction since LHFEM can enhance Boundary Clarity and separate objects inside/outside better than traditional methods.

$$I_{re}^{i} = 1 - \sigma\left(F_{e}^{i}\right) \tag{16}$$

where $\sigma(\cdot)$ denotes the sigmoid function. Such a presentation impedes responses at edge areas as well; in other words, the distance-effect is reinforced by it to some extent. After integrating edge attention later on, reverse attention can be used by LHFEM to maintain contours sharp and interior consistency; subsequently, these equations are built upon it.

Additionally, at the same time as edges are identified using laplace pyramid decompositions; A pyramid whose basis is a Gaussian function and each level subtracts the neighbours creates high-frequency components. During implementation, for simplicity's sake, only the first Laplace-Adjoint Layer is employed in the forward pass. The input data can be smooth via a two-dimensional 5x5 Gaussian kernel that gives higher importance to the center pixels and gradually less significant weights for their neighbours. Eliminate noise, keep the original structure of it intact. After smoothening, the results are downsampled and upsampled to obtain I_{up}^{i}. Finally, perform pixel-wise subtraction of the input and I_{up}^{i} to obtain the residual image that captures high-frequency edge I_{e}^{i} components removed by Gaussian smoothing and downsampling.

$$I_{up}^{i} = \mathrm{GF}\left(\mathrm{Up}\left(\mathrm{Dn}\left(\mathrm{GF}\left(\sigma\left(F_{e}^{i}\right)\right)\right)\right)\right) \tag{17}$$

$$I_{e}^{i} = \sigma\left(F_{e}^{i}\right) - I_{up}^{i} \tag{18}$$

where $\mathrm{GF}(\cdot)$ denotes Gaussian filtering operator, $\mathrm{Up}(\cdot)$ denotes upsampling operator, $\mathrm{Dn}(\cdot)$ denotes downsampling operator, $\sigma(\cdot)$ denotes sigmoid function, and F_{e}^{i} denotes the edge feature map at scale i.

Compute reverse-attention and edge-attention weights first; then multiply these by corresponding feature maps from LFH-MoI to produce background-enhanced and edge-refined outputs. Then connect the above contents together and pass through a CBRe-enhancement block for joint enhancement, add an additional attention mask to reduce residual noise further and stress key regions.

$$F_{fuse}^{i} = \mathrm{CBR}\left(\mathrm{Cat}\left(F_{e}^{i} \times I_{re}^{i}, F_{e}^{i} \times I_{e}^{i}\right)\right) \tag{19}$$

where $\mathrm{CBR}(\cdot)$ consists of a 1×1 convolutional layer, batch normalization (BN), and ReLU function.

3.4 Loss Functions

3.4.1 Teacher Optimization

The teacher network is trained with a mixed-loss function including pixel-wise cross-entropy, Dice loss and auxiliary cross-entropy to achieve dense-pixel level semantic segmentation. By combining class-to-discrimination, regions-overlap and intermediate-feature-guidance respectively to enhance the model's convergence stability and mask consistency together.

The first kind of cross-entropy loss is to minimise the difference between predicted class probabilities and one-hot-encoded ground-truth labels. Formally, it is as follows:

$$L_{ce} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_k^n \log \hat{y}_k^n \tag{20}$$

where N is the number of training samples, K is the total number of classes, y_k^n is the ground-truth indicator, and $\hat{y}_k^n$ denotes the predicted probability.

To enhance the alignment of the predicted segmented image with respect to the ground-truth mask more rigorously, an additional dice loss item will be introduced.

$$L_{dice} = 1 - \frac{2}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{\hat{y}_k^n y_k^n}{\hat{y}_k^n + y_k^n} \tag{21}$$

In addition to the auxiliary supervision term that gradually conducts semantic representation learning through the intermediate decoder block. Bilinear interpolation or feature combination generates the auxiliary prediction, which has a loss function of:

$$L_{aux} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_k^n \log d_k^n \tag{22}$$

where d_k^n is the aggregate output of the auxiliary head. The total loss of the teacher model optimisation is a weighted sum.

$$L_{\text{total}} = L_{ce} + L_{dice} + L_{aux} \tag{23}$$

3.4.2 Student Optimization

Within the teacher-student model, the optimisation of the Student Network adopts a distillate-based knowledge-distillation approach [49, 50]. Overall Objective: Supervised at both ends, there is also weak-label supervised training to preserve the relationship of class interaction between teachers and students' predictions. Students' targets are different from their Teachers'; They focus more on imparting good knowledge; Retained are the accuracy verification Conditions stipulated in Standard cross-entropy and Dice.

$$L_{\text{hard}} = L_{\text{ce}} + L_{dice} \tag{24}$$

Aligning the distribution of students' model with that of teachers', a soft label loss is defined

as follows: The inter-class relation loss promotes global semantic consistency through the comparison of probability distribution means with Pearson correlations:

$$L_{\text{inter}} = 1 - \text{mean}\left(\text{Pearson}\left(\hat{y}^{\text{T}}_{\text{student}}, \hat{y}^{\text{T}}_{\text{teacher}}\right)\right) \tag{25}$$

where $\text{Pearson}(\cdot)$ is the Pearson function.

