EMSNet: Efficient Multimodal Symmetric Network for Semantic Segmentation of Urban Scene from Remote Sensing Imagery

Yejian Zhou, Yachen Wang, Jie Su, Zhenyu Wen, Puzhao Zhang, Wenan Zhang

***Abstract*—High-resolution remote sensing imagery (RSI) plays a pivotal role in the semantic segmentation (SS) of urban scenes, particularly in urban management tasks such as building planning and traffic flow analysis. However, the dense distribution of objects and the prevalent background noise in RSI make it challenging to achieve stable and accurate results from a single-view. Integrating Digital Surface Models (DSM) can achieve high-precision SS. But this often requires extensive computational resources. It is essential to address the trade-off between accuracy and computational cost and optimize the method for deployment on edge devices. In this paper, we introduce an efficient multimodal symmetric network (EMSNet) designed to perform SS by leveraging both optical and DSM images. Unlike other multimodal methods, EMSNet adopts a dual encoder-decoder structure to build a direct connection between DSM data and the final result, making full use of the advanced DSM. Between branches, we propose a continuous feature interaction to guide the DSM branch by RGB features. Within each branch, multi-level feature fusion captures low spatial and high semantic information, improving the model's scene perception. Meanwhile, knowledge distillation (KD) further improves the performance and generalization of EMSNet. Experiments on the Potsdam and Vaihingen datasets demonstrate the superiority of our method over other baseline models. Ablation experiments validate the effectiveness of each component. Besides, the KD strategy is confirmed by comparing it with the Segment Anything Model (SAM). It enables the proposed multimodal SS network to match SAM's performance with only one-fifth of the parameters, computation, and latency.**

***Index Terms*—Remote Sensing Image Interpretation, Semantic Segmentation, Symmetric MultiModal, Segment Anything.**

I. Introduction

ACCURATE target segmentation in high-resolution images (HRI) is essential for various remote sensing applications [1]–[3]. With massive large-scale images captured by spaceborne imaging sensors, deep learning techniques have emerged as a prevalent approach to address this challenge. In this context, urban scene segmentation has attracted extensive attention due to its key role in land use, urban expansion monitoring, and disaster management [4]–[6]. Compared with natural scenes, urban scenes usually have more complex structures, diverse object categories, and higher occlusion, which increases the difficulty of the segmentation task. Consequently, there is a growing focus on devising deep neural networks that strike a balance between segmentation accuracy and computational efficiency, to adapt to these unique challenges, making it a prominent area of research.

Recently, SS has undergone notable advancements driven by convolutional neural networks (CNNs). CNNs excel in extracting intricate image features and accurately differentiating objects based on size and spatial characteristics. Nonetheless, accurately delineating the precise boundaries of targets using CNNs remains a persisting challenge. After that, Unet [7] is used to solve the segmentation task of medical images. It combines encoder and decoder features into a ladder structure, and has garnered attention for its remarkable performance across diverse scenarios. Based on Unet, various studies have proposed variant architectures to enhance the extraction performance of buildings from RSI [8]–[10].

The development of remote sensing technology has brought higher-resolution optical images, which present complex color and texture details in significant background noise and varying target sizes. Moreover, the challenges of top-view-only perspective and 2D space shadow interference further complicate feature extraction. To tackle the issues, dual-stream networks for segmentation have emerged. Existing research has primarily concentrated on optical and DSM images [11]–[13]. Xiu et al. [14] proposed MDAFNet, which utilizes optical images to generate DSM and integrates optical and DSM for segmentation. Fan et al. [15] proposed an information exchange mechanism between optical features and DSM features, enabling their interaction and representation in a shared feature space. This approach facilitates the extraction of complementary information from different modalities. Cui et al. [16] introduced a multimodal gated segmentation network based on frequency decomposition, where correlations were established through low-frequency components, followed by a gating mechanism to fuse modality-specific features.

However, the extended update cycle of DSM hinders its ability to provide real-time change information. Some studies have shifted focus to optical and SAR images [17]–[19].

Yejian Zhou, Yachen Wang, and Wenan Zhang are with College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310023, P.R.China.

Jie Su and Zhenyu Wen are with Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310023, P.R.China.

Puzhao Zhang is with Key Laboratory of Collaborative Intelligence Systems of Ministry of Education of China, Xidian University, Xi'an 710071, P.R.China, and also with Department of Urban Planning and Environment, Royal Institute of Technology KTH, Stockholm 100 44, Sweden;

This work was supported by the National Nature Science Foundation of China under Grant 62471438, 62003251, and 62472387, the Zhejiang Provincial Natural Science Foundation of China (Grant No. LY23F010012), Zhejiang Provincial Natural Science Foundation of Major Program (Youth Original Project) under Grant LDQ24F020001. (Corresponding author: Jie Su, email: jieamsu@gmail.com;)

SAR systems operate independently of weather conditions, guaranteeing consistent updates and providing abundant texture information. In addition, to achieve higher versatility and accuracy, some researchers focus on large-scale models [20]. Du et al. [21] integrated the diffusion model with Mamba, achieving promising results in the semantic segmentation of optical and synthetic aperture radar (SAR) images. Kan et al. [22] proposed MGFNet, a method for multimodal semantic segmentation that leverages a gating mechanism for effective feature integration. Liu et al. [23] combined the strengths of convolutional neural networks (CNNs) and transformers, proposing an internal self-attention mechanism to facilitate multimodal information fusion for land cover classification. Segment Anything Model (SAM) [24] has elevated SS to unprecedented levels, demonstrating exemplary performance across challenging scenarios.

Although the segmentation algorithms mentioned previously have achieved notable advancements in performance, they frequently demonstrate high complexity, often overlooking considerations related to computational load and parameter scale. As a result, these high-performance models tend to feature numerous parameters and require substantial computational resources, leading to increased memory requirements during deployment and reduced processing speed. Despite the rapid progress in artificial intelligence, which facilitates the integration of deep learning models with edge devices, such as in-vehicle or spaceborne applications [25], [26], designing lightweight models capable of achieving high-precision segmentation within the constraints of limited hardware resources remains a formidable challenge.

There is a pressing need for a novel architectural paradigm to strike a balance between model complexity, computational efficiency, and analytical capability. Therefore, we propose a solution: the efficient multimodal symmetric network (EMSNet). EMSNet efficiently uses RGB images and DSM data, achieving high performance and computational efficiency. EMSNet consists of two parallel networks, containing a symmetrical encoder-decoder. This design allows for obtaining results of different attributes, which are then integrated into a comprehensive segmentation map. Due to the limitations of DSM data, the results of the DSM branch have more errors. Therefore, we introduce the SAGate as the bridge to facilitate continuous feature interaction, using RGB features to guide the DSM branch in achieving accurate segmentation. Within each branch, multi-level feature fusion enhances spatial and semantic features and prevents information loss during upsampling. Furthermore, inspired by the structures of SAM and EMSNet, a knowledge distillation strategy is designed to enable EMSNet to show superior performance in complex SAR scenes. Extensive experiments on two DSM datasets confirm that EMSNet significantly outperforms existing models, and results on the SAR dataset show that the distillation strategy improves the generalization of EMSNet.

In summary, our contributions are summarised as follows:

- The proposed EMSNet comprises two parallel networks, linking RGB and DSM data to the final result and fully utilizing advanced attributes of multimodal data. We explore the optimal fusion strategy between RGB and DSM, and continuous and multi-level feature interactions guide the DSM stream with RGB features to prevent performance degradation caused by lightweight. Experiments show that EMSNet achieves SOTA performance with minimal cost, ensuring real-time application of edge devices.
- To enhance the generalization of EMSNet for diverse scenes, we use a knowledge distillation (KD) strategy to allow the model to utilize the advanced encoding and decoding information provided by the SAM in complex scenes. The strategy effectively improves the spatial understanding of the model and produces accurate segmentation results. This approach combines KD strategy and SAM and provides new insights and solutions for efficient multimodal networks applicable in multiple scenes.

Organization. In Section II, we review the existing work on semantic segmentation. We propose the details of EMSNet and knowledge distillation strategy with SAM in Section III. The discussion of experiments is given in Section IV. In Section V, we conclude our paper.

II. Related work

Recently, semantic segmentation for remote sensing has made rapid development. Depending on the input data, the research can be divided into unimodal and multi-modal methods. On one hand, unimodal networks often use specialized components to enhance performance. Addressing downsampling challenges in large-size images, Huynh et al. [27] proposed Magnet, a progressive semantic segmentation framework. Magnet processes refined segmentation outputs in multiple stages, preserving crucial details during the downsampling. Lee et al. [28] devised a semi-supervised learning framework. They leverage targeted data augmentation and a well-designed objective function to enable the model to excel in recognizing small and complex-shaped buildings. For agricultural applications, Zhang et al. [29] introduced the Fine Pyramid Scene Parsing Network (PSPNet), designed for PoISAR images. Meanwhile, the transformer has gradually gained attention in unimodal methods and has been widely used in building extraction tasks [30], [31].

On the other hand, with the advancement of satellite image resolution and recognition of complex remote sensing scenes, unimodal algorithms fall short, paving the way for multimodal models that harness the strengths of diverse data sources to achieve enhanced performance. For RGB and DSM images, Liu et al. [32] introduced AFNet, employing a multi-layer architecture featuring a scale feature attention module. AFNet effectively enhanced small object features. Ma et al. [33] proposed a multimodal method, MSFNet, which leverages cross-attention for fine-resolution segmentation. Zhou et al. [12] used DSM data to achieve unsupervised segmentation of remote sensing optical images. Iyer et al. [34] integrated information similarity from two modalities based on graph networks and created a fusion graph to achieve segmentation. Wang et al. [35] proposed a multimodal feature self-attention fusion module that applies to various data types, including DSM images. For RGB and SAR images, Wu et al. [36] devised a cross-fusion module to blend optical and

SAR features and proposed multimodal aggregation to achieve specialized high-level feature fusion. Li et al. [18] proposed a progressive fusion framework to extract buildings and optimize the edge details of buildings. Xiao et al. [37] proposed a modal intrinsic noise suppression module that effectively eliminated noise in SAR images. Yao et al. [38] proposed a general multimodal framework, ExViT, to address the land cover classification problem. ExViT integrates separable convolutions with a visual transformer to process multimodal images in parallel and introduces a cross-modal attention mechanism to facilitate the exchange of information across heterogeneous modalities. Hong et al. [39] introduced a new multimodal dataset comprising hyperspectral, multispectral, and SAR images. Additionally, they developed a high-resolution domain adaptation network and a novel loss function for solving the class imbalance problem.

Significant progress has been made in remote sensing segmentation, but it is difficult for these networks to break through hardware bottlenecks, thus the necessity of designing lightweight models. More importantly, lightweight achieved with simplified structures will inevitably cause a loss of accuracy, which needs to be targeted and efficiently improved to enhance the expressive power of the model.

III. Method

In this section, we first introduce the overall framework of the proposed method. Subsequently, we integrate the architecture of EMSNet to analyze its feasibility in DSM scenarios. Finally, we expound on the SAM-based distillation strategy for complex SAR scenarios, ensuring that EMSNet maintains its exemplary performance.

