scientific reports

Check for updates

OPEN

SAM2-ARAFNet: adapting SAM2 with an attention-enhanced residual ASPP fusion network for high-resolution remote sensing semantic segmentation

Wenbin Shi[1], Jiayin Ding[2,3], Jingsheng Lei[2,3] & Yong Ji[2]✉

High-resolution remote sensing image segmentation plays a crucial role in fields such as environmental surveillance, disaster impact analysis, and spatial resource management, yet the pronounced variability within classes, intricate scene structures, and substantial computational burden of modern deep models often impede their practical use. To mitigate these limitations, this study introduces SAM2-ARAFNet, a segmentation framework derived from Segment Anything Model 2 (SAM2) and equipped with lightweight adapter modules for efficient parameter tuning, together with an Attention-Enhanced Residual Atrous Spatial Pyramid Pooling (ResASPP) component that enriches multi-scale semantic representation. For deployment on resource-limited platforms, a tailored distillation strategy is further employed to compress the fine-tuned SAM2 model into a compact student network based on EfficientNet_b0. Experiments conducted on the ISPRS Vaihingen and Potsdam benchmarks demonstrate clear performance gains: SAM2-ARAFNet attains mIoU values of 85.43% and 87.44%, exceeding widely used baselines such as PSPNet by 4.93% and 4.03%. In addition, the distilled student model reduces parameters by 97% (from 222.98 M to 6.68 M) while preserving more than 99% of the teacher network's accuracy, illustrating its capability to deliver high-quality segmentation with markedly improved computational efficiency, and confirming its suitability for edge-focused remote sensing scenarios.

With the continuous advancement of satellite-based Earth observation, high-resolution remote sensing imagery has become an essential tool across numerous application domains. These data support a wide spectrum of tasks, including cloud identification, assessments of urban infrastructure, agricultural resource management, and analysis of road and land-surface conditions[1–3]. Cloud detection, for example, represents a fundamental preprocessing step in many remote sensing workflows, ensuring the reliability of downstream interpretation[4]. In urban studies, remote sensing facilitates the evaluation of buildings, transportation networks, and other critical structures, enabling informed decision-making for maintenance and planning[5]. Similarly, in agricultural applications, remote sensing imagery contributes to monitoring vegetation status, analyzing spatial patterns of land utilization, and improving irrigation strategies, thereby enhancing overall agricultural productivity[6].

Despite the extensive use of remote sensing data in numerous application domains, achieving accurate remote sensing image segmentation (RSIS) remains difficult owing to the inherent complexity of the imagery. Remote sensing scenes frequently display pronounced variations within the same category, only subtle distinctions between different classes, and quality degradation issues such as noise, blurring, or occlusion[7]. These characteristics, combined with the limited availability of large, well-annotated datasets and the challenge of building models capable of generalizing across diverse geographic conditions, continue to impede robust segmentation performance. Although recent advances have improved segmentation accuracy, constraints persist in terms of robustness, adaptability, and the reliance on manual annotation or task-specific model refinements.

To cope with these difficulties, a broad spectrum of segmentation methodologies has been proposed, which are commonly grouped into threshold-driven methods, classical machine learning techniques, and deep

[1]School of Computer, Hangzhou Dianzi University, Hangzhou 310018, China. [2]School of Computer Science and Technology, Zhejiang University of Science and Technology, Hangzhou 310023, China. [3]Jiayin Ding and Jingsheng Lei contributed equally to this work. ✉email: 222308855024@zust.edu.cn

learning-based solutions[8–10]. Threshold-oriented approaches depend on handcrafted heuristics and expert-defined criteria, making them labor-intensive and hard to extend across different settings[11,12]. Traditional machine learning algorithms, including Support Vector Machines (SVMs) and Random Forests (RF), have shown effectiveness in several remote sensing applications[13,14]; however, their transferability is often restricted. SVM performance is highly influenced by kernel choice[15], while RF models trained in one region may exhibit poor portability to other regions due to their sensitivity to the subset of features used during node splitting[16]. Deep learning–based methods possess stronger representation capacity, yet they also face obstacles arising from the scarcity of labeled data and the need to adapt to heterogeneous remote sensing domains[17]. To alleviate these issues, approaches grounded in domain adaptation and the integration of prior knowledge have gained traction, enabling models to leverage available information more effectively and capture domain-specific structural cues, thereby enhancing generalization and robustness across varied remote sensing scenarios[18–20].

Recent advances in vision foundation models (VFMs)[21–24] have further transformed the landscape of image segmentation. Among them, the Segment Anything Model (SAM1)[21] and its improved version, SAM2[22], represent significant milestones. SAM2 enhances SAM1 by leveraging a larger training corpus and architectural refinements, yet it still produces class-agnostic masks in the absence of human-provided prompts. This characteristic poses a substantial challenge when transferring SAM2 to downstream applications requiring class-aware or task-specific segmentation. Enhancing its adaptability to such tasks remains an active area of research. Existing efforts to tailor SAM1 or SAM2 for downstream use include parameter-efficient tuning via adapters[25,26] and the incorporation of additional conditioning signals such as text prompts[27–29] or in-context examples[30,31].

Despite the notable progress enabled by SAM2, its direct application to remote sensing image segmentation (RSIS) remains non-trivial. The inherent characteristics of remote sensing data introduce several obstacles that limit SAM2's effectiveness in this domain. The main challenges can be summarized as follows:

(1) Domain gap: SAM2 is pretrained predominantly on natural-image datasets, leading to a mismatch when transferred to remote sensing imagery. Remote sensing scenes differ substantially from natural images in terms of spatial resolution variability, atmospheric interference, viewing geometry, and object appearance, all of which hinder SAM2's ability to generalize effectively.
(2) Loss of global context: Due to the extremely large spatial extent of high-resolution remote sensing images, both training and inference commonly rely on patch-wise processing. This practice inevitably discards important long-range contextual dependencies present in the full-scene imagery, weakening the model's ability to capture holistic spatial relationships.
(3) Resource constraints: Conventional deep neural networks often require substantial computational power and memory, making them difficult to deploy on resource-limited platforms such as UAVs, satellites, and edge devices.

Without cloud-based computation, running these models locally becomes impractical, which restricts their usability in real-time or field-based applications.

For the first difficulty, drawing upon ideas from SAM2UNet[32], we incorporated lightweight adapter components to adjust the parameters of SAM2, thereby improving its suitability for remote sensing image segmentation tasks. To overcome the second limitation, we integrated the multi-level feature representation capability of SAM2 with an attention-augmented residual Atrous Spatial Pyramid Pooling (ASPP) structure, forming a newly designed architecture named SAM2-ARAFNet. Addressing the final challenge, SAM2-ARAFNet was further adopted as the teacher network, and a distillation framework was introduced to derive a compact and computationally efficient student model tailored for edge-oriented deployment.

The major contributions of this work are summarized as follows:

- A novel architecture, SAM2-ARAFNet, is introduced by employing SAM2 as the feature encoder and embedding an attention-enhanced residual ASPP module. This design enriches multi-scale contextual modeling and strengthens feature discrimination, enabling accurate segmentation across varied remote sensing conditions.
- Comprehensive evaluations on the ISPRS Vaihingen and Potsdam datasets show that both the teacher network and the distilled student model achieve significant accuracy improvements, outperforming a range of contemporary segmentation frameworks.
- A distillation strategy is developed to support lightweight deployment. The resulting student model exhibits a substantial reduction in parameter count—from 222.98 M in the teacher model to 6.68 M—yielding an overall compression of 97% without severely compromising segmentation performance.

Related work

Multi-scale remote sensing semantic segmentation

In recent years, advances in deep learning have markedly transformed the development of remote sensing semantic segmentation. A core technique driving this progress is multi-scale feature extraction, which enables models to capture fine-grained details and coarse-grained contextual information simultaneously. Existing segmentation strategies are commonly organized into three broad categories: Multi-scale feature extraction method based on convolutional neural networks (CNNs),Multi-scale feature extraction method based on Transformer,and Multi-scale feature extraction method based on hybrid architecture.

Multi-scale feature extraction method based on CNN

Early multi-scale feature extraction primarily relied on the hierarchical structure of CNNs and multi-scale pooling strategies. For instance, PSPNet[33] aggregates global features at different scales through a pyramid pooling module, effectively alleviating classification ambiguity for large-scale objects. CRENet[34] proposes a local feature

alignment mechanism to mitigate the semantic gap between shallow and deep features in encoder-decoder architectures. Shallow features hold fine spatial details but lack high-level semantics, whereas deep features have strong semantic discriminability yet suffer from resolution loss. The alignment module bridges this discrepancy via adaptive transformation and similarity constraint, preserving both spatial precision and semantic accuracy during fusion. In contrast, CRFNet[35] fuses multiscale contextual information with stochastic potential functions to enhance segmentation accuracy: its multiscale module captures hierarchical features for objects of varying sizes, while stochastic potentials model pixel-wise probabilistic dependencies to refine boundaries and suppress segmentation noise.

Although CNNs excel at modeling fine-grained spatial structures, their inherent locality—stemming from the fixed receptive field of convolution kernels—makes it challenging to capture broader contextual information. This is an essential prerequisite for accurate interpretation of large-scale remote sensing scenes, where target semantics often depend on long-range spatial correlations between different ground objects.

Multi-scale feature extraction method based on transformer

The emergence of Transformers has provided new perspectives for multi-scale feature modeling. CG-Swin[36] employs Swin Transformer as an encoder and captures multi-scale semantic information effectively through adaptive scaling of window attention, but it faces significant computational overhead when processing high-resolution remote sensing images. SSDT[37] designs a scale-decoupled Transformer module to separate semantic patterns at different spatial resolutions, yet its ability to extract features of small-scale objects relies on predefined scale divisions, limiting its flexibility.