Intra-class relation loss model consistency in outputs of a same class via transposing them:

$$L_{\text{intra}} = 1 - \text{mean}\left(\text{Pearson}\left(\hat{y}_{\text{student}}, \hat{y}_{\text{teacher}}\right)\right) \tag{26}$$

4 Experimental Results and Analysis

4.1 Training Environment

SAM2-MAFLNet adopts the teacher-student distillation method of multi-level hierarchical feature learning. A batch size of 4 for training is set. The initial learning rate is set to 6×10^{-4}; The backbones are optimised with a reduced learning rate of 6×10^{-5} to maintain stability in adapting the pre-trained encoders. The new added modules have a high update frequency; however, those in the backbones are cautiously optimised. The remaining parts of the networks use cosine annealing schedules to slowly reduce their learning rates, promoting smoother convergence and enhancing generalisation performance.

4.2 Datasets

To assess the entire proposed framework through experiments on the ISPRS Vaihingen and ISPRS Potsdam datasets. As a rule of thumb, the Vaihingen dataset trains using 15 pictures, evaluates on Image30, and test with 17 others. For Potsdam, 22 tiles were trained; Image 2_10 was reserved for validation; and the rest of the official test Tiles were assessed. Figure 3 shows some typical segmented outcomes from the Vaihingen set.

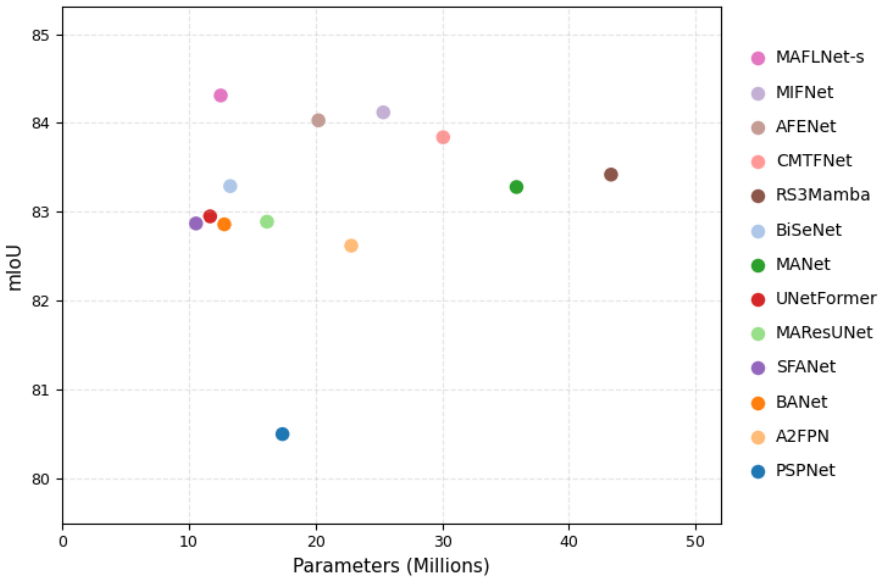

Figure 3: Network performance in experiments using the Vaihingen dataset.

•ISPRS Vaihingen

The ISPRS Vaihingen dataset is a typical test-bed for very-high resolution Urban Semantic Segmentation. There are 33 aerial image tiles, each about 2494×2064 pixels in size and at a ground sampling interval close to 9 cm. Each tile includes near-infrared, red and green channels along with the DSM and nDSM data. In the experiment, we crop these images into small pieces for training; additionally, some standard data-augmentation methods were used in this study to enhance model generalisation performance.

•ISPRS Potsdam

The ISPRS Potsdam dataset consists of 38 Ultra High-Resolution Satellite Imagery whose size is 6000*6000 pixels; The ground sample error range was from 5 cm to as much as several tens times larger than this value. It annotates the six types of semantic classes: impervious Surfaces, Buildings, Low Vegetation, Trees, Cars and Clutter/Background, along with RGB-IR and elevation information. According to the conventional assessment methods, after using RGB input data; We control memory consumption by dividing large images into smaller parts; Randomly flip and perform a mosaic-like enhancement to increase robustness.