A. Overview

The overall operating framework of the proposed method is illustrated in Figure 1. In Stage 1, the EMSNet serves as a segmentation network designed for low computational complexity and few parameters. For simple DSM scenes, the outputs from the RGB and DSM branches complement each other, enabling EMSNet to achieve optimal performance. However, while the use of simple CNNs effectively reduces parameters and computational demands, it often compromises the quality of feature representation. In complex SAR scenarios, radar characteristics such as scattering and interference pose challenges for EMSNet, making it difficult to extract accurate features, resulting in suboptimal performance and hindering the generalization. Therefore, in stage 2, we propose a SAM-based distillation strategy to improve EMSNet's performance in SAR scenarios. SAM is the first large-scale model with excellent performance in visual segmentation, which has powerful feature extraction and zero-shot capabilities. This strategy enables EMSNet to approximate the performance of SAM in the SAR scenario but with much fewer parameters, computations, and delays.

B. EMSNet Details

The easy accessibility of RGB images enables real-time monitoring of targets, capturing local details within the area. Conversely, DSM has a long acquisition period and provides a macro view of the integrated status and changes, reflecting the overall characteristics and long-term trends of the area. To leverage the advantages of both modalities and achieve modal complementarity, we propose the dual-stream network EMSNet, which uses RGB and DSM images for urban resource segmentation.

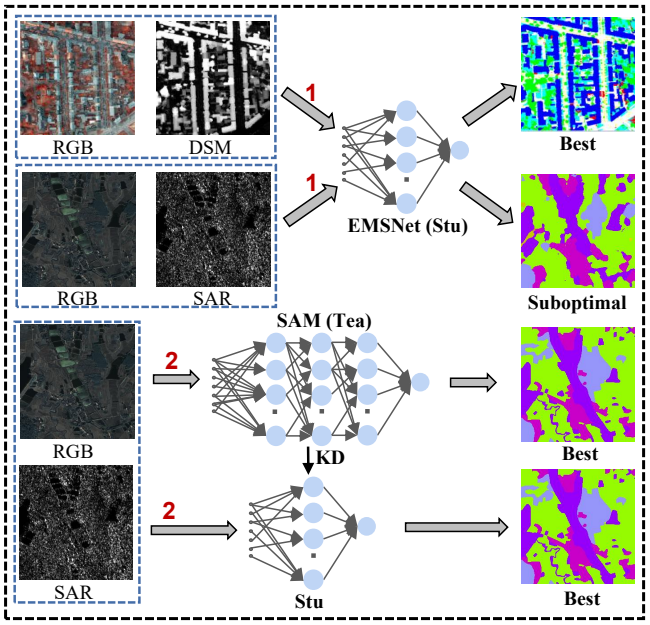

Fig. 1. The overall operating framework of proposed method.

Unlike traditional multimodal networks, EMSNet employs an encoder-decoder structure in both independent streams, as illustrated in Figure 2. This independent structure captures different features and unique information from RGB and DSM data, enabling each modality's advantages to be reflected in the final result. The encoder uses symmetric feature extraction to generate feature maps with matching dimensions from RGB and DSM images, denoted as FPR_{1-5} and FPD_{1-5}. The decoder mirrors the encoder, with FPR_5 and FPD_5 as the inputs. The outputs of the two decoding streams are refined through trans-convolution blocks and multi-level feature fusion. Given the challenges of generating a complete segmentation map from the DSM image, we employ SAGate as a bridge to continuously introduce RGB features into the DSM stream for guidance. The RGB stream results focus on intra-class segmentation, subdividing each target locally. In contrast, the DSM stream results emphasize inter-class segmentation, using height information to distinguish edges of categories effectively. The integration of the two branch results amplifies the role of DSM data, fully leveraging the advantages of multimodal data.

1) Dual-branch Feature Extraction: In designing the encoder for EMSNet, we prioritized lightweight architecture, employing a simplified residual structure as shown in Figure 2. The encoder achieves a computational cost of $5.8G$ and a parameter count of $8.6M$, significantly lower than ResNet34, which requires $9.2G$ and $21.8M$ respectively. The input X is processed through three convolutional layers to produce

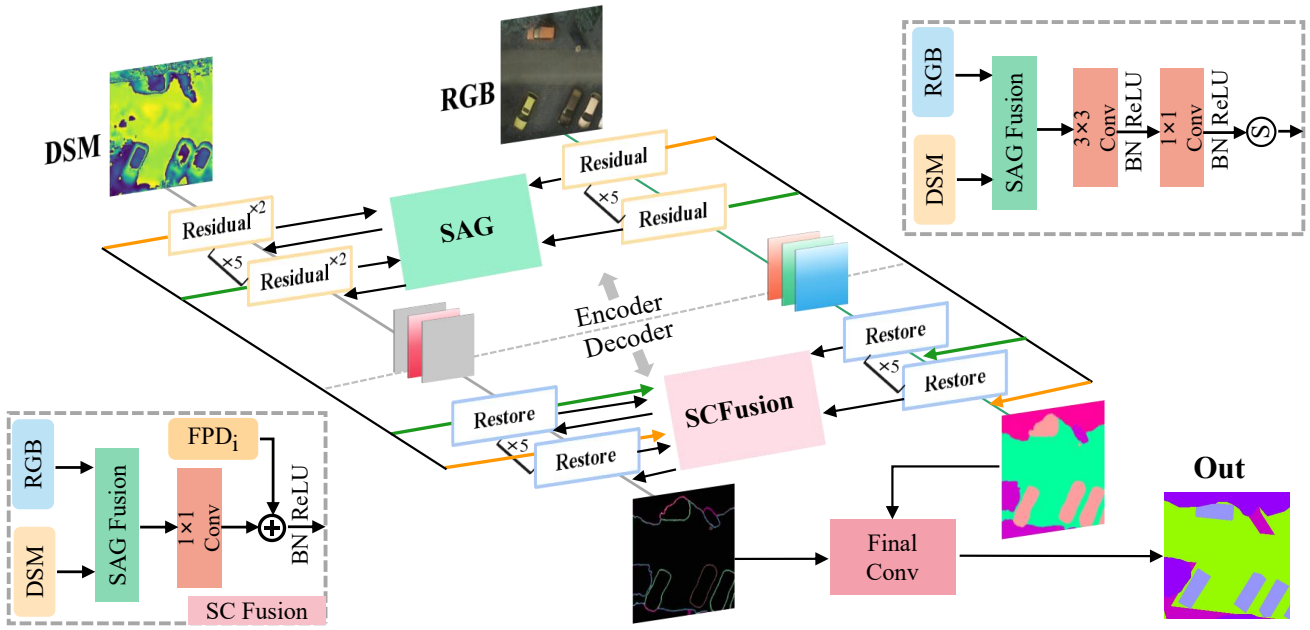

Fig. 2. The overall framework of our proposed model, EMSNet. The model entails dual-stream structure for both the encoder and decoder, with the Separation-and-Aggregation Gate (SAG) being crucial components for cross-modal feature fusion. The encoder comprises five residual layers, '×2' represents the number of residual blocks. Similarly, the decoder employs a five-layer structure mirroring the encoder. The '+' and 'S' represent channel concatenation and sigmoid function, respectively.

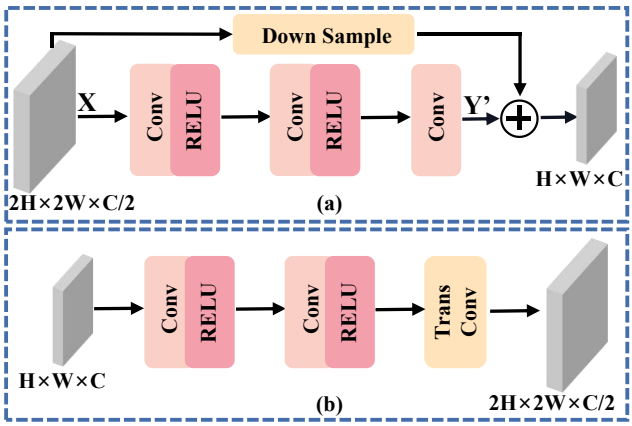

Fig. 3. The pipeline of Residual and Restore. (a) Residual Layer. (b) Restore Block.

$Y^{'}$, a feature map with reduced channels and dimensions. On the residual side, a 1×1 convolution is used for a skip connection, generating $X^{'}$. The final output features are obtained by combining $X^{'}$ and $Y^{'}$. Mathematically, it can be expressed as

$$\boldsymbol{Out} = \mathrm{RE}LU(Conv^{*3}(\boldsymbol{x}) + DS(\boldsymbol{x})) \tag{1}$$

where $Conv^{*3}$ denotes three convolution layers, DS denotes down-sample of residual side.

The RGB and DSM branches consist of five-layer residual structures. In the DSM branch, the initial convolutional channel is set to 1, and each layer comprises two residual blocks, ensuring the extraction of DSM features. The dimensions of the input RGB and DSM images are denoted as (H, W). Following the five-layer residual, the feature maps FPR_{1-5} and FPD_{1-5} exhibit dimensions of $(H/2, W/2)$, $(H/4, W/4)$, $(H/8, W/8)$, $(H/16, W/16)$, and $(H/32, W/32)$.

Given the limited contextual information available in DSM images, we incorporate RGB features at each layer of the DSM branch to guide its processing. RGB features convey rich visual information, such as color and texture, while DSM features capture height-related data. The integration of these two modalities enhances the DSM branch's capability to interpret object shapes and structures, enabling a more holistic understanding of the scene. At each layer, the RGB features are fused with DSM features to form the input for the subsequent layer in the DSM branch. This fusion process can be mathematically represented as follows:

$$\boldsymbol{FPD_i} = SAG(\boldsymbol{FPR_{i-1}}, RES^{\times 2}(\boldsymbol{FPD_{i-1}})) \tag{2}$$

where $RES^{\times 2}$ denotes the two residual blocks at each layer of the DSM branch, and SAG represents the SAGate feature fusion module.

2) Feature Reconstruction: The decoder reconstructs the high features aiming to match the dimensions of the original input. This process involves upsampling and convolution, as commonly employed in prior research. Initially, upsampling involves bilinear interpolation, though parameter-free, resulting in a significantly increased computational load per subsequent convolution. Additionally, reducing channels through convolution before employing upsampling to recover the size can introduce inaccuracies in the final results. When the generated pixels exhibit excessive sameness, potentially leading to the loss of target edges.

Therefore, we introduced restore blocks within the five layers of the decoder, depicted in Figure 2. Initially, a con-

volutional layer is employed to reduce the channels of high features, followed by applying a trans-convolutional layer to restore the dimension of the features before output. This effectively addresses the above problem with minimal parameter costs. The restore block enhances the representation of EMSNet, yielding superior results, particularly given the complex structure of RSI. Meanwhile, To improve the perception of details, we propose applying multi-level feature fusion within the two branches. This involves passing low-level fine-grained features and high-level abstract semantics from the encoder to the decoder, thereby enriching the spatial detail information. Meanwhile, the decoder mirrors the structure of the encoder, with the DSM branch continuously guided by RGB features to enhance feature recovery. Therefore, the RGB and DSM decoded stream can be represented as follows:

$$\boldsymbol{RSR_i} = RS(Cat(\boldsymbol{RSR_{i+1}}, \boldsymbol{FPR_i})) \tag{3}$$

$$\boldsymbol{RSD_i} = RS(Cat(\boldsymbol{RSD_{i+1}}, SAG(\boldsymbol{RSR_{i+1}}, \boldsymbol{FPD_i}))) \tag{4}$$

where RS stands for restore block, and Cat represents the concatenation operation.