These Transformer-based architectures capitalize on the unique capability of the self-attention mechanism to capture extensive contextual interactions and long-range feature dependencies across the entire feature map, without being constrained by the fixed receptive field inherent in convolutional operations. By computing the similarity between every pair of feature tokens and assigning adaptive attention weights, these architectures can effectively model the spatial correlations between distant but semantically related ground objects—such as the connection between scattered residential buildings and surrounding road networks, or the correlation between fragmented wetland patches and adjacent water bodies. This characteristic is particularly advantageous when dealing with the structural complexity of remote sensing imagery, where target objects often exhibit large variations in size, irregular spatial distributions, and intricate inter-object relationships that are difficult for traditional convolutional neural networks to fully characterize.

Multi-scale feature extraction method based on hybrid architecture

In recent years, hybrid architectures have become a mainstream direction for multi-scale feature extraction, with their core concept being the integration of CNN's fine-grained local features with Transformer's global scale modeling capability. Representations[38,39] have recently attracted increasing attention. UNetFormer [38] integrates the local feature extraction capability of CNNs with the global context modeling strength of Transformers, fusing multi-scale features through a U-shaped architecture. However, it does not adequately account for differences in semantic importance among features at different scales during fusion, leading to persistent scale confusion issues.whereas CMTFNet[39] adopts a hybrid fusion strategy to integrate CNN local descriptors with Transformer global semantics. Its CNN branch extracts fine-grained features, the Transformer branch captures long-range dependencies via self-attention, and a cross-branch alignment module reconciles discrepancies for boosted remote sensing analysis performance.

Building upon the strengths of this hybrid direction, our approach further incorporates the pretrained Segment Anything Model 2 (SAM2)—a vision foundation model trained on a large-scale mask corpus—thereby improving generalization capability and adaptability beyond conventional architectures.

Segment anything model

The Segment Anything Model (SAM)[21] represents a large-scale segmentation paradigm built around prompt-based interaction, capable of generating class-agnostic masks in response to various forms of user input. Benefiting from training on an extensive corpus comprising millions of images and a vast collection of annotated masks, SAM demonstrates strong generalization ability in natural-image scenarios. SAM2[22] further expands this framework by extending its applicability to both static imagery and video sequences, enabling coherent prompt-driven segmentation across time.

A growing body of work has investigated how to adapt SAM and SAM2 to task-specific domains. For instance, Wu et al.[40] introduced adapter-based refinements to enhance SAM's suitability for medical image segmentation, while Chen et al.[25] incorporated domain knowledge to tackle challenges such as shadowed or camouflaged object identification. Guo et al.[41] presented MDSAM, which improves salient object detection by leveraging multi-scale and hierarchical features. Despite these advances, most adaptations remain tailored to natural-image resolutions and exhibit limited effectiveness when applied to high-resolution remote sensing imagery, where precise boundary delineation and fine-grained spatial detail are essential.

These limitations motivate the central objective of this study: adapting SAM2 to remote sensing semantic segmentation through parameter-efficient tuning and enhanced multi-scale contextual modeling, thereby narrowing the gap between vision foundation models and remote sensing applications.

Knowledge distillation

Knowledge distillation (KD) refers to techniques that transfer representational or predictive ability from a high-capacity network to a more compact model in order to improve computational efficiency without substantial performance loss. Since its introduction by Hinton et al.[42] and subsequent extensions in numerous studies[43–45], KD has evolved into a widely used paradigm for model compression. Current formulations generally obtain

supervisory signals from a teacher network, and related literature can be broadly viewed as either distilling information from a stronger teacher model or leveraging additional auxiliary cues to enhance the student's learning process. This work focuses on distillation strategies tailored for semantic segmentation, where dense spatial predictions and structure-aware consistency impose challenges beyond those encountered in image-level classification.

A number of studies have explored ways to improve the effectiveness and efficiency of knowledge distillation, particularly when transferring information from complex teacher networks. Representative works include SKD[46], IFDV[47], and CIRKD[48]. SKD[46] emphasizes structured knowledge transfer that aligns with the dense prediction characteristics of segmentation tasks. IFDV[47] encourages the student network to preserve inter-class relational distances present in the teacher's feature space. CIRKD[48] extends this concept by modeling pixel-level and region-level structural dependencies across complete images.

Other research focuses on distillation strategies that rely on auxiliary information. For example, LGD[49] and LG3D[50] introduce label-guided distillation through manually crafted label encoders and mapping functions, primarily within object detection pipelines. In contrast, the method proposed in this work does not require such label encoders or mapping modules for semantic segmentation. Additionally, KD-Net[51] presents a framework for transferring knowledge from multi-modal sources to a single modality in medical image segmentation, demonstrating the versatility of distillation in cross-modal scenarios.

Methodology

Building on the advancements of SAM2Unet[32], this paper introduces SAM2-ARAFNet, a model designed for Remote Sensing Semantic Segmentation. It is specifically crafted to process high-resolution images, using adapters for parameter-efficient fine-tuning (PEFT) while maintaining the feature extraction capabilities of the pre-trained encoder.

As shown in Fig. 1, Due to the limited availability of remote sensing datasets poses challenges for retraining Hiera (hierarchical transformer)[52] from scratch, potentially diminishing its feature extraction capabilities. To maintain its effectiveness, we implement a frozen backbone strategy during training. High-resolution remote sensing images are characterized by significant noise and heterogeneity. To facilitate efficient fine-tuning while ensuring adaptability to remote sensing tasks, SAM2-ARAFNet incorporates the adapter design from SAM2-Unet. Adapters are integrated before the multiscale modules of the Hiera encoder, enabling dynamic recalibration of feature distributions and minimizing the number of tunable parameters. SAM2-ARAFNet employs a frozen encoder strategy, with its PEFT approach akin to low-rank adaptation[53] (LoRA). While LoRA primarily targets the attention mechanism within Transformer architectures, SAM2-ARAFNet positions its adapter module

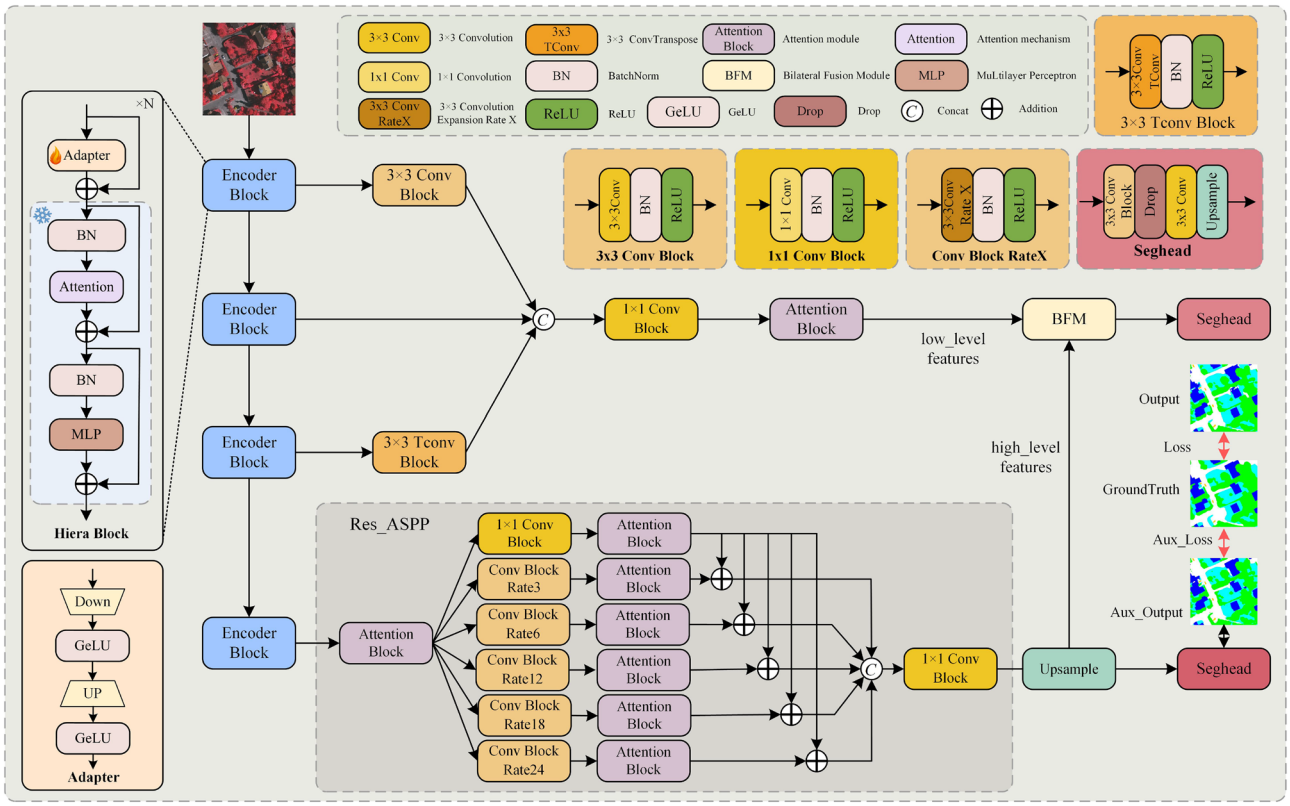

Fig. 1. Overall architecture of the proposed SAM2-ARAFNet model, illustrating the integration of the SAM2 backbone with the ARAFNet module for Remote Sensing Semantic Segmentation, and Attention Block shown in Fig. 2.

before the multi-scale modules of the Hiera encoder to refine feature mapping and enhance model adaptability, optimizing it for remote sensing label tasks.