4.3 Evaluation Metrics

Overall accuracy (OA), Mean Intersection over Union (mIoU) and mean F1-score are used to evaluate model performance. Combined with the assessment of pixel accuracy, regional overlaps and classes' segmentation accuracy; hence they are typically used for evaluating remote-sensing semantic-segmentation results across various fields.

$$\mathrm{OA} = \frac{\sum_{k=1}^{K}\left(TP_k + TN_k\right)}{\sum_{k=1}^{K}\left(TP_k + FP_k + TN_k + FN_k\right)} \tag{27}$$

$$\mathrm{mIoU} = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k + FP_k + FN_k} \tag{28}$$

$$\mathrm{Precision}_{\mathrm{macro}} = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k + FP_k} \tag{29}$$

$$\mathrm{Recall}_{\mathrm{macro}} = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k + FN_k} \tag{30}$$

where TP_k denotes the number of true-positive pixels for class k, FP_k and FN_k denote false-positive and false-negative pixels for class k, respectively.

4.4 Comparison with State-of-the-Art Methods

4.4.1 Experimental Results on ISPRS Vaihingen Dataset

According to Tables 1 and Figures 4, the developed teacher network MAFLNet-T obtained the best comprehensive results for the ISPRS Vaihingen test set. It has reached 91.92% mF1, 85.06% mIOU and 93.35% OA; Compared with other well-known baselines it outperformed them by 0.68%, 0.94% and 0.69%, respectively. The category-level results also show that, compared to other methods, the proposed one has a much better performance under adverse conditions; For example, it performs well on Sparse Vegetation 85.64% /75.39%(F1/IOU),Trees

(90.08%/(82.88)),Cars (90.13%)/(82.47)%etc, The improvement of this combination makes It More capableofenhancingthe context understanding ability and edge stability Degree simultaneously.

Table 1: Semantical Segmentation Results of The Vaihingen Datasets.

Method	Class F1 / IoU (%)					MF1	MIoU	OA
	Imp.surf.	Building	Low.veg.	Tree	Car	(%)		
PSPNet[29]	95.19/90.81	94.05/88.77	83.37/71.48	89.60/81.15	82.55/70.28	88.95	80.5	91.58
BiSeNet[51]	95.81/91.96	95.30/91.01	83.96/72.35	89.98/81.78	88.50/79.36	90.71	83.29	92.36
BANet[52]	95.80/91.69	95.61/91.48	83.11/71.25	89.57/81.12	88.60/79.54	90.64	83.2	92.12
A2FPN[53]	95.73/91.81	95.27/90.96	83.48/71.64	89.60/81.16	87.33/77.51	90.28	82.62	92.14
MANet[54]	95.77/91.88	95.32/91.06	83.45/71.60	90.02/81.85	88.88/79.99	90.69	83.28	92.25
MAResUNet[55]	95.72/91.78	95.31/91.04	83.67/71.93	89.78/81.46	87.79/78.23	90.45	82.96	92.19
UNetFormer[35]	95.72/91.91	95.39/91.23	83.90/72.33	89.77/81.43	88.84/79.72	90.81	83.45	92.26
SLCNet[56]	95.80/91.92	95.47/91.33	84.13/72.64	89.94/81.71	89.03/80.07	90.92	83.53	92.3
GCDNet[57]	95.84/92.01	95.68/91.72	83.65/71.90	89.79/81.47	89.50/81.00	90.89	83.62	92.36
CMTFNet[58]	95.74/91.84	95.81/91.83	83.88/72.24	90.07/81.93	89.43/80.58	91.03	83.84	92.49
SFANet[59]	95.72/91.86	95.64/91.66	83.72/72.41	90.19/81.93	88.67/74.41	90.98	83.47	92.22
MIFNet[60]	95.87/92.10	96.03/92.02	84.26/72.80	90.10/81.74	89.75/81.40	91.24	84.12	92.66
RS3Mamba[61]	95.91/92.15	95.86/92.06	83.58/71.79	90.85/81.57	89.75/81.59	90.75	83.42	92.26
AFENet[62]	95.69/91.92	95.45/91.33	83.24/71.27	90.25/81.89	88.90/79.90	90.41	82.91	92.11
MAFLNet-T	94.47/93.30	96.45/93.25	85.64/75.39	90.08/82.88	90.13/82.47	91.92	85.06	93.35
MAFLNet-S	95.90/92.12	95.62/91.60	85.05/73.99	90.89/83.30	89.23/80.55	91.34	84.31	92.79