The final output can be expressed as:

$$\boldsymbol{Out} = Sig(Conv_C^{\times 2}(SAG(\boldsymbol{RSR_1}, \boldsymbol{RSD_1}))) \tag{5}$$

where Sig represents the $Sigmoid$ function, while $Conv^{\times 2}$ denotes two convolution layers and C is the output channel, which is 6 in this paper.

3) Separation-and-Aggregation Gate Module: To promote the integration of features across two modalities, we employ the SAG [40] for feature fusion, delineated in Figure 4. The SAG contains two operations: Feature Separation (FS) and Feature Aggregation (FA).

The FS is applied to calibrate input feature maps denoted as RGB_{in} and DSM_{in}. First, the primary objective is to attain less noisy filter maps, achieved through global pooling and MLP operations. RGB_{in} and DSM_{in} are cascaded and pooled, yielding the cross-modality attention vector $I = (I_1 ... I_{2C})$. This vector serves as the global descriptor for the entire input. Subsequently, the cross-modality gates W_{rgb} and W_{dsm} are derived from vector I using an MLP network. Finally, filtered maps RGB_{filter} and DSM_{filter} result from the multiplication of inputs with the cross-modality gates. Through an addition operation, accurate feature maps RGB_{rec} and DSM_{rec} are obtained from the filter maps and inputs. The concept is to leverage visual data from the RGB feature to reduce noise in DSM and utilize height information from the DSM feature to calibrate the texture of RGB. This calibration ensures the robustness of fused feature maps by integrating favorable information from the inputs.

The FA is employed to combine features from distinct modalities. The color and edge features from the RGB image and the height feature from the DSM are harmonized at a suitable spatial location. RGB_{rec} and DSM_{rec} are concatenated, yielding two spatial-wise gates, G_{rgb} and G_{dsm}, derived from the concatenated vector using two distinct mapping functions. Following applicating the softmax function, the weights A_{rgb} and A_{dsm} are obtained.

$$\boldsymbol{A_{rgb}^{i,j}} = \frac{\boldsymbol{e^{G_{rgb}^{i,j}}}}{\boldsymbol{e^{G_{rgb}^{i,j}} + e^{G_{dsm}^{i,j}}}}, \boldsymbol{A_{dsm}^{i,j}} = \frac{\boldsymbol{e^{G_{dsm}^{i,j}}}}{\boldsymbol{e^{G_{rgb}^{i,j}} + e^{G_{dsm}^{i,j}}}} \tag{6}$$

The final merged feature F_{out} is obtained by multiplying the inputs and the weights. The formula is expressed as

$$\boldsymbol{F_{out}} = \boldsymbol{RGB_{in}} \times \boldsymbol{A_{rgb}} + \boldsymbol{DSM_{in}} \times \boldsymbol{A_{dsm}} \tag{7}$$

C. Knowledge Distillation

1) Motivation for Knowledge Distillation: As mentioned earlier, although DSM images provide stable and accurate height data, they are updated infrequently and are unsuitable for dynamic monitoring. In addition, DSM images are obtained through optical sensors or LiDAR, and clouds and insufficient lighting can lead to incomplete data or reduced quality. In contrast, SAR images are unaffected by weather and lighting conditions and offer wide coverage, effectively compensating for the limitations of DSM images. However, SAR images present challenges due to complex scattering mechanisms, resulting in reflections, shadows, and speckle noise. EMSNet needs more complex architectures to achieve generalization in SAR scenarios, which contradicts our goal of maintaining lightweight. These challenges limit the practicality of EMSNet. Improving the performance in SAR scenarios while ensuring lightweight is another serious problem that this paper urgently needs to solve. Knowledge distillation (KD) is a trustworthy solution that can improve performance without changing the structure of the model. Recent advancements, such as large-scale models like the Segment Anything Model (SAM), have significantly improved the effectiveness and insight of KD techniques.

2) Fine-tuning the Teacher Model: As shown in Figure 5, in our implementation, we fine-tune the SAM using the RSI to handle SAR images more effectively. A notable fine-tuning technique is Low-Rank Adaptation (LoRA). Compared with conventional methods that adjust all parameters in SAM, LoRA allows SAM to update a small number of important parameters during training. This approach not only preserves SAM's robust segmentation performance but also reduces computational overhead. Specifically, we utilize pre-trained weights and many parameters of the SAM encoder. Each transformer block incorporates a LoRA bypass, allowing for the targeted adjustment of specific parameters. Given the significance of the attention mechanism in transformers, we apply LoRA to the query, key, and value projections within this mechanism. In our practice, we focus on applying LoRA to the projection layers of both the value and query, which yields impressive results. Furthermore, we utilize the default prompts in the prompt encoder, as additional prompts necessitate more extensive annotations. The LoRA process unfolds as follows:

$$\begin{aligned} \Delta W &= B \times A \\ W'x &= Wx + \Delta Wx \end{aligned} \tag{8}$$

where $W \in \mathbb{R}^{a \times b}$ is a pre-trained weight matrix, $A \in \mathbb{R}^{r \times b}$ and $B \in \mathbb{R}^{a \times r}$ are two trainable low-rank decomposition matrices,

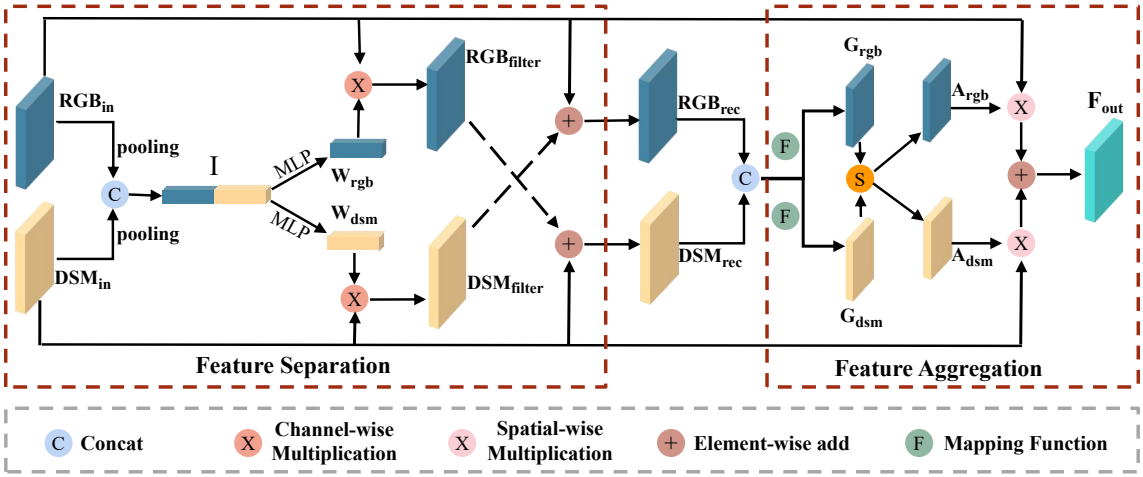

Fig. 4. The architecture of the Separation-and-Aggregation Gate (SAG), which contains two parts, Feature Separation (FS) and Feature Aggregation (FA).

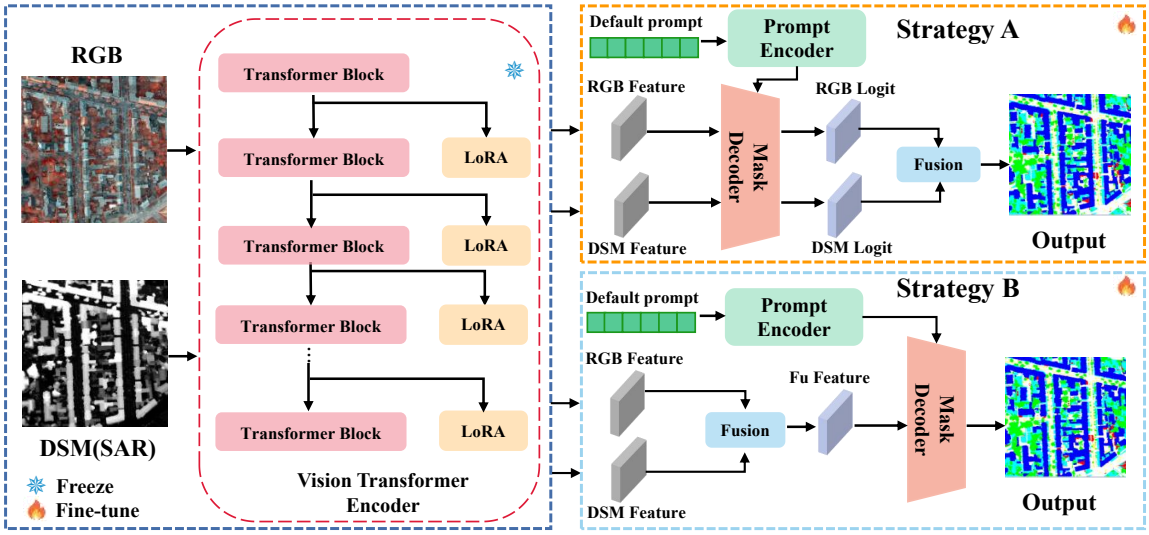

Fig. 5. Two fine-tuning strategies are explored for the Segment Anything Model (SAM). The left encoder remains frozen, and its parameters are not updated. Strategy A and Strategy B represent two distinct methods for fine-tuning the decoder.

ΔW is the update part, W' is the new weight matrix, and r is the rank of LoRA, which is set to 4 in our work.

In Figure 5, Strategies A and B represent two distinct approaches for retraining the mask decoder. In strategy A, the features from the two modalities are inputted to the mask decoder. Subsequently, we employ SAGate to integrate the two logit information streams, producing the final output. This retraining of the decoder enables it to leverage multimodal information, resulting in optimal performance effectively. Conversely, we seek to minimize SAM's computational load by fusing the features before inputting them into the decoder in strategy B. However, this early fusion risks obscuring the inherent information of each modality, which can impair the decoder's ability to fully comprehend each modality's characteristics, ultimately leading to suboptimal results.

3) Knowledge Transfer Process: To improve the efficiency of EMSNet in SAR scenarios, we employ KD to extract knowledge from SAM. Two prevalent distillation methods, logit-based and feature-based, can be used for various tasks, including segmentation. We dynamically combine these methods to achieve optimal results, as illustrated in Figure 6.