In the decoder stage, SAM2-ARAFNet incorporates an attention-enhanced residual ASPP mechanism to effectively capture multi-scale contextual information. This approach allows the model to focus on relevant features across different spatial dimensions, improving segmentation accuracy by refining boundaries and enhancing feature representation. The attention mechanism selectively emphasizes important regions, ensuring that the network can adaptively weigh the significance of various features, thereby optimizing the overall performance in complex visual tasks.

Encoder and adapter design

Hiera is a Transformer backbone network engineered for efficient visual feature extraction. Unlike conventional convolutional neural networks (CNNs), Hiera captures both local details and global semantics through hierarchical feature modeling. Its layered computation strategy minimizes computational costs while preserving robust representational capabilities. Our model employs the SAM2-pretrained Hiera Transformer as an encoder, leveraging its multi-scale feature learning to produce high-quality segmentation features.

The overall structure of Hiera is illustrated in Fig. 1. Given the substantial number of parameters in the SAM2-pre-trained Hiera backbone, updating all weights directly would incur considerable computational overhead and risk overfitting to limited remote sensing datasets. To enhance parameter efficiency while maintaining the pre-trained encoder's feature extraction capability, we strictly follow a frozen backbone strategy: the entire Hiera backbone (all layers and parameters) is set to non-trainable, and only compact adapter modules (inserted ahead of each multi-scale stage of Hiera) are trainable. Each adapter is implemented as a bottleneck-style transformation: it first projects features into a lower-dimensional space via a linear layer with GeLU activation [53], then restores the dimensionality through another linear layer with GeLU. This design dynamically recalibrates feature distributions to adapt to remote sensing data, while minimizing the volume of trainable parameters.

Decoder with ResASPP and bilateral fusion

As shown in Fig. 1,the decoder's structure can be divided into three main components: low-level feature processing, high-level feature processing, and a fusion module. Low-level feature processing is responsible for integrating and managing the first three layers of the decoder's output, ensuring effective transmission and utilization of foundational information. High-level feature processing focuses on the final layer of the decoder's output, utilizing our proposed attention-enhanced residual ASPP module to refine the feature representation and enhance semantic information capture. Ultimately, these processed features are combined through the fusion module to achieve a more precise and efficient decoder functionality.

During low-level feature refinement, we design a multi-branch aggregation mechanism that strengthens spatial detail modeling and improves the descriptive quality of early-layer representations. The decoder's first-layer output is first processed with a 3×3 convolution to capture fine-grained local cues, while the third-layer output is passed through a 3×3 transposed convolution to restore spatial resolution and recover structural details. The resulting feature maps from the first and third layers are then integrated with the second-layer output to form a unified multi-scale representation, after which an additional 3×3 convolution reorganizes the fused feature space to reduce redundancy and promote more coherent information flow. To further refine the merged representation, an attention module adaptively adjusts channel responses, emphasizing informative components and weakening noise, ultimately producing a more discriminative low-level feature description.

In the high-level feature extraction phase, the incoming representations are first refined through a CBAM module, as illustrated in Fig. 2, which adaptively emphasizes texture-, color-, and structure-related cues. The enhanced features then pass through a cascade of five dilated convolution blocks, where each block consists of an atrous convolution paired with a CBAM refinement step and is accompanied by a residual pathway constructed with 1-1 convolutions and additional attention modules. The dilation settings for these blocks are configured to 3, 6, 12, 18, and 24, employing 3-3 kernels to capture progressively expanding receptive fields. To gather multi-scale contextual signals, the outputs of the dilated branches are merged with the residual pathway through element-wise addition, yielding five intermediate representations that collectively form a CBAM-augmented

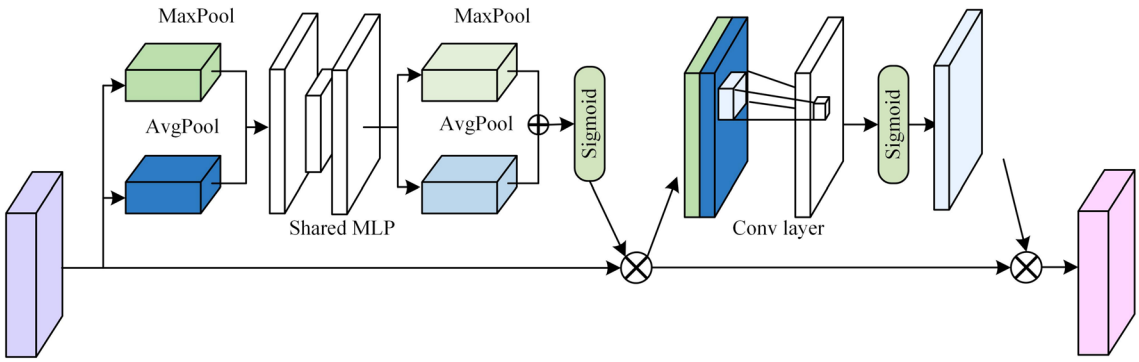

Fig. 2. Architecture of block attention module.

residual ASPP structure. These intermediate outputs are subsequently concatenated along the channel axis, and a concluding 1-1 convolution is applied to integrate the fused information, generating the final set of high-level feature embeddings.

To integrate representations from different semantic levels, a Bilateral Fusion Module (BFM) is introduced. As illustrated in Fig. 3, the module accepts two feature streams and processes them symmetrically through dual interaction branches. In the first branch, the initial input is refined by depthwise and standard convolutions, and its response is modulated by a sigmoid-generated gating signal derived from the second branch. Conversely, the second branch performs an analogous sequence of convolutions on the other input and applies a gating operation based on the output of the first branch. The two modulated outputs are then combined through element-wise addition, after which a convolution layer followed by a ReLU nonlinearity produces the fused feature representation.

Loss function

Teacher training methods

In this study, remote sensing semantic segmentation is treated as a dense prediction problem in which each pixel is assigned a semantic category. To optimize the teacher network under this formulation, a multi-component loss function is adopted, consisting of a standard pixel-level cross-entropy term, a Dice loss that promotes consistency in region-level predictions, and an auxiliary cross-entropy loss that further guides the learning of intermediate representations.

The cross-entropy term serves as the primary supervision signal by penalizing the mismatch between the categorical labels and the network's predicted probability outputs through the negative log-likelihood, formulated as

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

where N denotes the number of samples and K is the total set of semantic classes. The variables y_k^n and $\hat{y}_k^n$ indicate the reference label and the model's estimated probability for class k at the n-th pixel.

To reinforce consistency between the predicted maps and the annotated masks, a Dice-based objective is additionally employed. This term encourages greater correspondence in spatial regions by promoting agreement in foreground assignment, and is defined as

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

To guide the intermediate layers of the decoder toward learning more discriminative representations, an auxiliary prediction branch is incorporated. This branch fuses the outputs from different decoder stages via bilinear upsampling and element-wise combination, and its supervision signal is provided through an additional cross-entropy objective formulated as

$$L_{\text{aux}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_k^n \log d_k^n, \tag{3}$$

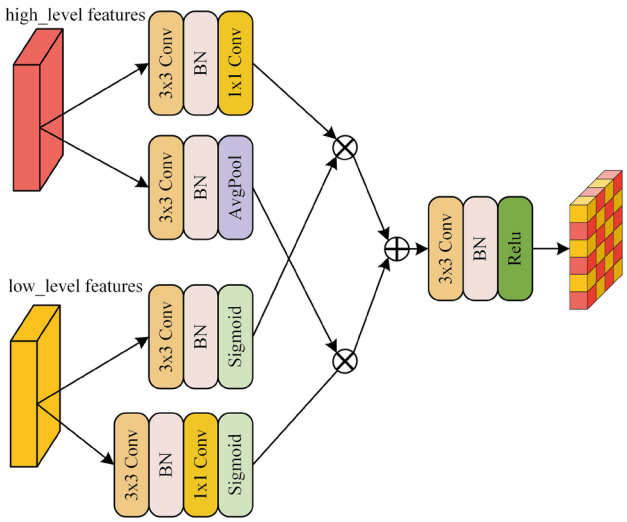

Fig. 3. Architecture of the proposed BFM module.

where d_k^n denotes the auxiliary prediction corresponding to class k at pixel n. The overall optimization target for the teacher network integrates all components into a unified objective:

$$L_{\text{total}} = L_{\text{ce}} + L_{\text{dice}} + \alpha L_{\text{aux}}, \tag{4}$$

with α serving as the balancing coefficient for the auxiliary term. In our implementation, α is set to 0.4, consistent with prior practice in UNetFormer[54].

Students training methods

Within the teacher–student learning paradigm, the student network is optimized using a distillation scheme derived from the DIST[55] formulation, which enables the transfer of semantic representations learned by the teacher into a compact student model. The overall objective integrates the supervision from ground-truth annotations with an additional soft-label constraint that captures both inter-class and intra-class structural relationships reflected in the teacher's output distribution.