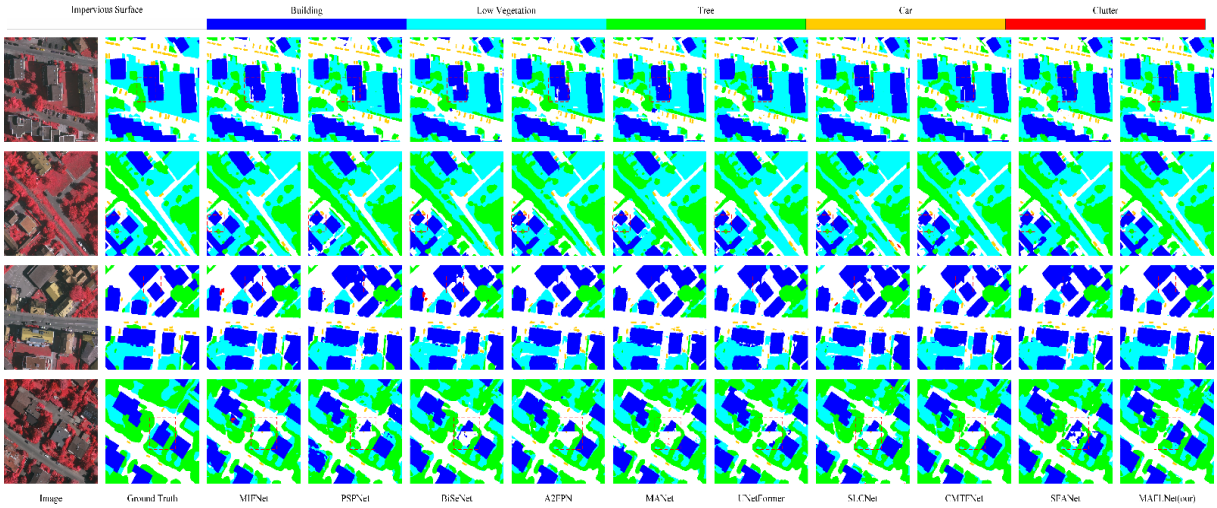

Figure 4: Semantic Segmentation Results for Examples from the Vaihingen Dataset.

MAFLNet-S, a light-weighted student network, also obtains outstanding performance in the Vaihingen dataset with just 12.52M parameters. It has obtained 91.34%, 84.31% mIoU, and 92.79% OAt values compared with some competitors' baseline systems as well as more substantial ones. Based on these results, it can be concluded from the proposed distillation approach that many teachers' representations have been transmitted successfully to students with minimal loss of precision. Overall, the Vaihingen experiments support the significance of the proposed teacher-student Design. MAFLNet-T achieves the best results to date; while MAFLnet-S is deployed easier and has acceptable efficiency-cost trade-off.

4.4.2 Results on the ISPRS Potsdam Dataset

MAFLNET-T has shown superior results to all the others tested in the ISPRS Potsdam dataset. As reported in Table 2 and Figure 5, it reaches 93.64% mF1, 88.21% mIoU, and 92.59% OA. The Improvement is more apparent in subcategories of low vegetation (89.14% / 80.40%, F1/F2) and tree areas (90.37 % / 82.44), respectively.

Table 2: Semantic Segmentation Results of the ISPRS Potsdam Dataset.

Method	Class F1 / IoU (%)					MF1	MIoU	OA
	Imp.surf.	Building	Low.veg.	Tree	Car	(%)		
PSPNet[29]	92.57/86.17	94.29/89.20	86.07/75.55	86.76/76.62	94.45/89.49	90.83	83.41	89.61
BiSeNet[51]	93.77/88.27	96.07/92.43	87.00/76.99	88.39/79.20	96.04/92.39	92.25	85.86	91.07
BANet[52]	93.32/87.48	95.95/92.21	86.65/76.45	88.61/79.54	95.78/91.90	92.06	85.52	90.73
A2FPN[53]	93.33/87.49	95.58/91.54	86.76/76.62	88.22/78.92	95.76/91.86	91.93	85.28	90.73
MANet[54]	93.88/88.47	96.42/93.08	87.16/77.25	88.77/79.81	96.03/92.36	92.45	86.19	91.22
MAResUNet[55]	93.44/87.69	96.19/92.65	86.88/76.60	88.28/79.02	95.73/91.81	92.1	85.59	90.82
UNetFormer[35]	90.86/83.24	93.11/87.10	82.99/70.93	82.08/69.60	93.23/87.32	88.45	79.64	87.03
SLCNet[56]	93.04/86.98	95.84/92.01	86.80/76.68	88.81/79.97	95.61/91.58	92.02	85.42	90.66
GCDNet[57]	93.97/88.62	96.36/92.98	87.13/77.19	88.62/79.56	95.57/91.52	92.33	85.98	91.24
CMTFNet[58]	93.80/88.32	96.54/93.32	87.81/78.28	88.82/79.89	96.15/92.59	92.63	86.48	91.42
SFANet[59]	93.75/88.24	96.46/93.17	86.86/76.77	88.50/79.38	95.87/92.07	92.29	85.93	90.98
MIFNet[60]	94.18/89.00	97.05/94.28	87.31/77.49	89.17/80.46	96.53/93.30	92.85	86.9	91.55
RS3Mamba[61]	93.24/87.33	96.93/92.11	86.37/76.01	88.07/78.69	96.17/92.63	91.95	85.35	90.55
AFENet[62]	94.12/88.89	96.77/93.73	87.36/77.55	88.70/79.70	96.29/92.86	92.65	86.54	91.44
MAFLNet-T	94.83/90.17	97.42/94.98	89.14/80.40	90.37/82.44	96.41/93.07	93.64	88.21	92.59
MAFLNet-S	94.31/89.23	96.71/93.64	88.01/78.58	89.58/81.12	96.13/92.56	92.94	87.02	91.76