KD involves the simulation of the teacher model from the student, emphasizing structural similarity. MGD [41] introduces a new KD paradigm, generating teacher features instead of mere simulation. First, a random mask selectively covers the pixels of the student features, and then teacher features are generated through a simple block. In our context, we select the feature maps from SAM as teacher features, denoted as $Feature_{tea}$. For the five-layer feature maps of EMSNet, we choose the maps with the same dimensions as $Feature_{tea}$ (though the number of channels differs), denoted as $Feature_{stu}$. Using MGD, we perform feature-based KD to generate teacher features. It is crucial to note that the pre-trained encoder in SAM is not entirely stable for processing RSI, as training data differs from remote sensing data. This instability can cause SAM to generate error features, risking

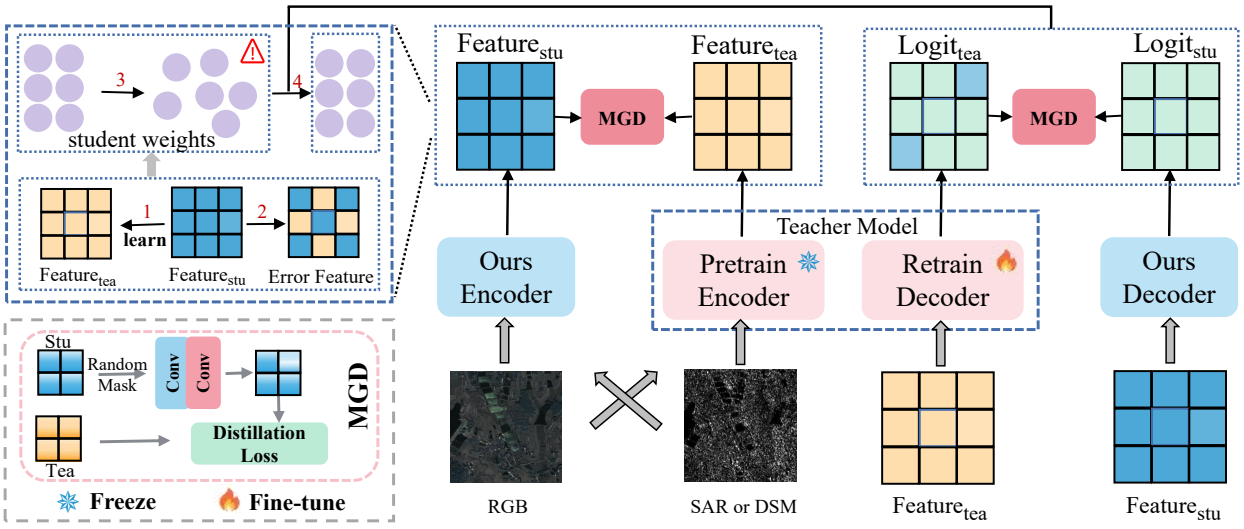

Fig. 6. Details of the knowledge distillation process. The right side offers a macroscopic overview of the entire process, while the left side provides a granular depiction of feature learning. Here, the purple circle represents the parameter update status of the student model. The specifics of the key component, Masked Generative Distillation (MGD), are shown in the pink box.

parameter update confusion in the student model, as depicted in step 3 of Figure 6. To mitigate this risk, we also introduce logit-based KD. Due to the mask decoder being fine-tuned with remote sensing datasets, it can correct error features and generate new logits. This approach helps constrain the stability of the entire network's parameter updates and reinforces feature extraction.

In general, considering the EMSNet's structure and the training process of SAM, KD occurs in the four branches of the encoder and decoder. However, because the frozen encoder will produce error features and the simple mask decoder cannot yield reliable logits, we use the loss of the final result and the ground truth as a threshold. When the distillation loss at the feature or logit stage exceeds a two-fold threshold, these losses are excluded from backpropagation. This dynamic integration of feature-based and logit-based KD ensures a more efficient distillation.

D. Loss Function

The loss function is composed of two components: LossM and LossD. LossM, a cross-entropy function, quantifies the discrepancy between the prediction and the ground truth. Given the inherent imbalance in datasets across various categories, we assigned weights to these categories. On the other hand, LossD is formulated as the mean squared error (MSE) function, signifying the distillation loss. The expressions for LossM and LossD are presented below

$$LossM(\boldsymbol{p},\boldsymbol{g}) = -\sum_{i,j} \boldsymbol{g}_{(i,j)} \times log\,\boldsymbol{p}_{(i,j)} \tag{9}$$

where p represents the prediction, and g represents the ground truth.

$$LossD(\boldsymbol{s},\boldsymbol{t}) = \frac{1}{w \times h} \sum_{i=0}^{w} \sum_{j=0}^{h} (\boldsymbol{s}_{(i,j)} - \boldsymbol{t}_{(i,j)})^2 \tag{10}$$

where s is student features, t is teacher features, w and h represent width and height. Total loss can be expressed as

$$Loss = LossM + (LossD_1 + ... + LossD_4)/4 \tag{11}$$

where $LossD_i$ represents the distillation errors of the four features.

IV. Experiment

In this section, we first evaluate the feasibility of EMSNet in segmentation tasks by comparing it with seven other classic two-stream networks on the Potsdam and Vaihingen datasets. Second, to validate the effectiveness of the distillation strategy, we conduct experiments on the Potsdam dataset. Third, facing complex SAR images and a broader range of scene categories, we compare the proposed method with other methods on the WHU and DFC datasets to assess the generalization of our approach. Subsequently, we perform ablation experiments to ascertain the importance of each key component. Finally, we calculate the computation, parameters, and latency of EMSNet and compare them with other methods.

A. Execution Details and Datasets

Training Details. The proposed method is developed using the PyTorch framework, which is publicly accessible. The maximum epoch of training iterations is 60 to achieve convergence. The initial learning rate is 1.0×10^{-3} and is decayed by a factor of 0.1 every 20 epochs. To improve the computational efficiency, all experiments are conducted on the GPU and utilize a dual-core parallel network. We optimized the network parameters using the Adam optimizer.

Metrics. We employ three recognized metrics for model evaluation: Overall Accuracy (OA), Kappa coefficient, and Intersection-over-Union (IoU). OA quantifies the ratio of accurately predicted pixels in the ground truth. The Kappa coefficient as a reliable measure for segmentation results,

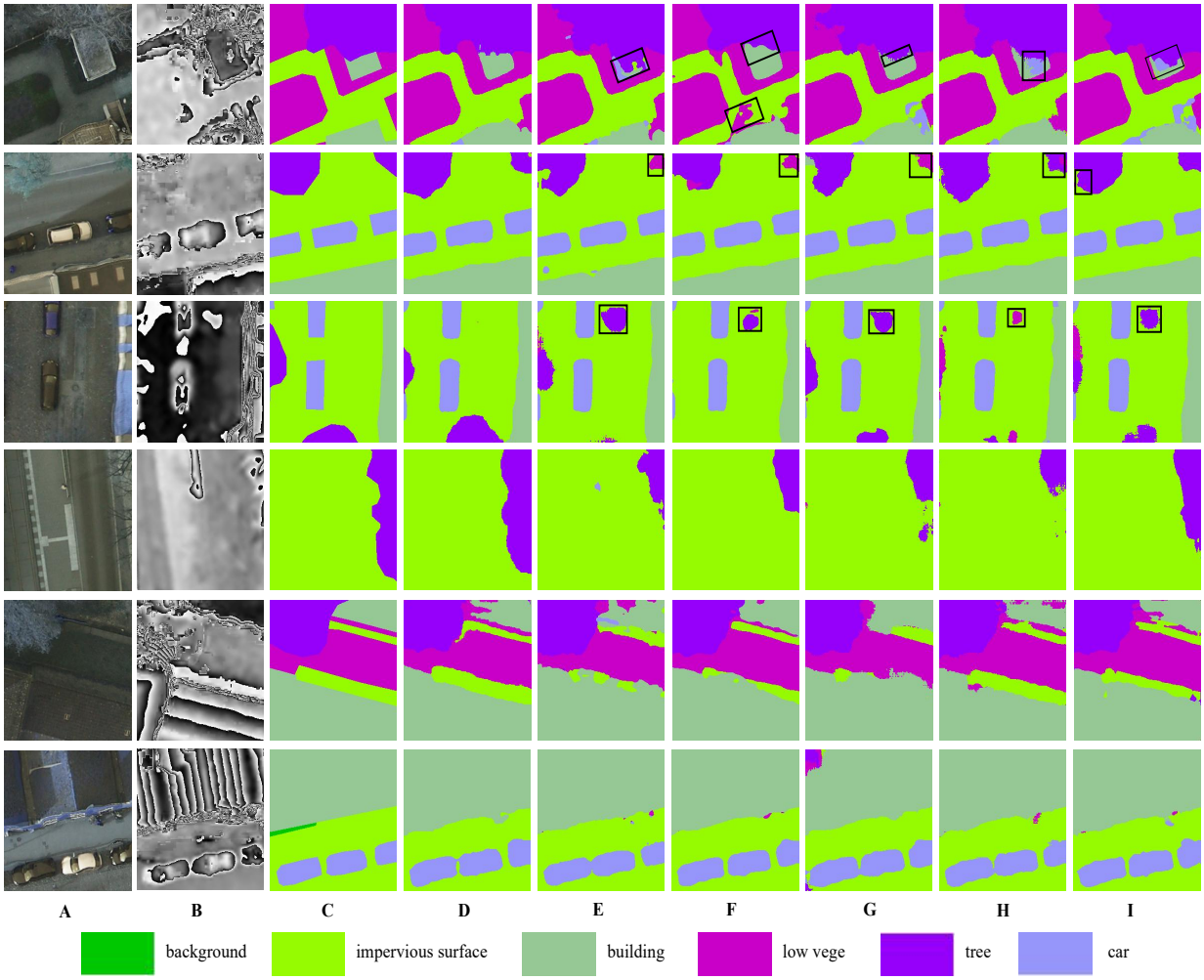

Fig. 7. Segmentation visual results with other multi-modal models on the Potsdam dataset. A: Optical images. B: DSM images. C: Ground Truth. D: EMSNet (Ours) results. E: MCANet results. F: ESANet results. G: RedNet results. H: ACNet results. I: CMGFNet results. The black box shows the difference in local details.

is derived from the confusion matrix. It ensures recall for a smaller percentage of categories when OA is high. IoU, commonly utilized in segmentation tasks, measures the dissimilarity between the outcome and the ground truth.

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$Kappa = \frac{(P_0 - P_e)}{1 - P_e} \tag{13}$$

$$IoU = \frac{Pre \cap Gt}{Pre \cup Gt} \tag{14}$$

where the TP, TN, FP, and FN correspond to True-Positive, True-Negative, False-Positive, and False-Negative, respectively. P_0 denotes the ratio of the sum of diagonal elements to the sum of all elements in the matrix, and P_e signifies the ratio of the sum of elements in the same row and column to the square of all the elements. Here, Pre and Gt represent the predicted result and ground truth, respectively.

$$mIoU = \frac{1}{C} \sum_{1}^{C} IoU_i \tag{15}$$

where C is the total number of classes.

Comparative Baselines. In this paper, we compare our proposed method with several typical multi-modal algorithms. Among them, RedNet, ACNet, and ESANet are suitable for segmenting cars, pedestrians, low vegetation, etc. CMGFNet and BuildFormer are designed to extract buildings. MCANet is used for land cover segmentation in urban areas.