In contrast to the teacher network, whose training involves an auxiliary branch to guide intermediate layers, the student model relies solely on pixel-level supervision based on cross-entropy and Dice criteria. The corresponding hard-label objective is expressed as

$$L_{\text{hard}} = L_{\text{ce}} + L_{\text{dice}}. \tag{5}$$

To further narrow the discrepancy between teacher and student predictions, the soft-label branch introduces two complementary terms. The first term characterizes semantic dependencies across categories by evaluating the similarity of class-level probability distributions through the mean Pearson correlation, thereby promoting coherent global behavior.

$$L_{\text{inter}} = 1 - \text{mean}\Big(\text{Pearson}(\hat{y}^T_{\text{student}}, \hat{y}^T_{\text{teacher}})\Big) \tag{6}$$

The intra-class relation loss models consistency within each class by transposing the output tensors:

$$L_{\text{intra}} = 1 - \text{mean}\Big(\text{Pearson}(\hat{Y}^T_{\text{student}}, \hat{Y}^T_{\text{teacher}})\Big) \tag{7}$$

where:

$$\hat{Y}^T_{\text{student}} = (\hat{y}^T_{\text{student}})^\top, \quad \hat{Y}^T_{\text{teacher}} = (\hat{y}^T_{\text{teacher}})^\top$$

Here, T is the temperature scaling factor for softening the output logits. In our implementation, T is set to 0.5 .The final total loss for training the student model is given by:

$$L_{\text{total}} = L_{\text{hard}} + L_{\text{inter}} + L_{\text{intra}} \tag{8}$$

By excluding the auxiliary loss, the student model remains simpler and more efficient, while the DIST-based soft constraints ensure that it inherits the structural knowledge learned by the teacher.

Experimental results

Datasets

To rigorously evaluate the performance of SAM2-ARAFNet, we conduct experiments on two high-resolution remote sensing benchmarks, namely the ISPRS Vaihingen and Potsdam datasets. The composition of each dataset is summarized in Table 1, where the image identifiers used for training, validation, and testing are listed together with their associated land-cover categories. The rich variety of urban environments contained in these datasets enables a comprehensive examination of the model's ability to generalize across different scene types.

ISPRS Vaihingen dataset

The ISPRS Vaihingen dataset[56] comprises 33 aerial image tiles released within the ISPRS benchmark. These tiles provide very high spatial detail, with a ground sampling distance of 0.5 m and annotations covering six land-cover classes, including impervious areas, buildings, low vegetation, trees, vehicles, and clutter. Although the tile dimensions are not uniform, they typically measure around 2494 by 2064 pixels, and the underlying data

Datasets	Split			Category
	Train	Validation	Test	
ISPRS Vaihingen	1,3,5,7,11,13,15,17, 21,23,26,28,32,34,37(15)	30(1)	2,4,6,8,10,12,14,16,20, 22,24,27,29,31,33,35,38(17)	Imp. Surf.,building,low vegetation, tree,car,clutters (6)
ISPRS Potsdam	2_11,2_12,3_10,3_11,3_12,4_10,4_11, 4_12,5_10,5_11,5_12,6_7,6_8,6_9,6_10, 6_11,6_12,7_7,7_8,7_9,7_11,7_12(22)	2_10(1)	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_15,7_13(14)	Imp.Surf.,building,low vegetation, tree,car,clutters (6)

Table 1. Overview of category definitions and dataset splits.

achieve an effective ground resolution of roughly 9 cm. Each scene contains a three-channel true orthophoto (near-infrared, red, and green) together with a corresponding digital surface model, both aligned on a unified spatial grid to maintain geometric coherence.

ISPRS Potsdam datasett
The ISPRS Potsdam dataset[56] contains 38 aerial image tiles acquired over the city of Potsdam, Germany, and is widely used for evaluating high-resolution semantic segmentation methods. It provides densely annotated labels for six land-cover categories, including impervious surfaces, buildings, low vegetation, trees, vehicles, and clutter. The dataset offers multi-spectral true orthophotos in several spectral configurations (IR–R–G, R–G–B, and R–G–B–IR), along with a digital surface model, all produced with a ground sampling distance of 5 cm. The orthophotos were derived from a larger ortho-mosaic, while the digital surface model was generated as part of the same photogrammetric workflow to ensure spatial consistency. In this work, only the TOP imagery and the corresponding segmentation masks—without boundary annotations—are utilized. Each tile in the dataset has a resolution of 6000×6000 pixels.

Experimental setup

All experiments were conducted on a workstation configured with a single NVIDIA TITAN RTX GPU. To ensure the reproducibility of experimental results, a fixed random seed (42) was uniformly set for all critical processes, including model training, testing, and data augmentation. This ensures that all random operations are reproducible. Training adopted the AdamW optimizer together with a cosine-annealing strategy for learning-rate adjustment. To enhance optimization stability, the backbone and the auxiliary modules were assigned distinct learning rates of 6×10^{-4} and 6×10^{-5}, respectively, and a weight decay of 0.01 was applied. Following common practice in high-resolution remote-sensing segmentation[54,57], the training data were augmented through random geometric transformations, horizontal and vertical flips, perturbations to brightness and contrast, resizing-based cropping, and light sharpening.

For the ISPRS Vaihingen and Potsdam benchmarks, we adopted the RGB images from the Vaihingen and Potsdam datasets, and image tiles of size 512×512 were randomly extracted during training. Networks were trained for at most 200 epochs with a batch size of 8. Performance assessment considered both computational cost and predictive capability. Model efficiency was examined through the number of trainable parameters, whereas predictive quality was evaluated using three commonly adopted metrics: mean intersection-over-union (mIoU), mean F1 score (mF1), and overall accuracy (OA). In line with previous work, the mIoU, mF1 and OA calculations covered the five principal semantic classes, excluding the clutter category.

The F1 metric is formulated as:

$$F_1 = 2 \cdot \frac{P\,R}{P + R}, \tag{9}$$

with the class-averaged precision P and recall R computed by:

$$P = \frac{1}{k} \sum_{i=1}^{k} \frac{p_{ii}}{\sum_{j=1}^{k} p_{ji}}, \qquad R = \frac{1}{k} \sum_{i=1}^{k} \frac{p_{ii}}{\sum_{j=1}^{k} p_{ij}}. \tag{10}$$

This formulation captures the harmonic interplay between precision and recall, enabling a comprehensive assessment of the model's classification capability.

Overall accuracy (OA) was computed using:

$$\mathrm{OA} = \frac{\sum_{i=1}^{k} p_{ii}}{\sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij}}, \tag{11}$$

which quantifies the ratio of pixels assigned to the correct class relative to the entire set of annotated pixels.

The mean IoU is given by:

$$\mathrm{mIoU} = \frac{1}{k} \sum_{i=1}^{k} \frac{p_{ii}}{\sum_{j=1}^{k} p_{ij} + \sum_{j=1}^{k} p_{ji} - p_{ii}}, \tag{12}$$

and provides a more discriminative evaluation than OA, particularly for classes with relatively small spatial coverage.

Quantitative and qualitative evaluation of semantic segmentation

Performance on the Vaihingen benchmark
The performance of SAM2-ARAFNet on the Vaihingen benchmark is summarized in the quantitative results reported in Table 2. The model achieves an mF1 of 91.99% and an mIoU of 85.43%, reflecting strong segmentation capability across the major land-cover classes. To provide a comprehensive assessment, comparisons are made with a wide range of CNN-based architectures, including PSPNet[33], BiSeNet[58], MANet[61], MAResUNet[62], A2FPN[60], SLCNet[63], and GCDNet[64]. Furthermore, to evaluate the benefit of integrating convolutional and transformer representations, we also include several hybrid designs such as BANet[59], UNetFormer[54],

Method	Backbone	Class F1/IoU %					mF1	mIoU	OA
		Imp.surf.	Building	Low.veg.	Tree	Car	%	%	%
PSPNet[33]	ResNet18	95.19/90.81	94.05/88.77	83.37/71.48	89.60/81.15	82.55/70.28	88.95	80.50	91.58
BiSeNet[58]	ResNet18	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[59]	ResNet18	95.56/91.50	95.24/90.90	83.21/71.25	89.57/81.11	88.60/79.54	90.44	82.86	92.02
A2FPN[60]	ResNet18	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[61]	ResNet50	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[62]	ResNet18	95.72/91.78	95.31/91.04	83.67/71.93	89.78/81.46	87.79/78.23	90.45	82.89	92.19
UNetFormer[54]	ResNet18	95.68/91.72	95.25/90.92	83.85/72.20	89.77/81.43	87.93/78.47	90.50	82.95	92.21
SLCNet[63]	ResNet50	95.80/91.94	95.47/91.33	84.13/72.61	89.94/81.71	88.93/80.07	90.86	83.53	92.38
GCDNet[64]	ResNet101	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[64]	ResNet50	95.74/91.84	95.93/92.17	84.03/72.45	90.07/81.93	89.40/80.83	91.03	83.84	92.49
SFANet[65]	efficientnet_b3	95.66/91.69	95.70/91.76	83.32/71.41	89.79/81.48	87.64/78.00	90.42	82.87	92.19
MIFNet[66]	ResNeXt	96.03/92.36	95.87/92.07	84.26/72.80	90.10/81.99	89.75/81.40	91.20	84.12	92.66
STUNet[67]	Swin transformer	95.92/92.15	95.63/91.54	84.52/73.12	90.35/82.33	89.21/80.15	91.12	83.85	92.67
DMANet[68]	Swin transformer	96.05/92.41	95.87/91.98	84.89/73.68	90.62/82.77	89.87/80.92	91.46	84.35	92.93
SAM2-ARAFNet (ours)	sam2	96.34/92.94	96.40/93.06	85.70/74.98	91.07/83.61	90.44/82.54	91.99	85.43	93.33

Table 2. Quantitative evaluation on the Vaihingen benchmark.