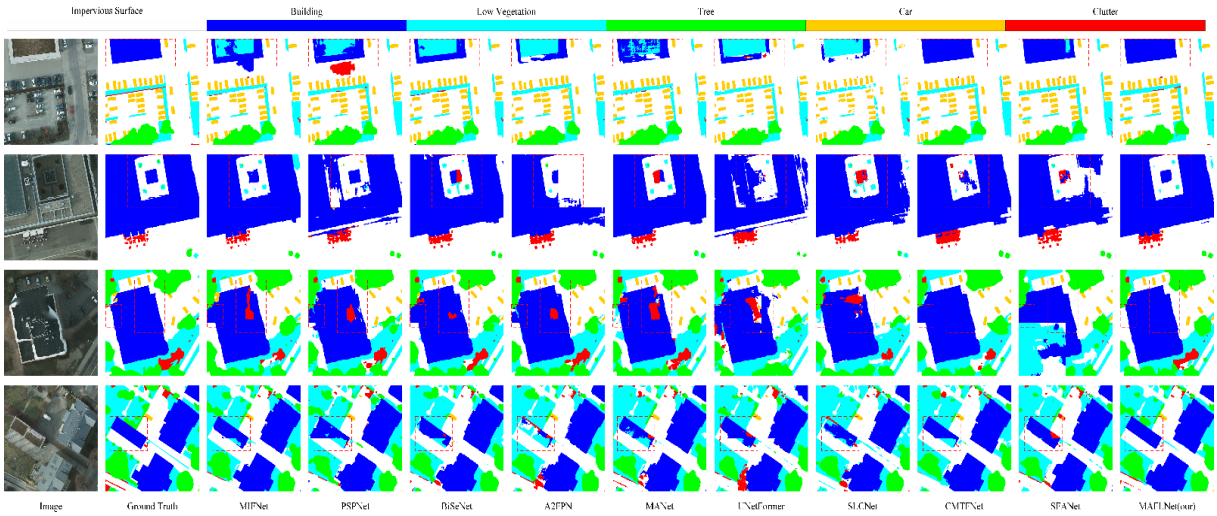

Figure 5: Examples of semantic segmentation outcomes for the Potsdam dataset.

MAFLNET-T can achieve a satisfactory accuracy of 94.83%/90.17%, 97.42%/94.98% in terms of imperviousness/Buildings for most primary urban type areas, respectively. Compared to MIFNet, this study's model has increased mF1 scores by 0.79% and mIoU scores by 1.31%, presenting more accurate road boundary judgments and retaining smaller vehicle shapes relatively better qualitatively (as shown in Figures 5-6).

MAFLNet-S students have stable performance on the Potsdam dataset at 92.94% mF1, 87.02% mIoU, and 91.76% OA using just 12.52 million parameters. Although lighter than the teacher, it is also above several typical baseline models; Its performance remains at a level relatively close to that of the most effective heavy architectures in all experiments.

4.5 Ablation Study

Quantitative assess how much influence, in each major portion separately tested through ablation experiments with ISPRS Vaihingen dataset; MACA mainly improves the long-distance context-aware reasoning ability and channel collaboration in the proposed Design; Meanwhile, LHFEM refines boundaries and reconstructs high-frequency structures with its strengths. Separate and examine one by one; then determine which part of the overall performance enhancement caused it.

Table 3 and Figures 6a, b demonstrate that the ablated model MACK has enhanced performance on both metrics: The precision recall F-measure (mF1) increased by +1.50%, while the mean intersection over union (mean IoU) also improved by +2.42%. Rich Spatial-Channel Interactions help networks address large amounts of ambiguity caused by multiple contexts more robustly.

Table 3: Ablation experiments of key parts in the ISPRS Vaihingen dataset.