- RedNet [42]: One method utilizes a skip connection structure to establish spatial links between the encoder and decoder. Agent blocks are employed to reduce encoding channels, with RGB and depth maps as input.
- ACNet [43]: ACNet proposes a three-parallel branch architecture, inserting an attention auxiliary module into each encoder layer. This design balances feature distribution, enabling the network to focus more on the effective area of the image. Segmentation is conducted using RGB and depth maps.
- TFNet [44]: A two-stream image fusion network processes multispectral and panchromatic images. TFNet aims to fuse these features and reconstruct a pan-sharpened image from the fused feature map.

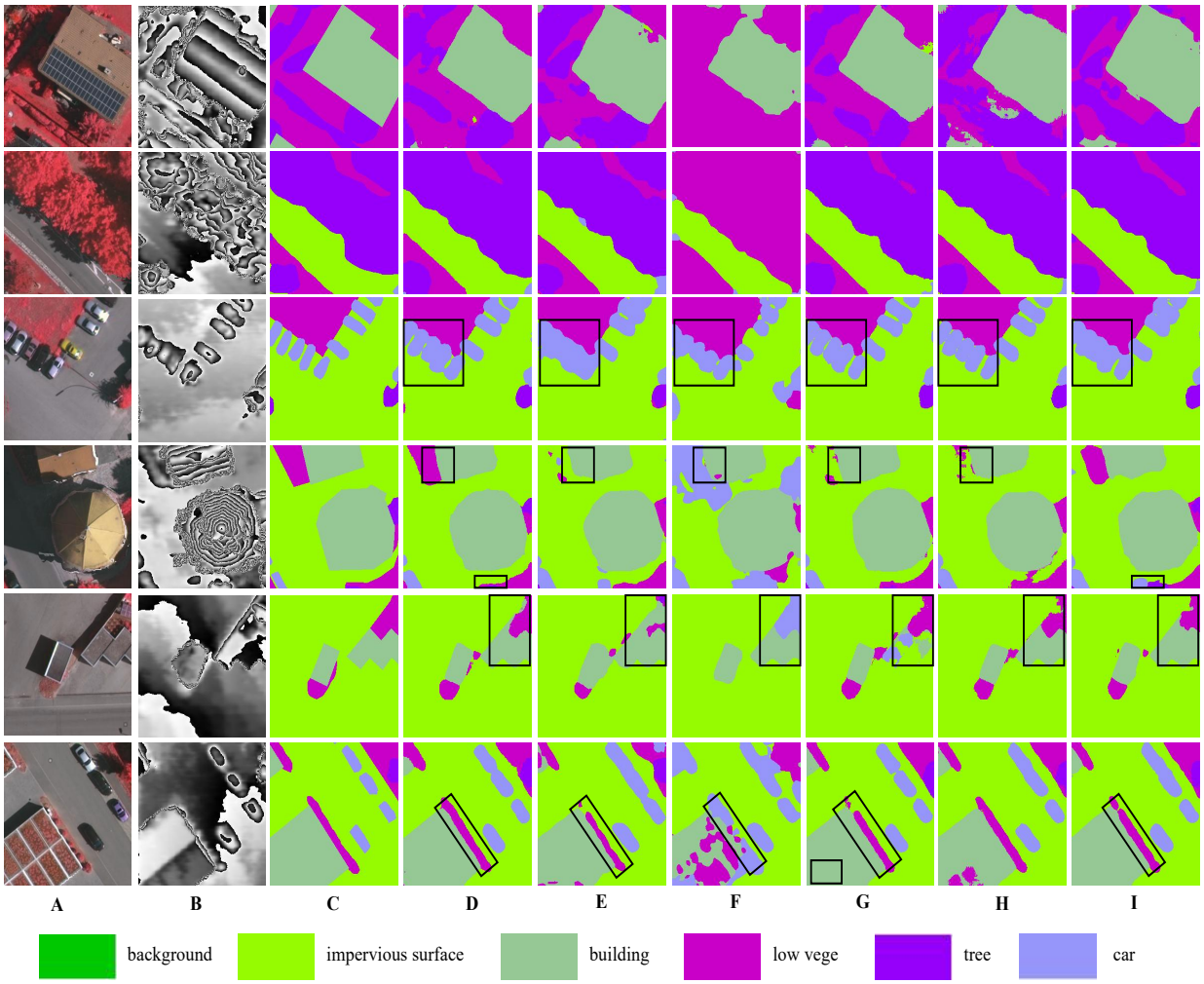

Fig. 8. Segmentation visual results with other multi-modal models on the Vaihingen dataset. A: Optical images. B: DSM images. C: Ground Truth. D: EMSNet (Ours) results. E: MCANet results. F: ESANet results. G: RedNet results. H: ACNet results. I: CMGFNet results. The black box shows the difference in local details.

- ESANet [45]: ESANet introduces a unique learnable upsampling technique, distinct from bilinear upsampling. The decoder integrates the Unet structure to perform segmentation on RGB and depth maps.
- CMGFNet [46]: Designed for building extraction using RGB and DSM images. The two encoders generate modality-specific features, achieving high-precision extraction through feature fusion. The encoder utilizes residual depth-separable convolution for upsampling, improving computational efficiency.
- MCANet [47]: An approach incorporates a cross-modal attention mechanism and a multi-level feature fusion module for land use segmentation using optical and SAR images.
- PACSCNet [48]: A progressive symmetric cascade network with a dual-pyramid decoder is used to extract and merge similarities in cross-modal features.
- FTransUnet [13]: A multi-level multimodal method provides a robust and effective multimodal fusion backbone for semantic segmentation by integrating CNN and Vit into a unified fusion framework.

Datasets. We employ the Potsdam, Vaihingen, WHU-Opt-Sar, and DFC2023 datasets for method evaluation. The Potsdam and Vaihingen encompass RGB and DSM images. In contrast, the WHU-Opt-Sar and DFC2023 contain RGB and SAR images.

- Potsdam Dataset: The dataset comprises 38 IR-RGB and DSM images, each with a resolution of 5 cm and dimensions of 6000×6000 pixels. It is categorized into six classes: impervious surface, building, low vegetation, tree, car, and background.
- Vaihingen Dataset: The Vaihingen consists of 33 image pairs, featuring a resolution of 9 cm and dimensions of 2000×2500 pixels. Each RGB image is accompanied by corresponding DSM data, and the categories align with the Potsdam dataset.
- WHU-Opt-Sar Dataset: The WHU dataset encompasses an extensive area of approximately $50,000 km^2$ in Hubei Province, China. This dataset comprises 100 NIR-RGB and SAR images, each with dimensions of 5556×3704 pixels. A total of seven classes: farmland, city, village, impervious surface, forest, road, and other.

TABLE I
QUANTITATIVE COMPARISON RESULTS WITH OTHER DUAL-STREAM BASELINES ON THE POTSDAM DATASET.

Method	OA ↑	Kappa ↑	mIoU ↑	Accuracy/IoU									
				imp surface		building		low vege		tree		car	
MCANet	91.0	87.5	78.1	90.9	85.6	99.3	88.5	94.6	88.1	61.0	56.5	96.1	71.8
ESANet	91.9	88.8	81.0	91.9	87.7	98.8	92.3	93.8	88.0	76.3	55.0	98.0	82.1
RedNet	91.1	87.6	80.5	91.2	88.6	97.6	84.7	93.4	86.6	72.8	59.4	98.9	83.2
ACNet	92.2	89.3	82.6	91.1	87.6	97.4	92.5	94.2	88.3	92.8	66.1	98.3	78.4
CMGFNet	92.8	90.0	82.0	91.5	88.8	99.4	92.9	95.7	90.1	81.9	66.0	98.7	72.4
PACSCNet	92.0	88.9	82.5	93.2	90.1	96.4	90.4	92.1	87.0	91.7	59.1	97.9	85.8
FTransUnet	91.5	88.4	81.8	89.8	87.3	97.4	90.2	93.5	87.0	91.3	62.1	98.9	82.2
Ours	**93.1**	**90.4**	**82.7**	91.7	90.1	98.2	92.0	96.4	90.4	87.7	72.6	98.0	68.7

TABLE II
QUANTITATIVE COMPARISON RESULTS WITH OTHER DUAL-STREAM BASELINES ON THE VAIHINGEN DATASET.

Method	OA ↑	Kappa ↑	mIoU ↑	Accuracy/IoU									
				imp surface		building		low vege		tree		car	
MCANet	93.6	84.3	56.9	97.6	93.0	92.1	88.4	12.9	12.1	94.5	85.9	99.5	05.1
ESANet	94.6	86.7	67.2	97.4	93.8	95.2	89.3	22.1	19.3	99.2	81.5	95.7	52.0
RedNet	94.7	87.1	62.3	97.6	93.9	93.5	90.4	33.8	28.1	98.1	84.2	96.8	14.7
ACNet	94.8	87.1	64.5	97.8	93.9	94.8	89.8	21.6	19.5	98.9	84.3	97.9	35.1
CMGFNet	93.9	84.6	62.9	98.6	93.3	91.3	89.2	06.7	05.3	90.1	79.5	75.5	47.1
PACSCNet	86.0	80.8	65.5	83.9	72.0	88.5	83.7	70.7	51.6	90.7	84.5	98.0	35.5
FTransUnet	85.8	80.2	66.5	86.2	75.8	93.8	85.6	39.7	33.2	93.8	80.4	98.3	57.5
Ours	**94.9**	**87.4**	**70.2**	97.7	93.8	95.3	89.4	32.1	27.8	93.1	88.2	73.4	51.8

TABLE III
QUANTITATIVE ANALYSIS OF DIFFERENT SAM TRAINING STRATEGIES ON THE POTSDAM DATASET.

Method	OA ↑	Kappa ↑	mIoU ↑	Accuracy/IoU									
				imp surface		building		low vege		tree		car	
A	**94.6**	**92.0**	**88.9**	**89.6**	**88.2**	**99.2**	**92.4**	**92.7**	**84.3**	96.9	**92.3**	97.9	**87.5**
B	93.5	90.4	87.5	88.1	85.7	97.7	90.3	92.2	83.7	**98.3**	**92.3**	**98.7**	85.5

TABLE IV
QUANTITATIVE ANALYSIS OF SAM-BASED DISTILLATION STRATEGIES ON THE POTSDAM DATASET.

Method	OA ↑	Kappa ↑	mIoU ↑	Accuracy/IoU									
				imp surface		building		low vege		tree		car	
Original	93.1	90.4	82.7	**91.7**	**90.1**	98.2	92.0	**96.4**	**90.4**	87.7	72.6	98.0	68.7
Kd-A-e	92.9	89.3	83.3	90.6	85.6	99.2	91.3	84.8	79.4	73.8	73.0	97.3	87.5
Kd-A-d	91.6	87.6	81.1	84.1	82.2	**99.6**	**92.4**	93.4	66.2	80.3	78.1	97.6	86.6
Kd-B-ed	93.1	89.8	85.9	88.3	84.7	97.4	91.0	93.3	81.1	91.4	88.3	**98.2**	84.1
Kd-A-ed	**94.1**	**91.2**	**87.6**	88.8	87.4	99.2	92.1	89.8	79.3	**97.6**	**91.5**	98.0	**87.9**

- DFC2023 Dataset: The DFC2023 dataset provides satellite images, digital surface models, and semantic labels of buildings in 17 cities from six continents. The dataset contains 1773 images, each sized at 512×512 pixels, with accurately annotated semantic labels.