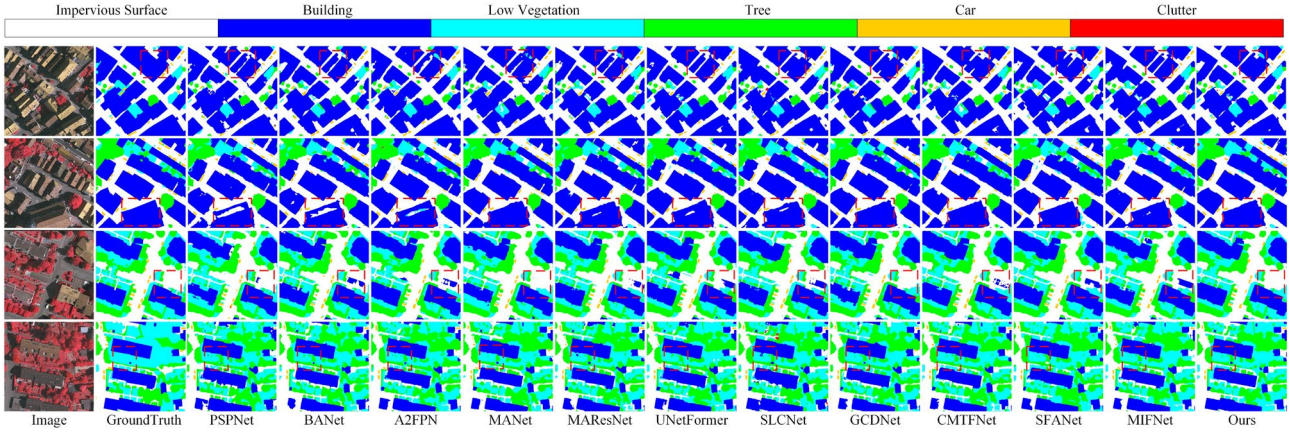

Fig. 4. Visualization results on the Vaihingen dataset, with key distinctions marked by red boxes.

CMTFNet[69], SFANet[65],STUNet[67],and DMANet[68]. According to Table 2, SAM2-ARAFNet outperforms the majority of existing CNN-only and hybrid transformer–CNN models. Since the proposed network adopts a combined CNN–transformer structure, the comparisons with BANet, UNetFormer, CMTFNet, and SFANet are particularly indicative of its advantages.

Qualitative comparisons with state-of-the-art approaches further highlight the robustness of the proposed model under challenging conditions. SAM2-ARAFNet demonstrates a strong ability to mitigate large intraclass variations caused by occlusion and structural complexity. For example, in the red-marked regions in the first two rows of Fig. 4, the geometric arrangement of nearby buildings leads several competing models to misclassify vehicles whose contours resemble background structures. In contrast, SAM2-ARAFNet correctly distinguishes these vehicles. Similarly, in the lower rows, subtle differences in color and texture make it difficult for other methods to accurately delineate building boundaries, whereas SAM2-ARAFNet succeeds in producing clearer and more consistent segmentation of adjacent structures.

Evaluation on the Potsdam dataset

SAM2-ARAFNet is evaluated against a series of representative segmentation networks, as reported in Table 3. The proposed model attains an mIoU of 87.44% and an mF1 of 93.18%, marking a substantial improvement compared with competing approaches. Among conventional CNN-based architectures, GCDNet shows relatively strong performance, yet SAM2-ARAFNet surpasses it by 1.46% in mIoU and 0.85% in mF1. When contrasted with MIFNet, the gains reach 0.54% and 0.33% for mIoU and F1, respectively, indicating that the proposed framework achieves a favorable trade-off between computational cost and segmentation accuracy. These quantitative advantages are consistent with the qualitative comparisons shown in Fig. 5.

On the Potsdam benchmark, the visual examples in Fig. 5 highlight several challenging scenarios. In the first two rows, the spectral and structural similarity between low vegetation and trees leads to frequent confusion

Method	Backbone	Class F1/IoU %					mF1	mIoU	OA
		Imp.surf.	Building	Low.veg.	Tree	Car	%	%	%
PSPNet[33]	ResNet18	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[58]	ResNet18	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[59]	ResNet18	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[60]	ResNet18	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[61]	ResNet50	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[61]	ResNet18	93.44/87.69	96.19/92.65	86.88/76.80	88.28/79.02	95.73/91.81	92.10	85.59	90.82
UNetFormer[54]	ResNet18	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[63]	ResNet50	93.04/86.98	95.84/92.01	86.80/76.68	88.81/79.87	95.61/91.58	92.02	85.42	90.66
GCDNet[64]	ResNet101	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[69]	ResNet50	93.80/88.32	96.54/93.32	87.81/78.28	88.82/79.89	96.14/92.57	92.63	86.48	91.38
SFANet[65]	efficientnet_b3	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[66]	ResNeXt	94.18/89.00	97.05/94.28	87.31/77.49	89.17/80.46	96.53/93.31	92.85	86.90	91.55
STUNet[67]	Swin Transformer	94.05/88.73	96.23/92.81	87.42/77.65	89.05/80.12	96.18/92.61	92.59	86.81	91.42
DMANet[68]	Swin Transformer	94.28/89.15	96.47/93.25	87.76/78.12	89.38/80.64	96.35/92.89	92.85	86.81	91.68
SAM2-ARAFNet (ours)	Sam2	94.69/89.92	97.10/94.37	88.42/79.24	89.53/81.05	96.18/92.64	93.18	87.44	92.13

Table 3. Quantitative evaluation on the potsdam benchmark.

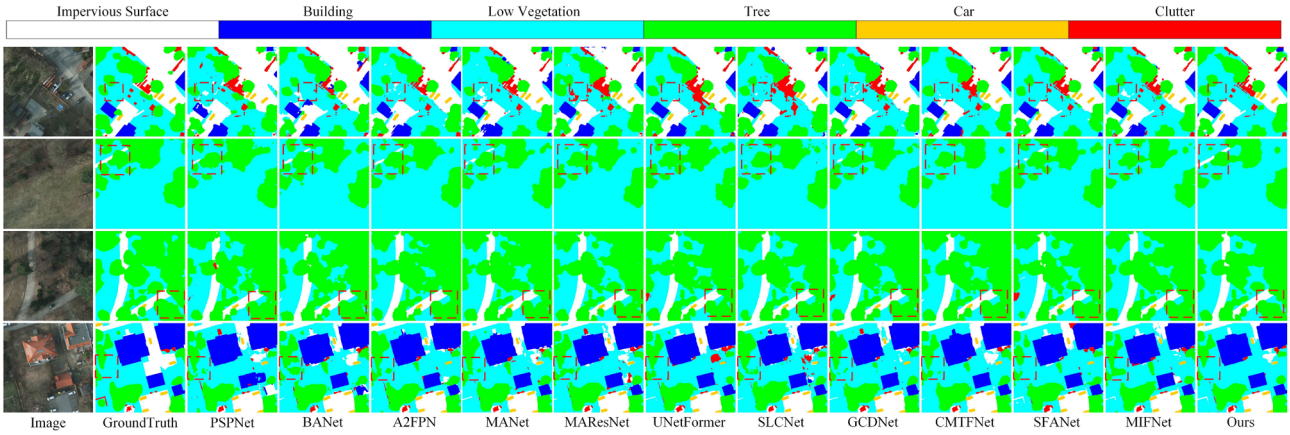

Fig. 5. Potsdam dataset visualizations with highlighted differences.

among existing methods. In such cases, SAM2-ARAFNet demonstrates clearer category boundaries and more stable differentiation between the two vegetation types. Moreover, clutters located adjacent to building facades often form spatial configurations that introduce additional ambiguity. As illustrated in the red-marked regions of the third and fourth rows, many methods struggle with boundary inconsistencies, misclassification, and partial omission. SAM2-ARAFNet effectively mitigates these issues, providing more coherent predictions and noticeably improving class separation in densely arranged urban scenes.

Performance comparison of the teacher and student models

In this experiment, the previously introduced SAM2-ARAFNet is treated as the teacher model, whereas a variant equipped with a lightweight backbone serves as the student model. As described in "Loss function", the teacher network is trained using a composite objective that includes pixel-level cross-entropy, a Dice loss for improving region consistency, and an auxiliary cross-entropy term that supervises intermediate decoder outputs.

To address the performance gap between the heavyweight teacher model (SAM2-ARAFNet) and the lightweight student model, we specifically tailor a relational knowledge distillation loss that aligns the student's semantic and structural feature representations with those of the teacher at the logits level. As elaborated in "Loss function", this custom distillation loss integrates two core relational constraints: inter-class relational constraints that preserve the discriminative distance between different semantic categories in remote sensing scenes and intra-class relational constraints that enhance the consistency of the same category . Unlike label-level distillation that only transfers hard classification targets, this logits-level distillation strategy enables the student to inherit not only the teacher's final classification decisions but also the fine-grained semantic hierarchies and structural dependencies encoded in the teacher's logits—information that is critical for handling the complex spatial layouts and ambiguous boundaries common in remote sensing image segmentation tasks.

By combining this tailored distillation loss with the standard hard-label loss, the student model (KD-ARAFNet) effectively mitigates the performance degradation caused by backbone lightweighting. The experimental results of KD-ARAFNet further validate that distillation is sufficient for our remote sensing segmentation task: the relational constraints embedded in the distillation loss are capable of transferring the teacher's domain-specific knowledge, making it unnecessary to introduce additional pixel-level or feature-level distillation modules.

From a practical standpoint, model inference performance is strongly influenced by computational complexity under identical hardware conditions. Complexity is commonly measured using indicators such as floating-point operations (FLOPs) and parameter count. To provide a comprehensive comparison, we report four metrics: the number of parameters, model size, FLOPs, and inference throughput (images processed per second). The parameter count and model size follow standard calculation protocols, while inference speed is measured as img/s. FLOPs and inference throughput are computed using the same experimental configuration as in earlier evaluations, with an input batch size of 8, three-channel input, and a spatial resolution of 512×512. A comparison between the prediction performance of the student and teacher models is presented in Table 4.