Method	MACA	LHFEM	MF1(%)	MIoU(%)	OA(%)
Baseline	-	-	89.95	82.14	92.32
Baseline+MACA	√	-	91.45	84.56	92.90
Baseline+LHFEM	-	√	91.42	84.51	93.02
SAM2-MAFLNet	√	√	91.92	85.66	93.35

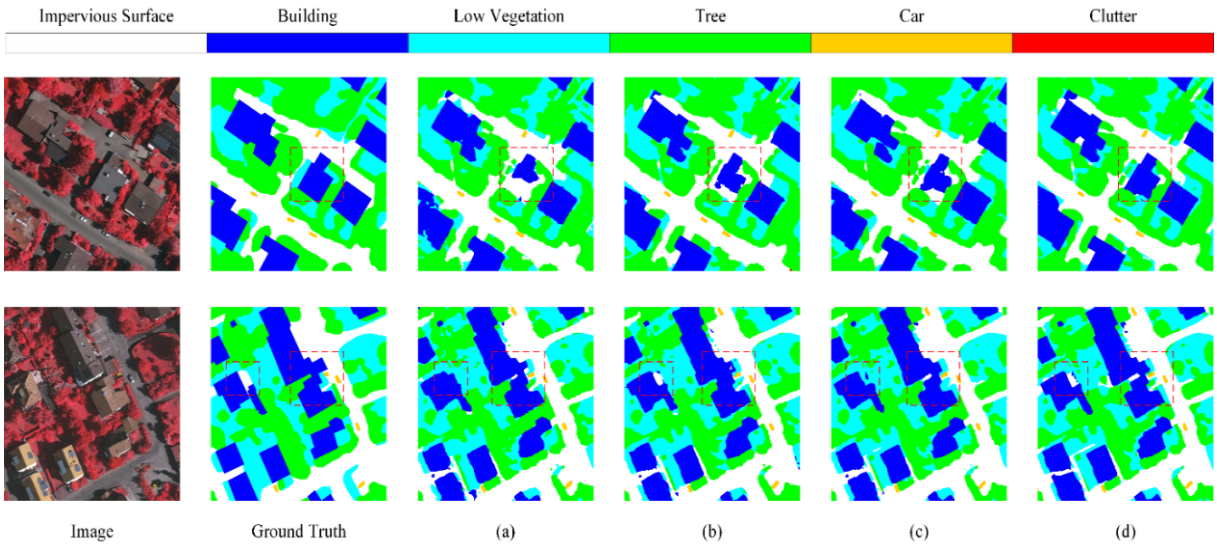

Figure 6: Ablation experiments of MACA and LHFEM module on Vaihingen dataset. (a) Baseline. (b) Baseline+MACA. (c) Baseline+LHFEM. (d) SAM2-MAFLNet.

LHFEM is also effective; its recognition rate achieves up to 91.42%mF1 and 84.51%mIoU. When MACA and LHFEM are used together, the complete SAM2-MAFLNet achieves the best result, namely 91.92% mF1, 85.66% mIoU, and 93.35% OA. Compared with the basic model, it shows an improvement of 1.97 points in mF1 and 3.52 points in mIoU; thus, context aggregation and high-frequency boundary expansion have provided corresponding

improvements.

5 Conclusion

SAM2-MAFLNet is the first work to propose a knowledge-distillation method for remote-sensing semantic-segmentation based on an SAM2-based teacher and a lightweight student network. MACA combined with long-distance range channel models and LHFEM boundary-aware high-frequency improvements can improve both context-based reasoning and details retained at very high resolutions of very-high-resolution images.

The experiments conducted on the ISPRS Vaihingen and Potsdam datasets show that our method has achieved good performance and stable operation; A student distilled from just 12.52M parameters is also highly accurate. Both MACA and LHFEM are required, or at least partially supported by the experiment's results. Future Research Directions: Cross-modal Generalization, Other Potential Efficiency Optimisations And Additional Tests Of The Remote Sensing Benchmarks.

About the Author

Tang Xiaoliang, born in Fuxin City, Liaoning Province, China in 1980. He received his Bachelor's Degree from Dalian University of Technology in China. Currently teaching in the School of Computer Science and Technology at Zhejiang University of Science and Technology. Primary Research Field of His is Computer Vision and Deep Learning.

References

[1] Sun, W., & Wang, R. (2018). Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM. IEEE Geoscience and Remote Sensing Letters, 15, 474-478.

[2] Zhang, S., Dai, X., Li, J., Gao, X., Zhang, F., Gong, F., Lu, H., Wang, M., Ji, F., Wang, Z., et al. (2022). Crop classification for UAV visible imagery using deep semantic segmentation methods. Geocarto International, 37, 10033-10057.

[3] Han, Z., Li, X., Wang, X., Wu, Z., & Liu, J. (2025). Building segmentation in urban and rural areas with MFA-Net: A multidimensional feature adjustment approach. Sensors, 25.

[4] Khan, M. Z., Gajendran, M. K., Lee, Y., & Khan, M. A. (2021). Deep neural architectures for medical image semantic segmentation: Review. IEEE Access, 9, 83002-83024.

[5] Peng, C., Wang, N., Li, J., & Gao, X. (2021). Soft semantic representation for cross-domain face recognition. IEEE Transactions on Information Forensics and Security, 16, 346-360.

[6] Zhang, Y., Lu, C., Wang, J., & Du, F. (2024). A large-scale extraction framework for mapping urban informal settlements using remote sensing and semantic segmentation. Geocarto International, 39, 2345135.