In experiments, each image in all datasets is segmented into multiple 256×256 images, subsequently partitioned in a 7:1:2 ratio for the training, validation, and test sets, respectively.

B. Performance Evaluation

In this section, we compare the proposed EMSNet with seven other multi-modal networks on the DSM dataset mentioned in Section IV-A.

The quantitative comparisons for the Potsdam and Vaihingen datasets are detailed in Table I and Tabel II. We evaluated seven dual-stream segmentation networks, which are MCANet, ESANet, RedNet, ACNet, CMGFNet, PACSCNet and FTransUnet. Notably, in all experiments, we excluded the background without impacting the training process or statistical measures. The results demonstrate EMSNet's superiority across all metrics, surpassing other methods in Overall Accuracy (OA), Kappa, and mIoU. MCANet designed for RGB and SAR images, doesn't perform optimally on these datasets. In the loss function, we use the weight matrix according to the percentage of each category. This strategy resulted in improved performance for the car with the lowest percentage. However, this emphasis may have adversely affected the results of other smaller percentage categories, such as trees in Potsdam and low vegetation in Faihingen. Consequently, future efforts should concentrate on optimizing performance for unbalanced

TABLE V
QUANTITATIVE COMPARISON RESULTS OF DIFFERENT METHODS ON THE WHU-OPT-SAR DATASET.

Method	OA ↑	Kappa ↑	mIoU ↑	Accuracy/IoU													
				farmland		city		village		water		forest		road		others	
SAM	81.7	74.2	50.7	68.4	62.1	58.1	34.1	83.0	50.8	88.0	76.6	90.8	83.2	91.0	24.3	58.7	23.7
MCANet	79.0	70.4	47.1	70.1	59.9	57.4	25.0	65.3	44.1	85.0	71.8	87.9	81.4	51.9	34.5	47.9	13.2
BuildFormer	80.0	71.8	48.1	65.9	59.0	56.9	30.3	75.6	49.2	87.6	71.8	90.4	82.1	66.9	23.9	58.9	20.6
TFNet	80.2	72.1	**49.8**	68.2	60.2	74.8	43.4	75.4	51.3	85.9	69.5	88.5	81.7	69.7	20.6	66.5	22.4
Ours	78.4	69.7	46.2	61.9	56.4	58.3	24.6	65.8	40.7	86.4	70.8	91.2	82.7	68.1	30.4	64.4	17.8
Ours-kd	**81.2**	**73.3**	49.1	70.2	63.5	57.5	26.5	68.2	44.8	85.0	73.6	91.8	82.5	72.3	26.9	61.5	26.1

TABLE VI
QUANTITATIVE COMPARISON RESULTS OF DIFFERENT METHODS ON THE DFC2023 DATASET.

Method	OA ↑	Kappa ↑	mIoU ↑	Accuracy/IoU			
				background		building	
SAM	92.8	80.0	82.3	95.1	91.0	85.3	73.5
BuildFormer	90.0	71.4	75.8	94.3	87.8	75.8	63.8
MCANet	89.9	70.6	75.2	95.2	87.9	72.5	62.6
TFNet	89.8	71.3	75.7	93.6	87.6	77.2	63.8
Ours	89.6	69.1	74.2	95.7	87.6	69.4	60.8
Ours-kd	**90.7**	**73.2**	**77.1**	95.0	88.6	76.5	65.6

categories.

Figure 7 and Figure 8 depict the visual comparison results between our method and other baselines. In Figure 7, the first four rows show that all baselines exhibit more instances of misrecognition and underrecognition for the low vege and tree categories. While our method shares a similar bias, it effectively mitigates these errors, aligning with the findings from the quantitative comparison. The fifth row in Figure 7 highlights that our method delivers superior visuals in scenarios where different categories are interleaved. The utilization of a linear interpolation upsampling method by other baselines introduces errors in complex, alternating, multi-category images, which might be related to our inference in Section III-B2. The final row illustrates that the algorithms yield commendable recognition results for images with distinct target contours, with some baselines displaying minimal misrecognized regions.

C. Knowledge Distillation Results

In this part of the experiment, we evaluated the impact of two training strategies on SAM and assessed the feasibility of different distillation techniques using the Potsdam dataset.

The training results of SAM on the Potsdam dataset are in Table III. A and B correspond to the strategies depicted in Figure 5 strategy A and B, respectively. The result validates that empowering the decoder to assimilate information from RGB and DSM features leads to enhanced accuracy. In strategy A, SAM effectively learns the distribution of optical and DSM data, thereby improving its interpretation of the entire scene. This finding corroborates our hypothesis in Section III-C2 and introduces further insights for subsequent knowledge transfer.

Table IV displays the quantitative results of EMSNet following knowledge distillation on the Potsdam dataset. With e and d representing knowledge transfer for encoding features and decoding logits, Kd-A and Kd-B denote distillation using Strategy A and Strategy B. As previously discussed, relying just on feature distillation results in poorer performance. In addition, SAM's decoder is a simplified version of the transformer structure and has undergone fine-tuning, but it is difficult to generate reliable logits entirely. Consequently, employing distillation for logits weakens the functionality of EMSNet, yielding subpar results. The final findings indicate that optimal performance of the entire distillation strategy is attained only when both aspects complement each other. The analysis reveals a substantial improvement in all three metrics: OA, Kappa, and mIoU increased by 1%, 0.8%, and 5.1%, respectively. Compared to other baselines, it outperformed by 1.3%, 1.2%, and 5.2%, respectively. This noteworthy enhancement validates the effectiveness of our ultimate knowledge distillation strategy. In Figure 9, visualization results post-distillation depict significant improvement in EMSNet's recognition of categories like trees, cars, etc., leading to a reduction in visual errors.

D. Complex Scenario Assessment

In this section, to verify the generalization of the proposed KD method, we conducted further experiments using the WHU and DFC2023 datasets, which contain RGB and SAR images. The former contains more categories, and the latter is used for building extraction.

Tables V and VI show the qualitative comparison results on the two datasets, and Figure 10 shows the visualization results of the WHU dataset. Due to the complexity of SAR images, EMSNet performs poorly on the WHU and DFC2023, which is expected and acceptable. To keep the lightweight of EMSNet, we avoid convolution layer stacking and complex structures like transformer and attention mechanisms, which pose challenges for feature extraction from SAR images. Nevertheless, by integrating SAM, EMSNet still outperforms other advanced methods, underscoring the effectiveness of our distillation strategy. It is important to note that the WHU dataset contains more categories, and presents a significant challenge for segmentation, resulting in EMSNet performing substantially below other baselines. Consequently, the effect of KD is more pronounced. In contrast, the building is the only foreground in the DFC2023 dataset, leading to a smaller performance gap between EMSNet and other baselines.

These results demonstrate the effectiveness of the proposed KD algorithm in SAR scenarios. The lightweight design of EMSNet reduces computing resource requirements, which is

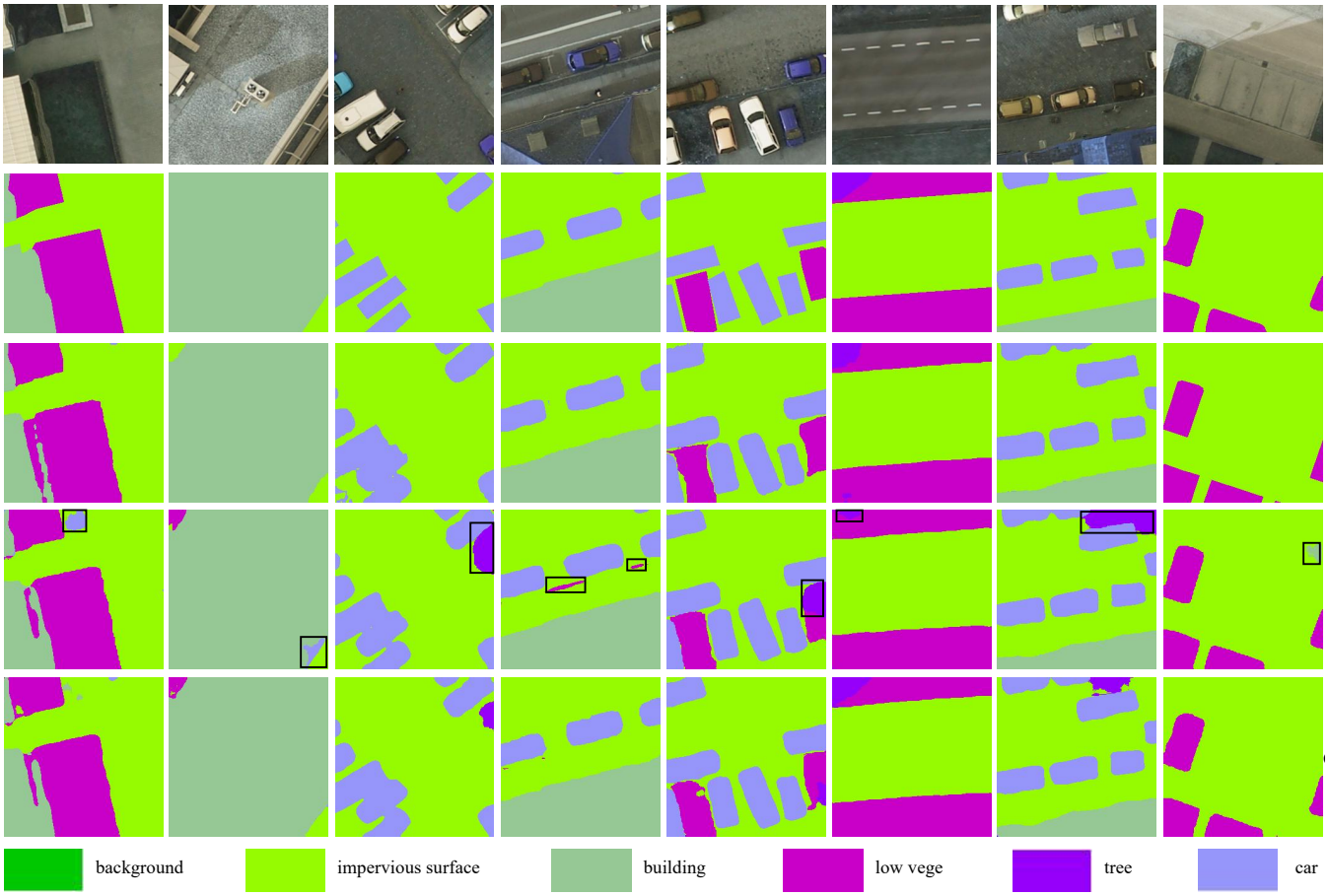

Fig. 9. The knowledge distillation visual results on the Potsdam Dataset. From top to bottom are Optical images, Ground Truth, SAM results, EMSNet results, EMSNet-kd results. The kd represents knowledge distillation. The black boxes highlight detail improvements.

crucial for practical applications. The proposed distillation strategy optimizes EMSNet's performance while maintaining its simplicity. These advantages make EMSNet an attractive choice in resource-constrained and enhance the practical application of multimodal algorithms.