It is important to highlight that several backbone candidates were examined as alternative architectures for the student model, as summarized in Table 4. Among them, EfficientNet_b0 was ultimately chosen to form the distilled version of the network, KD-ARAFNet. With this replacement, the parameter count is reduced to 3.0% of that of the teacher model before distillation, and the model size decreases from 852.16 MB to 26.13 MB. By removing redundant parameters and compressing neuron representations, the distillation process significantly reduces computational overhead, improves inference speed, and lowers memory usage. These characteristics make the student model particularly suitable for deployment in resource-limited scenarios, including embedded hardware and edge-computing platforms. In addition, the lighter architecture enables faster training cycles and model updates, which is beneficial for applications requiring continual adaptation. As demonstrated in "Quantitative and qualitative evaluation of semantic segmentation", the distilled student model retains performance that is highly comparable to that of the teacher network, indicating that the distillation strategy successfully preserves essential discriminative information while improving efficiency.

Figure 6 provides qualitative comparisons before and after knowledge distillation. The predictions produced by the undistilled student model are displayed in the middle column, while those generated after applying distillation are shown in the fourth column. The rightmost column contains the segmentation outputs of the teacher network. From the visual results, it is evident that distillation effectively conveys the teacher's structural and semantic knowledge to the student, resulting in outputs that more closely match the teacher's predictions. Furthermore, the distilled student model exhibits consistently higher accuracy than its non-distilled counterpart. For clarity, we refer to the EfficientNet_b0-based student model without knowledge distillation as *KD-ARAFNet (Baseline)*, while the version trained with the full distillation pipeline is denoted as *KD-ARAFNet (Distilled)*.

As shown in Table 5, the student model without incorporating distillation (*KD-ARAFNet Baseline*) records mF1, mIoU, and OA scores of 90.78%, 83.43%, and 92.41%, respectively. These metrics amount to roughly 98.7%, 97.7%, and 99.0% of the corresponding results obtained by the teacher network. The performance decline mainly stems from replacing the original backbone with a lightweight alternative, which inevitably reduces the model's representational strength. Once knowledge distillation is introduced, the student model (*KD-ARAFNet Distilled*) demonstrates a notable recovery in performance, improving to 91.48% mF1, 84.59% mIoU, and 93.02% OA. Following distillation, the student model preserves more than 99.4% of the teacher's accuracy in terms of OA, significantly narrowing the gap between the two. These findings confirm that, although the student architecture substantially reduces parameters and computational demands, incorporating distillation enables it to maintain recognition performance that closely approximates that of the teacher network.

Ablation study

To clarify the contribution of individual modules within the overall architecture, a set of controlled ablation experiments was performed in which selected components were removed or replaced. The examination centers on the effects of the multi-scale atrous convolution block, the attention-augmented ResASPP module, the shallow feature fusion pathway, and the BFM unit. All ablation configurations were trained under identical backbone

Backbone	F1 (%)	mIoU (%)	OA (%)	Params (M)	Params reduction (%)	Flops (G)	FPS (img/s)	Size (MB)
SAM2	91.99	85.43	93.33	222.98	–	1712.47	20.01	852.16
Efficientnet_b0 (Selected)	91.48	84.59	93.02	6.68	97.00	56.36	305.85	26.13
Efficientnet_b1	91.64	84.83	93.16	9.16	95.89	64.03	235.69	35.81
Mobilenetv3_large	91.13	83.99	92.71	9.94	95.54	57.80	357.14	38.75
Mobilenetv3_small	90.30	82.64	92.29	5.47	97.55	43.23	586.04	21.40
Resnet18	91.64	84.82	92.97	15.76	92.93	129.95	352.87	60.52
Resnet50	91.39	84.38	92.80	39.18	82.43	295.68	128.05	150.83
Segformer_b0	91.45	84.52	92.93	6.15	97.24	60.90	206.30	23.73
Segformer_b1	91.61	84.79	93.06	17.81	92.01	123.37	130.32	68.33
Ghostnetv3	91.09	83.91	92.66	4.02	98.20	50.71	188.25	16.41

Table 4. Comparison of model complexity, inference speed, and prediction performance between teacher and student models.

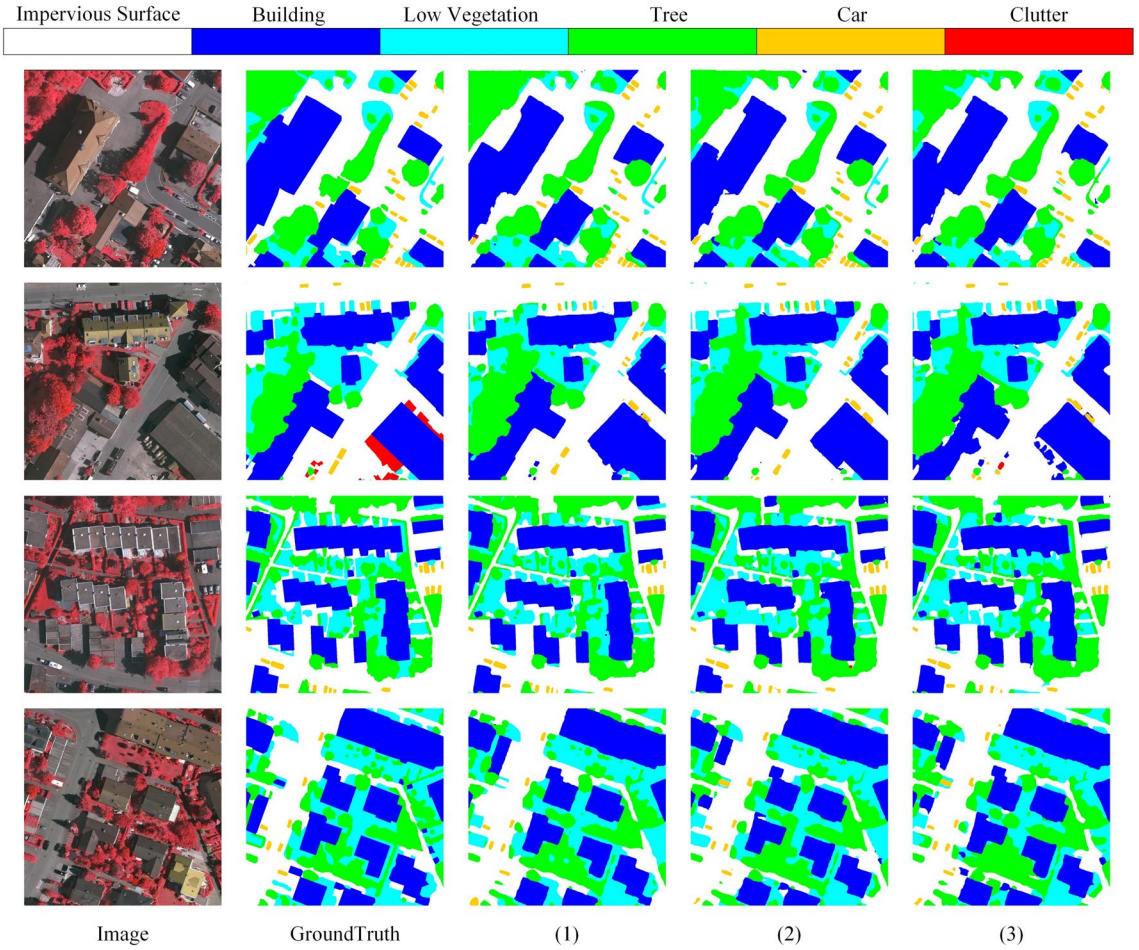

Fig. 6. Visual comparison results of the module ablation study. (1) SAM2-ARAFNet (2) KD-ARAFNet (Distilled) (3) KD-ARAFNet (Baseline).

Method	Class F1/IoU %					mF1	mIoU	OA
	Imp.surf.	Building	Low.veg.	Tree	Car	%	%	%
SAM2-ARAFNet	96.34/92.94	96.41/93.06	85.70/74.98	91.08/83.61	90.44/82.55	91.99	85.43	93.34
KD-ARAFNet (Distilled)	96.19/92.66	96.26/92.79	85.00/73.91	90.58/82.77	89.38/80.80	91.48	84.59	93.02
KD-ARAFNet (Baseline)	95.80/91.94	95.70/91.75	83.82/72.15	89.96/81.76	88.61/79.54	90.78	83.43	92.41

Table 5. Comparison of classification accuracy before and after knowledge distillation.

and decoder settings to maintain comparability, and the corresponding quantitative outcomes are reported in Table 6.And the ablation performance of each module is shown in the Fig. 7.

From Table 6, it is evident that multi-scale atrous convolutions and attention mechanisms within the ResASPP structure play a critical role in enhancing high-level semantic representations. Removing the ResASPP module leads to a significant performance drop, while eliminating all attention modules within ResASPP results in a moderate yet notable decline, demonstrating that attention further strengthens multi-scale feature fusion. Comparing branch and trunk attention, we observe that trunk attention contributes more to global semantic refinement, whereas branch attention primarily enhances local details.

The low-level feature fusion path, implemented via skip connections, is also crucial for accurate segmentation. When this path is removed but BFM is retained, performance decreases, highlighting the importance of low-level spatial details for boundary reconstruction. Removing both the low-level path and the BFM module exacerbates the performance drop, confirming that both components are necessary for preserving fine-grained spatial information.