[7] Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., Yokoya, N., Li, H., Ghamisi, P., Jia,

X., et al. (2024). SpectralGPT: Spectral remote sensing foundation model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46, 5227-5244.

[8] Yang, R., Zheng, C., Wang, L., Zhao, Y., Fu, Z., & Dai, Q. (2023). MAE-BG: Dual-stream boundary optimization for remote sensing image semantic segmentation. Geocarto International, 38, 2190622.

[9] Hafner, S., Nascetti, A., Azizpour, H., & Ban, Y. (2022). Sentinel-1 and Sentinel-2 data fusion for urban change detection using a dual stream U-Net. IEEE Geoscience and Remote Sensing Letters, 19, 1-5.

[10] Li, Z., Chen, B., Wu, S., et al. (2024). Deep learning for urban land use category classification: A review and experimental assessment. Remote Sensing of Environment, 311, 114290.

[11] He, D., Liu, X., & Shi, Q. (2025). Visual-language reasoning segmentation (LARSE) of function-level building footprint across Yangtze River Economic Belt of China. Sustainable Cities and Society, 108, 106439.

[12] Ma, J., Yan, L., Chen, B., & Zhang, L. (2025). A tree crown segmentation approach for unmanned aerial vehicle remote sensing images on field programmable gate array (FPGA) neural network accelerator. Sensors, 25.

[13] Chen, B., Tong, A., Wang, Y., Zhang, J., Yang, X., & Im, S. K. (2025). LKAFFNet: A novel large-kernel attention feature fusion network for land cover segmentation. Sensors, 25.

[14] Zuo, Z., Shuai, B., Wang, G., Liu, X., Wang, X., Wang, B., & Chen, Y. (2016). Learning contextual dependence with convolutional hierarchical recurrent neural networks. IEEE Transactions on Image Processing, 25, 2983-2996.

[15] Chen, H., Qi, Z., & Shi, Z. (2022). Remote sensing image change detection with transformers. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-14.

[16] Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Tomizuka, M., Keutzer, K., & Vajda, P. (2020). Visual transformers: Token-based image representation and processing for computer vision. arXiv, abs/2006.03677.

[17] Guo, C., Fan, B., Zhang, Q., Xiang, S., & Pan, C. (2020). AugFPN: Improving multi-scale feature learning for object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 12592-12601).

[18] Ma, Y., Chen, S., Ermon, S., & Lobell, D. B. (2024). Transfer learning in environmental remote sensing. Remote Sensing of Environment, 301, 113924.

[19] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W. Y., et al. (2023). Segment anything. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 3992-4003).

[20] Ravi, N., Gabeur, V., Hu, Y. T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland,

C., Gustafson, L., et al. (2024). Sam 2: Segment anything in images and videos. arXiv, 2408.00714.

[21] Zhou, X., Liang, F., Chen, L., Liu, H., Song, Q., Vivone, G., & Chanussot, J. (2024). MeSAM: Multiscale enhanced segment anything model for optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 62, 1-15.

[22] Gui, B., Bhardwaj, A., & Sam, L. (2024). Evaluating the efficacy of segment anything model for delineating agriculture and urban green spaces in multiresolution aerial and spaceborne remote sensing images. Remote Sensing, 16, 414.

[23] Xiong, X., Wu, Z., Tan, S., Li, W., Tang, F., Chen, Y., Li, S., Ma, J., & Li, G. (2024). SAM2-UNet: Segment Anything 2 makes strong encoder for natural and medical image segmentation. CoRR, abs/2408.08870.

[24] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97 (pp. 2790-2799).

[25] Qiu, Z., Hu, Y., Li, H., & Liu, J. (2023). Learnable ophthalmology SAM. arXiv preprint, arXiv:2304.13425.

[26] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3431-3440).

[27] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (pp. 234-241). Cham.

[28] Wang, Q., Xie, J., Zuo, W., Zhang, L., & Li, P. (2021). Deep CNNs meet global covariance pooling: Better representation and generalization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43, 2582-2597.

[29] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 6230-6239).

[30] Al-Wahaibi, S. S. S., & Lu, Q. (2023). Improving convolutional neural networks for fault diagnosis by assimilating global features. In Proceedings of the 2023 American Control Conference (ACC) (pp. 4730-4734).

[31] Bertasius, G., Shi, J., & Torresani, L. (2016). Semantic segmentation with boundary neural fields. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3602-3610).

[32] Fu, J., Liu, J., Wang, Y., & Lu, H. (2017). Densely connected deconvolutional network for semantic segmentation. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP) (pp. 3085-3089).

[33] Mou, L., Hua, Y., & Zhu, X. X. (2019). A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 12408-12417).

[34] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 6000-6010).

[35] Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., & Atkinson, P. M. (2022). UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 190, 196-214.