E. Ablation Experiment

We conducted ablation experiments on the Potsdam dataset, dividing this section into two parts. One is to explore the contributions of RGB and DSM images to segmentation results, and the other is to examine the impact of different levels of guidance in the DSM branch on the outcomes.

Table VII displays the results of the ablation study. In this context, D and R indicate the provision of DSM and RGB images to EMSNet, respectively. In multi-category tasks, leveraging visual information such as color and texture usually yields better results. DSM images contain only height information and lack sufficient contextual details. As a result, RGB images with visual features significantly show better performance compared to DSM images. The e_i and d_i indicate guidance from RGB features at the i_{th} layer of the encoder and decoder in the DSM stream. We select the second and fifth layers to fusion features, representing low-level and high-level features, respectively, as these are used in other works commonly. The results show that feature fusion at different layers has a certain impact on the DSM branch, but feature guidance at each level is necessary to harness fully the potential of RGB data. Most existing works employ early and mid-fusion strategies. However, RGB and DSM data exhibit substantial attribute differences, making early fusion inadequate for fully leveraging the diverse information sources. In addition, mid-fusion leads to insufficient interaction between cross-modal data at specific stages. This results in DSM information being inconsistently rich, compressed, or even lost, thereby confusing parameter updates and potentially degrading performance. Consequently, these fusion strategies are unsuitable for the proposed two-stream parallel network.

It is worth noting that impervious surfaces lack significant color variations in RGB images. The height information from the DSM images can effectively distinguish between them, leading to improved results. For buildings with shadow noise, the DSM features help to reduce the noise, which enhances the results significantly for this category. Cars and low vegetation have similar heights in the DSM image, so a higher percentage of pixels for low vegetation leads to poorer car results.

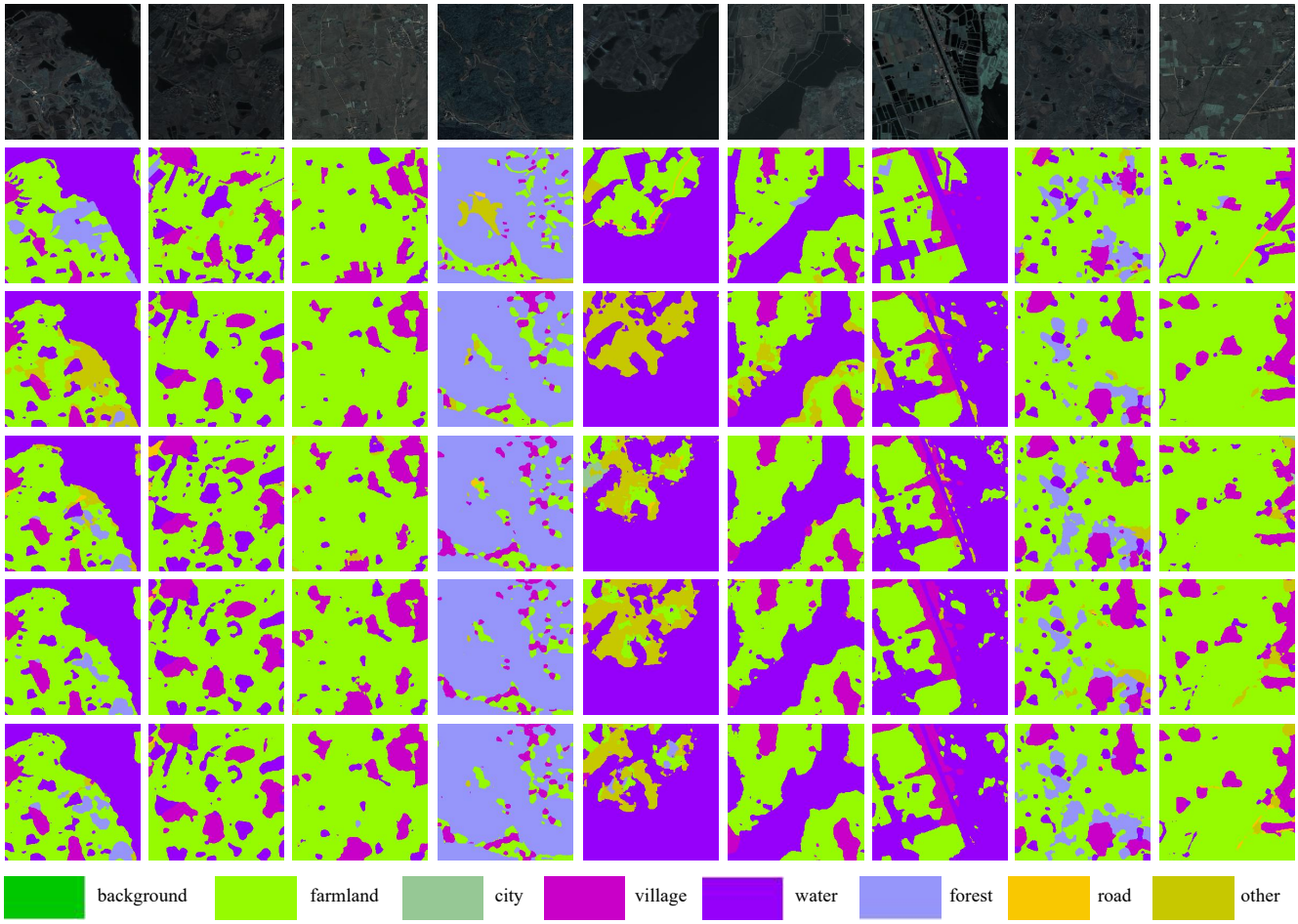

Fig. 10. The images from WHU-OPT-SAR dataset and the visual results produced by different models. From top to bottom are Optical images, Ground Truth, MCANet results, TFNet results, EMSNet (Ours) results, EMSNet-kd results. The kd represents knowledge distillation.

F. The Efficiency of Networks

In practical applications, the latency of algorithms holds significant importance, particularly in large-scale remote sensing data processing. Meanwhile, the number of parameters and computations determine the feasibility of deploying algorithms on mobile platforms with constrained resources. Therefore, in this experiment, we assessed these metrics.

Table VIII illustrates the computation, parameters, and latency of the above methods, all assessed on the RTX3090Ti. Latency tests involved 1958 images with a batch size of 2, mirroring real-world conditions closely. MCANet and SAM adopt complex attention structures, while RedNet, ACNet, and CMGFNet utilize deep residual structures in the encoder. Consequently, these baselines do not possess any advantage in terms of network efficiency. ESANet initially employs max-pooling to diminish the dimensionality of the input image, significantly reducing the computation of the entire network. However, this operation selectively retains only the most dominant features of the target, leading to the loss of other relevant information, which is detrimental to the segmentation task. Contrarily, EMSNet achieves a minimal parameter and reduces computational complexity through simplified feature extraction and reconstruction modules, lowering hardware requirements. Moreover, its utilization of multiple cross-modal feature fusion techniques ensures accuracy, leading to optimal segmentation performance. Additionally, its low latency ensures real-time effectiveness in practical applications.

G. Experiment Discussion

All the experiments presented above fully demonstrate the feasibility, effectiveness, and generalization of the proposed EMSNet. Additionally, EMSNet exhibits advantages over other methods in terms of computation, parameter, and latency, thus affirming its capability to achieve a superior balance between accuracy and efficiency in segmentation tasks. Although we have made progress in our study, it is important to note certain challenges in the experiment. In dealing with intricate remote sensing scenarios, leveraging multimodal data proves advantageous for enhancing performance. However, a notable issue arises from the imbalance in the proportion of multi-category pixels within existing datasets. This imbalance adversely affects the performance of categories with smaller proportions, presenting a common challenge for all networks. Although prior studies have attempted to address this issue by devising specialized loss functions, a complete resolution remains elusive. Future works should prioritize devising solu-

TABLE VII
EVALUATION METRICS SCORE OF ABLATION EXPERIMENT ON THE POTSDAM DATASET. EXPERIMENTS WERE CONDUCTED ON MULTI-MODAL DATA AND FEATURE FUSION STRATEGIES RESPECTIVELY.

Method	OA ↑	Kappa ↑	mIoU ↑	Accuracy/IoU									
				imp surface		building		low vege		tree		car	
D	51.1	34.8	35.2	56.8	40.4	99.8	55.8	16.7	13.8	05.6	02.2	99.3	63.7
R	87.4	82.5	78.2	91.0	78.4	81.3	70.7	92.1	86.4	84.3	70.6	98.6	85.0
D+R	90.1	86.3	78.8	90.8	88.0	98.9	83.4	89.0	84.0	79.3	58.4	97.6	80.1
D+R+$e_{2,5}$	87.0	81.8	76.7	91.3	76.9	75.1	66.0	95.4	89.6	78.2	67.7	98.6	84.1
D+R+$d_{2,5}$	88.2	83.4	75.0	90.3	82.1	87.7	73.6	96.0	88.1	50.2	44.9	98.3	86.5
D+R+e_{1-5}	92.0	89.0	79.7	91.3	88.4	98.4	93.4	93.7	88.1	84.7	64.0	98.5	64.6
D+R+d_{1-5}	90.9	87.4	78.4	90.1	88.8	98.9	89.6	91.2	85.5	84.2	54.6	98.0	73.3
D+R+e_{1-5}+$d_{2,5}$	92.5	89.7	81.6	90.3	89.0	97.3	91.3	97.0	89,6	85.7	71.8	98.3	66.3
D+R+$e_{2,5}$+d_{1-5}	92.7	89.9	82.0	90.7	89.1	96.8	91.3	97.2	90.3	87.3	72.6	98.7	66.6
Ours	**93.1**	**90.4**	**82.7**	91.7	90.1	98.2	92.0	96.4	90.4	87.7	72.6	98.0	68.7

TABLE VIII
QUANTITATIVE COMPARISON RESULTS OF NETWORK EFFICIENCY. CONSIDERING THREE ASPECTS: COMPUTATION (GFLOP), PARAMETERS (PARAMS), AND LATENCY.

Method	GFLOPs ↓	Params/M ↓	Latency/s ↓
SAM	64.76	90.91	31.58
MCANet	88.63	72.03	19.31
ESANet	**11.52**	54.41	9.35
RedNet	21.21	81.94	12.11
ACNet	26.47	116.6	14.83
CMGFNet	38.83	85.22	9.25
Ours	14.34	**15.26**	**7.61**

tions for the sample imbalance problem. This could involve designing plug-and-play modules applicable to all models or refining existing datasets to achieve balanced distributions. Such efforts will help advance the effectiveness of remote sensing algorithms in handling diverse scenarios.