Finally, the BFM module itself provides adaptive and effective high–low feature integration. Replacing it with simple addition or concatenation reduces accuracy, with concatenation slightly outperforming addition. Moreover, simultaneously removing attention, BFM, and ASPP and using basic fusion operations causes the

Method	Class F1/IoU %					mF1	mIoU	OA
	Imp.surf.	Building	Low.veg.	Tree	Car	%	%	%
SAM2-ARAFNet	96.34/92.94	96.41/93.06	85.70/74.98	91.08/83.61	90.44/82.55	91.99	85.43	93.34
w/o ResASPP module	96.13/92.55	96.24/92.74	84.80/73.61	90.68/82.95	87.84/78.31	91.14	84.03	92.89
w/o all attention in ResASPP	96.30/92.86	96.40/93.03	85.11/74.08	90.79/83.14	88.04/78.64	91.34	84.37	93.14
w/o branch attention in ResASPP	96.25/92.78	96.54/93.30	85.19/74.21	90.91/83.34	89.89/81.64	91.76	85.05	93.22
w/o trunk attention in ResASPP	96.43/93.12	96.50/93.24	85.70/74.97	90.60/82.81	89.65/81.24	91.79	85.10	93.29
w/o LowPath (BFM retained)	95.98/92.28	95.92/92.17	84.93/73.81	90.27/82.27	84.93/73.81	90.41	82.87	92.72
w/o LowPath (BFM removed)	95.72/91.79	95.96/92.23	84.53/73.21	90.27/82.26	82.28/69.89	89.75	81.88	92.54
w/o BFM (add)	95.73/91.80	95.90/92.13	84.36/72.95	90.10/81.98	83.21/71.25	89.86	82.02	92.48
w/o BFM (concat)	96.11/92.51	96.16/92.60	84.51/73.18	90.72/83.02	86.02/75.47	90.70	83.36	92.86
w/o attention, BFM, and ResASPP (add)	95.63/91.62	95.86/92.06	84.26/72.80	90.04/81.88	81.56/68.86	89.47	81.44	92.41
w/o attention, BFM, and ResASPP (Concat)	95.87/92.07	95.03/90.53	83.10/71.08	90.31/82.33	85.11/74.08	89.88	82.02	92.02

Table 6. Ablation study performance on the Vaihingen dataset.

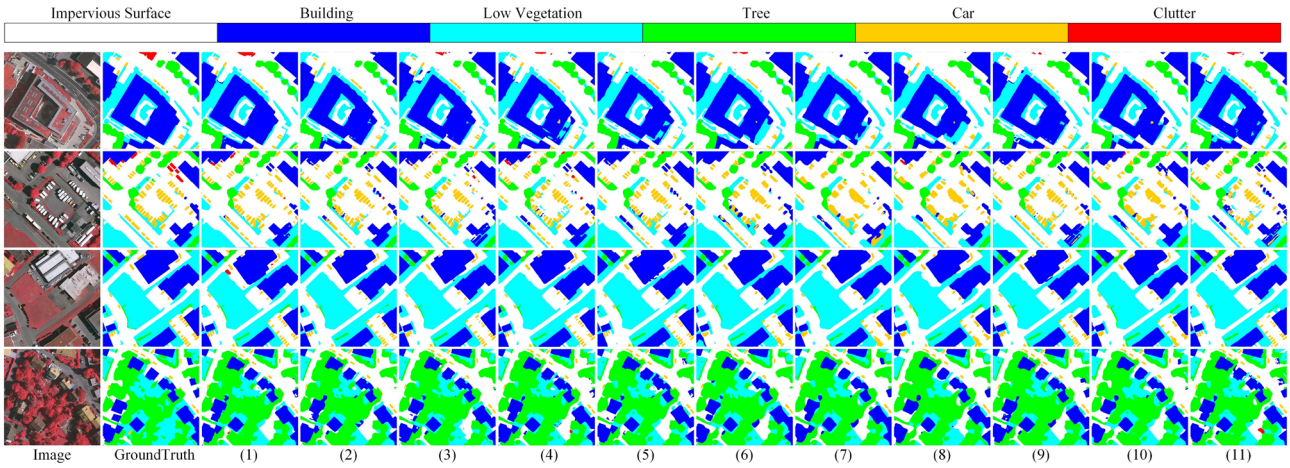

Fig. 7. Visual comparison results of the module ablation study. (1) SAM2-ARAFNet (2) w/o ResASPP module (3) w/o all attention in ResASPP (4) w/o branch attention in ResASPP (5) w/o trunk attention in ResASPP (6) w/o LowPath (BFM retained) (7) w/o LowPath (BFM removed) (8) w/o BFM (add) (9) w/o BFM (concat) (10) w/o attention, BFM, and ResASPP (add) (11) w/o attention, BFM, and ResASPP (Concat).

largest performance degradation, demonstrating that the synergy among multi-scale atrous convolutions, attention, and BFM is essential for achieving optimal segmentation performance.Overall, the ablation study validates that each component contributes positively, and their combined effect yields the best results.

Conclusion

This work introduces SAM2-ARAFNet, a remote sensing segmentation framework that integrates the capabilities of Segment Anything 2 through attention-enhanced residual atrous pyramid features and adapter-based parameter refinement. To support deployment on resource-constrained platforms, a tailored distillation scheme is further incorporated, enabling the transfer of structural and semantic cues from a high-capacity teacher model to a compact student network built upon EfficientNet.

Comprehensive evaluations on the ISPRS Vaihingen and Potsdam benchmarks verify the effectiveness of the proposed approach, showing notable gains in mIoU, mF1, and OA over representative convolutional and Transformer-based methods. The framework demonstrates strong robustness in complex urban scenes characterized by substantial intra-class variation, indistinct object boundaries, and categories with high spectral or spatial resemblance, yielding more reliable and fine-grained segmentation outcomes.

Moreover, the introduced distillation strategy leads to a substantial reduction in model parameters, decreasing the size by 97% (222.98 M to 6.68 M) while still preserving more than 99% of the teacher network's predictive capability. This indicates that SAM2-ARAFNet and its distilled counterpart (KD-ARAFNet) hold strong potential for deployment on devices with limited computational budgets, including UAV platforms and various edge-side systems.

The ablation experiments further demonstrate that all major components—ResASPP, the shallow-feature aggregation branch, and the bilateral fusion module—make meaningful contributions to the overall system behaviour, with their joint use yielding the most stable and accurate segmentation outcomes. When considered

together, these findings show that SAM2-ARAFNet offers an effective and reliable solution for high-resolution remote sensing semantic segmentation, achieving a favourable compromise between accuracy, robustness, and computational efficiency.

Looking forward, several research directions could further strengthen this line of work. The incorporation of prompt-guided cues or additional multimodal information may enhance contextual reasoning and improve adaptability in complex scenes. Furthermore, combining knowledge distillation with other compression paradigms—such as pruning, quantization, or sparsity-driven optimisation—may further reduce resource demands and facilitate deployment on UAV and edge devices. Finally, scaling experiments to broader and more heterogeneous datasets and extending the framework to tasks such as change detection, object extraction, or temporal analysis would help verify the scalability and practical utility of SAM2-ARAFNet.

Data availability

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Received: 6 December 2025; Accepted: 28 January 2026
Published online: 23 February 2026