[36] Zheng, X., Luo, Y., Fu, C., Liu, K., & Wang, L. (2024). Transformer-CNN cohort: Semi-supervised semantic segmentation by the best of both students. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA) (pp. 11147-11154).

[37] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. arXiv, cs.CV/2103.14030.

[38] Ghiasi, G., & Fowlkes, C. C. (2016). Laplacian pyramid reconstruction and refinement for semantic segmentation. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), Computer Vision – ECCV 2016 (pp. 519-534). Cham.

[39] Wang, C., Chen, S., Mi, D., Chen, Y., Zhang, Y., & Li, Y. (2025). SWDL: Stratum-wise difference learning with deep Laplacian pyramid for semi-supervised 3D intracranial hemorrhage segmentation. arXiv, abs/2506.10325.

[40] Srivastava, L., & Gakhar, I. (2025). LAqua: Laplacian pyramids for aquatic segmentation (Student abstract). In Proceedings of the AAAI Conference on Artificial Intelligence, 39, 29498-29500.

[41] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint, arXiv:1503.02531.

[42] Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2014). FitNets: Hints for thin deep nets. arXiv preprint, arXiv:1412.6550.

[43] Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., & Wang, J. (2019). Structured knowledge distillation for semantic segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2599-2608).

[44] Shu, C., Liu, Y., Gao, J., Yan, Z., & Shen, C. (2021). Channel-wise knowledge distillation for dense prediction. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 5291-5300).

[45] Liu, L., Huang, Q., Lin, S., Xie, H., Wang, B., Chang, X., & Liang, X. (2021). Exploring inter-channel correlation for diversity-preserved knowledge distillation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 8251-8260).

[46] Chen, P., Liu, S., Zhao, H., & Jia, J. (2021). Distilling knowledge via knowledge review. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5006-5015).

[47] Feng, Y., Sun, X., Diao, W., Li, J., & Gao, X. (2021). Double similarity distillation for semantic image segmentation. IEEE Transactions on Image Processing, 30, 5363-5376.

[48] Yang, C., Zhou, H., An, Z., Jiang, X., Xu, Y., & Zhang, Q. (2022). Cross-image relational knowledge distillation for semantic segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 12309-12318).

[49] Xia, Y., Xu, Y., Wang, C., & Stilla, U. (2021). VPC-Net: Completion of 3D vehicles from MLS point clouds. ISPRS Journal of Photogrammetry and Remote Sensing, 174, 166-181.

[50] Huang, T., You, S., Wang, F., Qian, C., & Xu, C. (2022). Knowledge distillation from a stronger teacher. In Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS). arXiv preprint, arXiv:2205.10536.

[51] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., & Sang, N. (2018). BiSeNet: Bilateral segmentation network for real-time semantic segmentation. arXiv, cs.CV/1808.00897.

[52] Wang, L., Li, R., Wang, D., Duan, C., Wang, T., & Meng, X. (2021). Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. Remote Sensing, 13.

[53] Li, R., Wang, L., Zhang, C., Duan, C., & Zheng, S. (2022). A2-FPN for semantic segmentation of fine-resolution remotely sensed images. International Journal of Remote Sensing, 43, 1131-1155.

[54] Li, R., Zheng, S., Zhang, C., Duan, C., Su, J., Wang, L., & Atkinson, P. M. (2022). Multiattention network for semantic segmentation of fine-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-13.

[55] Li, R., Zheng, S., Duan, C., Su, J., & Zhang, C. (2022). Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. IEEE Geoscience and Remote Sensing Letters, 19, 1-5.

[56] Yu, D., & Ji, S. (2023). Long-range correlation supervision for land-cover classification from remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 61, 1-14.

[57] Cui, J., Liu, J., Wang, J., & Ni, Y. (2023). Global context dependencies aware network for efficient semantic segmentation of fine-resolution remoted sensing images. IEEE Geoscience and Remote Sensing Letters, 20, 1-5.

[58] Wu, H., Huang, P., Zhang, M., Tang, W., & Yu, X. (2023). CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 61, 1-12.

[59] Hwang, G., Jeong, J., & Lee, S. J. (2024). SFA-Net: Semantic feature adjustment network

for remote sensing image segmentation. Remote Sensing, 16.

[60] Fan, J., Li, J., Liu, Y., & Zhang, F. (2024). Frequency-aware robust multidimensional information fusion framework for remote sensing image segmentation. Engineering Applications of Artificial Intelligence, 129, 107638.

[61] Ma, X., Zhang, X., & Pun, M. O. (2024). RS3Mamba: Visual state space model for remote sensing images semantic segmentation. arXiv, cs.CV/2404.02457.

[62] Gao, F., Fu, M., Cao, J., Dong, J., & Du, Q. (2025). Adaptive frequency enhancement network for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 1-1.