V. CONCLUSION

This paper presents EMSNet, an efficient segmentation model integrating high-resolution RGB and DSM images through dual-branch networks. To address accuracy loss due to the lightweight, we introduce RGB-guided continuous feature fusion, optimizing the outputs of the DSM branch. Concurrently, multi-level fusion effectively combines low-spatial and high-semantic information, yielding fine-grained segmentation maps. Moreover, we pioneer the application of a knowledge distillation strategy based on SAM, significantly enhancing EMSNet's performance and generalization. Experimental evaluations on the Potsdam and Vaihingen datasets demonstrate EMSNet's superior accuracy and efficiency compared to existing models. Furthermore, results from the WHU-OPT-SAR and DFC2023 datasets underscore the distillation strategy's potential to extend EMSNet's applicability to complex scenarios. In the future, we will focus on addressing the uneven distribution of categories and improving the performance of small sample sizes.

REFERENCES

[1] M. Dai, W. O. Ward, G. Meyers, D. D. Tingley, and M. Mayfield, "Residential building facade segmentation in the urban environment," *Building and Environment*, vol. 199, p. 107921, 2021.

[2] L. Mao, Z. Zheng, X. Meng, Y. Zhou, P. Zhao, Z. Yang, and Y. Long, "Large-scale automatic identification of urban vacant land using semantic segmentation of high-resolution remote sensing images," *Landscape and Urban Planning*, vol. 222, p. 104384, 2022.

[3] B. Wen, F. Peng, Q. Yang, T. Lu, B. Bai, S. Wu, and F. Xu, "Monitoring the green evolution of vernacular buildings based on deep learning and multi-temporal remote sensing images," in *Building Simulation*, vol. 16, no. 2. Springer, 2023, pp. 151–168.

[4] J. Yan, J. Liu, D. Liang, Y. Wang, J. Li, and L. Wang, "Semantic segmentation of land cover in urban areas by fusing multi-source satellite image time series," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[5] Y. Yang, G. Yuan, and J. Li, "Correlated mapping attention cooperative network for urban remote sensing image segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[6] S. Ren and Q. Liu, "Small target augmentation for urban remote sensing image real-time segmentation," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[8] N. He, L. Fang, and A. Plaza, "Hybrid first and second order attention unet for building segmentation in remote sensing images," *Science China Information Sciences*, vol. 63, pp. 1–12, 2020.

[9] A. Abdollahi, B. Pradhan, and A. M. Alamri, "An ensemble architecture of deep convolutional segnet and unet networks for building semantic segmentation from high-resolution aerial images," *Geocarto International*, vol. 37, no. 12, pp. 3355–3370, 2022.

[10] J. Yang, B. Matsushita, and H. Zhang, "Improving building rooftop segmentation accuracy through the optimization of unet basic elements and image foreground-background balance," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 201, pp. 123–137, 2023.

[11] H. AlMarzouqi and L. S. Saoud, "Semantic labeling of high resolution images using efficientunets and transformers," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[12] S. Zhou, Y. Feng, S. Li, D. Zheng, F. Fang, Y. Liu, and B. Wan, "Dsm-assisted unsupervised domain adaptive network for semantic segmentation of remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[13] X. Ma, X. Zhang, M.-O. Pun, and M. Liu, "A multilevel multimodal fusion transformer for remote sensing semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[14] X. Xiu, X. Ma, M.-O. Pun, and M. Liu, "Mdafnet: Monocular depth-assisted fusion networks for semantic segmentation of complex urban remote sensing data," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 6847–6850.

[15] J. Fan, J. Li, Z. Hua, F. Zhang, and C. Zhang, "Elevation information-guided multimodal fusion robust framework for remote sensing image segmentation," *IEEE Geoscience and Remote Sensing Letters*, 2024.

[16] J. Cui, J. Liu, Y. Ni, J. Wang, and M. Li, "Fdgsnet: A multi-modal gated segmentation network for remote sensing image based on frequency decomposition," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[17] B. Chu, J. Chen, J. Chen, X. Pei, W. Yang, F. Gao, and S. Wang, "Sdcafnet: A deep convolutional neural network for land-cover semantic segmentation with the fusion of polsar and optical images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 8928–8942, 2022.

[18] X. Li, G. Zhang, H. Cui, S. Hou, Y. Chen, Z. Li, H. Li, and H. Wang, "Progressive fusion learning: A multimodal joint segmentation framework for building extraction from optical and sar images," *ISPRS Journal of photogrammetry and remote sensing*, vol. 195, pp. 178–191, 2023.

[19] Y. Zhou, Y. Tan, Q. Wen, W. Wang, L. Li, and Z. Li, "Deep multimodal fusion model for building structural types recognition using multi-source remote sensing images and building-related knowledge," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[20] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu *et al.*, "Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery," *arXiv preprint arXiv:2312.10115*, 2023.

[21] W.-L. Du, Y. Gu, J. Zhao, H. Zhu, R. Yao, and Y. Zhou, "A mamba-diffusion framework for multimodal remote sensing image semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, 2024.

[22] K. Wei, J. Dai, D. Hong, and Y. Ye, "Mgfnet: An mlp-dominated gated fusion network for semantic segmentation of high-resolution multi-modal remote sensing images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 135, p. 104241, 2024.

[23] R. Liu, J. Ling, and H. Zhang, "Softformer: Sar-optical fusion transformer for urban land use and land cover classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 218, pp. 277–293, 2024.

[24] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[25] A. Kumar, T. Kashiyama, H. Maeda, H. Omata, and Y. Sekimoto, "Real-time citywide reconstruction of traffic flow from moving cameras on lightweight edge devices," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 192, pp. 115–129, 2022.

[26] H. Ijaz, R. Ahmad, R. Ahmed, W. Ahmad, Y. Kai, and W. Jun, "A uav assisted edge framework for real-time disaster management," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[27] C. Huynh, A. T. Tran, K. Luu, and M. Hoai, "Progressive semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 755–16 764.

[28] E. Lee, S. Jeong, J. Kim, and K. Sohn, "Semantic equalization learning for semi-supervised sar building segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[29] R. Zhang, J. Chen, L. Feng, S. Li, W. Yang, and D. Guo, "A refined pyramid scene parsing network for polarimetric sar image semantic segmentation in agricultural areas," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[30] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and cnn hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022.

[31] Z. Chen, D. Li, W. Fan, H. Guan, C. Wang, and J. Li, "Self-attention in reconstruction bias u-net for semantic segmentation of building rooftops in optical remote sensing images," *Remote sensing*, vol. 13, no. 13, p. 2524, 2021.

[32] R. Liu, L. Mi, and Z. Chen, "Afnet: Adaptive fusion network for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7871–7886, 2020.

[33] X. Ma, X. Zhang, M.-O. Pun, and M. Liu, "Msfnet: Multi-stage fusion network for semantic segmentation of fine-resolution remote sensing data," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 2833–2836.

[34] G. Iyer, J. Chanussot, and A. L. Bertozzi, "A graph-based approach for data fusion and segmentation of multimodal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4419–4429, 2020.

[35] T. Wang, G. Chen, X. Zhang, C. Liu, X. Tan, J. Wang, C. He, and W. Zhou, "Lmfnet: An efficient multimodal fusion approach for semantic segmentation in high-resolution remote sensing," *arXiv preprint arXiv:2404.13659*, 2024.

[36] W. Wu, S. Guo, Z. Shao, and D. Li, "Crofusenet: A semantic segmentation network for urban impervious surface extraction based on cross fusion of optical and sar images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 2573–2588, 2023.

[37] S. Xiao, P. Wang, W. Diao, X. Rong, X. Li, K. Fu, and X. Sun, "Mocg: Modality characteristics-guided semantic segmentation in multimodal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 2023.

[38] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (exvit) for land use and land cover classification: A multimodal deep learning framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[39] D. Hong, B. Zhang, H. Li, Y. Li, J. Yao, C. Li, M. Werner, J. Chanussot, A. Zipf, and X. X. Zhu, "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sensing of Environment*, vol. 299, p. 113856, 2023.

[40] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 561–577.

[41] K. Yue, J. Deng, and F. Zhou, "Matching guided distillation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 312–328.

[42] J. Jiang, L. Zheng, F. Luo, and Z. Zhang, "Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation," *arXiv preprint arXiv:1806.01054*, 2018.

[43] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1440–1444.

[44] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Information Fusion*, vol. 55, pp. 1–15, 2020.

[45] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Gross, "Efficient rgb-d semantic segmentation for indoor scene analysis," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 13 525–13 531.

[46] H. Hosseinpour, F. Samadzadegan, and F. D. Javan, "Cmgfnet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images," *ISPRS journal of photogrammetry and remote sensing*, vol. 184, pp. 96–115, 2022.

[47] X. Li, G. Zhang, H. Cui, S. Hou, S. Wang, X. Li, Y. Chen, Z. Li, and L. Zhang, "Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 106, p. 102638, 2022.

[48] X. Fan, W. Zhou, X. Qian, and W. Yan, "Progressive adjacent-layer coordination symmetric cascade network for semantic segmentation of multimodal remote sensing images," *Expert Systems with Applications*, vol. 238, p. 121999, 2024.

Yejian Zhou was born in Zhejiang Province, China, in 1993. He received the B.S. degree in electronic engineering and the Ph.D. degree in signal processing from Xidian University, in 2015 and 2020, respectively. He was a Visiting Ph.D. student with the Department of Urban Planning and Environment, KTH Royal Institute of Technology, from September 2019 to August 2020. He is currently an Associate Professor with College of Information Engineering, Zhejiang University of Technology. His research interests include ISAR imaging and image interpretation.

Wang Yachen was born in Shandong in 2000, received a bachelor's degree in engineering from Wuhan University of Science and Technology in 2022 and is currently pursuing a master's degree at Zhejiang University of Technology. His researched interests include semantic segmentation and height estimation etc.

Jie Su is currently an assistant professor with the Institute of Cyberspace Security and College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His research interests include deep learning, signal processing, and the IoT security. Su received his Ph.D. degree in computer science from Newcastle University U.K., in 2023.

Zhenyu Wen (Senior Member, IEEE) is currently the Tenure-Tracked Professor with the Institute of Cyberspace Security, and College of Information Engineering, Zhejiang University of Technology. His current research interests include IoT, crowd sources, AI systems, and cloud computing. For his contributions to the area of scalable data management for the Internet of Things, he was awarded the IEEE TCSC Award for Excellence in Scalable Computing (Early Career Researchers) in 2020.

Puzhao Zhang received a B.S. degree in intelligent science and technology from Xidian University in 2013 and his PhD in Geoinformatics from KTH Royal Institute of Technology in 2021. His research interests include machine learning, data mining, computer vision, and spatial-temporal analysis with their applications in optical and radar remote sensing, including change detection, urban expanding, land-cover classification, and wildfire monitoring.

Wen-An Zhang was born in Zhejiang Province, China, in 1982. He received the B.S. degree in automation and the Ph.D. degree in control theory and control engineering from the Zhejiang University of Technology in 2004 and 2010, respectively. Since 2020, he has been with the Zhejiang University of Technology, where he is currently a Professor with the Department of Automation. He was a Senior Research Associate with the Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong, from 2010 to 2011. His current research interests include multi-sensor information fusion estimation, and robotics. He has been a Subject Editor for Optimal Control Applications and Methods since September 2016. He was awarded an Alexander von Humboldt Fellowship in 2011–2012.