References

1. Xie, Z., Li, X., Ma, H., Wu, S. & Cui, D. PUNet: a lightweight parallel U-Net architecture integrating Mamba-CNN for high-precision image segmentation. *Sci. Rep.* **15**, 38954. https://doi.org/10.1038/s41598-025-22862-x (2025).
2. Wu, B., Chen, B., Jiang, X. & Liu, Z. Pruned U-Net with multi-scale feature fusion and attention for real-time UAV remote sensing of levee defects. *Sci. Rep.* **15**, 42354. https://doi.org/10.1038/s41598-025-26431-0 (2025).
3. Ji, Y., Shi, W., Lei, J. & Ding, J. DBRSNet: a dual-branch remote sensing image segmentation model based on feature interaction and multi-scale feature fusion. *Sci. Rep.* **15**, 27786. https://doi.org/10.1038/s41598-025-13236-4 (2025).
4. Yan, L. et al. A multilevel multimodal hybrid Mamba-Large strip convolution network for remote sensing semantic segmentation. *Remote Sens.* **17**, 2696. https://doi.org/10.3390/rs17152696 (2025).
5. Ge, X., Zhou, L. & Meng, D. DDNet: disaster damage detection for buildings based on dual-temporal joint attention network. *Sci. Rep.* **15**, 42513. https://doi.org/10.1038/s41598-025-26480-5 (2025).
6. Li, Z. et al. GPRNet: A geometric prior-refined semantic segmentation network for land use and land cover mapping. *Remote Sens.* **17**, 3856. https://doi.org/10.3390/rs17233856 (2025).
7. Yang, Z., Li, H., Wei, F., Ma, J. & Zhang, T. WSC-Net: A wavelet-enhanced Swin Transformer with cross-domain attention for hyperspectral image classification. *Remote Sens.* **17**, 3216. https://doi.org/10.3390/rs17183216 (2025).
8. Wang, W. et al. A review of road extraction from remote sensing images. *J. Traffic Transport. Eng. (Engl. Ed.)* **3**, 271–282 (2016).
9. Hossain, M. D. & Chen, D. Segmentation for object-based image analysis (OBIA): A review of algorithms and challenges from remote sensing perspective. *ISPRS J. Photogramm. Remote. Sens.* **150**, 115–134 (2019).
10. Kotaridis, I. & Lazaridou, M. Remote sensing image segmentation advances: A meta-analysis. *ISPRS J. Photogramm. Remote. Sens.* **173**, 309–322 (2021).
11. Dey, V., Zhang, Y. & Zhong, M. A review on image segmentation techniques with remote sensing perspective. *XXII ISPRS Congr.* **38** (2010).
12. Moser, G., Serpico, S. B. & Benediktsson, J. A. Land-cover mapping by markov modeling of spatial-contextual information in very-high-resolution remote sensing images. *Proc. IEEE* **101**, 631–651 (2012).
13. Chandra, M. A. & Bedi, S. Survey on SVM and their application in image classification. *Int. J. Inf. Technol.* **13**, 1–11 (2021).
14. Bagwari, N., Kumar, S. & Verma, V. S. A comprehensive review on segmentation techniques for satellite images. *Arch. Comput. Methods Eng.* **30**, 4325–4358 (2023).
15. Pal, M. & Mather, P. M. Support vector machines for classification in remote sensing. *Int. J. Remote Sens.* **26**, 1007–1011 (2005).
16. Juel, A., Groom, G. B., Svenning, J.-C. & Ejrnaes, R. Spatial application of random forest models for fine-scale coastal vegetation classification using object based analysis of aerial orthophoto and DEM data. *Int. J. Appl. Earth Obs. Geoinf.* **42**, 106–114 (2015).
17. Huang, B., Zhao, B. & Song, Y. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. *Remote Sens. Environ.* **214**, 73–86 (2018).
18. Pires de Lima, R. & Marfurt, K. Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. *Remote Sens.* **12**, 86 (2019).
19. Toldo, M., Maracani, A., Michieli, U. & Zanuttigh, P. Unsupervised domain adaptation in semantic segmentation: a review. *Technologies* **8**, 35 (2020).
20. Zeng, Q. & Geng, J. Task-specific contrastive learning for few-shot remote sensing image scene classification. *ISPRS J. Photogramm. Remote. Sens.* **191**, 143–154 (2022).
21. Kirillov, A. et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026 (2023).
22. Ravi, N. et al. SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024).
23. Wang, X. et al. SegGPT: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1130–1140 (2023).
24. Li, X. et al. OMG-Seg: Is one model good enough for all segmentation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27948–27959 (2024).
25. Chen, T. et al. SAM-Adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3367–3375 (2023).
26. Zhang, K. & Liu, D. Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023).
27. Huang, D. et al. AlignSAM: Aligning segment anything model to open context via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3205–3215 (2024).
28. Zhang, Y. et al. EVF-SAM: Early vision-language fusion for text-prompted segment anything model. arXiv preprint arXiv:2406.20076 (2024).
29. Li, W., Xiong, X., Xia, P., Ju, L. & Ge, Z. TP-DRSeg: improving diabetic retinopathy lesion segmentation with explicit text-prompts assisted SAM. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 743–753 (Springer, 2024).
30. Zhang, R. et al. Personalize segment anything model with one shot. arXiv preprint arXiv:2305.03048 (2023).
31. Liu, Y. et al. Matcher: Segment anything with one shot using all-purpose feature matching. arXiv preprint arXiv:2305.13310 (2023).
32. Xiong, X. et al. SAM2-UNet: Segment anything 2 makes strong encoder for natural and medical image segmentation. arXiv preprint arXiv:2408.08870 (2024).

33. Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2881–2890 (2017).
34. Fang, L., Zhou, P., Liu, X., Ghamisi, P. & Chen, S. Context enhancing representation for semantic segmentation in remote sensing images. *IEEE Trans. Neural Netw. Learn. Syst.* **35**, 4138–4152 (2022).
35. Pastorino, M., Moser, G., Serpico, S. B. & Zerubia, J. CRFNet: A deep convolutional network to learn the potentials of a CRF for the semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* (2024).
36. Meng, X. et al. Class-guided swin transformer for semantic segmentation of remote sensing imagery. *IEEE Geosci. Remote Sens. Lett.* **19**, 1–5 (2022).
37. Zheng, C. et al. SSDT: Scale-separation semantic decoupled transformer for semantic segmentation of remote sensing images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* (2024).
38. Wang, L. et al. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote. Sens.* **190**, 196–214 (2022).
39. Wu, H., Huang, P., Zhang, M., Tang, W. & Yu, X. Cmtfnet: Cnn and multiscale transformer fusion network for remote-sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–12 (2023).
40. Wu, J. et al. Medical SAM adapter: Adapting segment anything model for medical image segmentation. *Med. Image Anal.* **102**, 103547 (2025).
41. Gao, S., Zhang, P., Yan, T. & Lu, H. Multi-scale and detail-enhanced segment anything model for salient object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9894–9903 (2024).
42. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015).
43. Nguyen, T., Novak, R., Xiao, L. & Lee, J. Dataset distillation with infinitely wide convolutional networks. *Adv. Neural. Inf. Process. Syst.* **34**, 5186–5198 (2021).
44. Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A. A. & Wilson, A. G. Does knowledge distillation really work?. *Adv. Neural. Inf. Process. Syst.* **34**, 6906–6919 (2021).
45. Wang, L. & Yoon, K.-J. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 3048–3068 (2021).
46. Liu, Y., Shu, C., Wang, J. & Shen, C. Structured knowledge distillation for dense prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 7035–7049. https://doi.org/10.1109/TPAMI.2020.3001940 (2023).
47. Wang, Y., Zhou, W., Jiang, T., Bai, X. & Xu, Y. Intra-class feature variation distillation for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 346–362 (Springer, 2020).
48. Yang, C. et al. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12319–12328 (2022).
49. Zhang, P. et al. LGD: Label-guided self-distillation for object detection. *Proc. AAAI Conf. Artif. Intell.* **36**, 3309–3317 (2022).
50. Huang, Y. et al. Label-guided auxiliary training improves 3d object detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 684–700 (Springer, 2022).
51. Hu, M. et al. Knowledge distillation from multi-modal to mono-modal segmentation networks. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 772–781 (Springer, 2020).
52. Ryali, C. et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *Proceedings of the International Conference on Machine Learning (ICML)*, 29441–29454 (PMLR, 2023).
53. Hu, E.J. et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)* (2022).
54. Wang, L. et al. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote. Sens.* **190**, 196–214 (2022).
55. Huang, T., You, S., Wang, F., Qian, C. & Xu, C. Knowledge distillation from a stronger teacher. arXiv preprint arXiv:2205.10536 (2022).
56. International Society for Photogrammetry and Remote Sensing. Potsdam and vaihingen datasets. https://www.isprs.org/education/benchmarks/UrbanSemLab/ (2025) (accessed 20 Oct 2024).
57. Hanyu, T. et al. AerialFormer: Multi-resolution transformer for aerial image segmentation. *Remote Sens.* **16**, 2930 (2024).
58. Yu, C. et al. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 325–341 (2018).
59. Wang, L. et al. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. *Remote Sens.* **13**, 3065 (2021).
60. Hu, M., Li, Y., Fang, L. & Wang, S. A2-FPN: Attention aggregation based feature pyramid network for instance segmentation. arXiv preprint arXiv:2105.03186 (2021).
61. Fan, T., Wang, G., Li, Y. & Wang, H. MA-Net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access* **8**, 179656–179665 (2020).
62. 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.* **19**, 1–5. https://doi.org/10.1109/LGRS.2021.3063381 (2022).
63. Yu, D. & Ji, S. Long-range correlation supervision for land-cover classification from remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–14. https://doi.org/10.1109/TGRS.2023.3324706 (2023).
64. Cui, J., Liu, J., Wang, J. & Ni, Y. Global context dependencies aware network for efficient semantic segmentation of fine-resolution remoted sensing images. *IEEE Geosci. Remote Sens. Lett.* **20**, 1–5. https://doi.org/10.1109/LGRS.2023.3318348 (2023).
65. Hwang, G., Jeong, J. & Lee, S. J. SFA-Net: Semantic feature adjustment network for remote sensing image segmentation. *Remote Sens.* **16**, 3278 (2024).
66. Fan, J., Li, J., Liu, Y. & Zhang, F. Frequency-aware robust multidimensional information fusion framework for remote sensing image segmentation. *Eng. Appl. Artif. Intell.* **129**, 107638 (2024).
67. He, X. et al. Swin Transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–15. https://doi.org/10.1109/TGRS.2022.3144165 (2022).
68. Deng, C., Liang, H., Qin, X. & Wang, S. Dma-net: Dynamic morphology-aware segmentation network for remote sensing images. *Remote Sens.* **17**, https://doi.org/10.3390/rs17142354 (2025).
69. Wu, H., Huang, P., Zhang, M., Tang, W. & Yu, X. CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–12. https://doi.org/10.1109/TGRS.2023.3314641 (2023).

Author contributions

W.S., J.D., J.L., and Y.J. jointly conceived the study. W.S. implemented the proposed method, conducted experiments, and performed data analysis. J.D. and J.L. designed the network architecture and contributed to algorithm optimization. Y.J. supervised the overall project and provided critical revisions. W.S. drafted the main manuscript text, and J.D. and J.L. contributed to revising and improving the manuscript. All authors reviewed and approved the final manuscript.

Funding
This work was supported by the Zhejiang Provincial Science and Technology Program (Grant No. 2024C01109).

Declarations

Competing interests
The authors declare no competing interests.

Datasets used in this study
This study utilizes two publicly available benchmark datasets widely used in remote sensing semantic segmentation research. 1. **ISPRS Potsdam Dataset** This high-resolution urban semantic segmentation dataset is available from the ISPRS benchmark website: https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx 2. **ISPRS Vaihingen Dataset** This aerial image semantic labeling dataset is also publicly accessible at: https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx These datasets were used solely for research and evaluation purposes in accordance with the ISPRS licensing terms.

Additional information
Correspondence and requests for materials should be addressed to Y.J.