Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.060860

ARTICLE

CG-FCLNet: Category-Guided Feature Collaborative Learning Network for Semantic Segmentation of Remote Sensing Images

Min Yao[1,*], **Guangjie Hu**[1] **and Yaozu Zhang**[2]

[1]College of Information Engineering, Shanghai Maritime University, Shanghai, 201306, China
[2]R&D Department, Shanghai Freesense Technology Co. Ltd., Shanghai, 200000, China

*Corresponding Author: Min Yao. Email: yaomin1987@hotmail.com

Received: 12 November 2024; Accepted: 24 February 2025; Published: 16 April 2025

ABSTRACT: Semantic segmentation of remote sensing images is a critical research area in the field of remote sensing. Despite the success of Convolutional Neural Networks (CNNs), they often fail to capture inter-layer feature relationships and fully leverage contextual information, leading to the loss of important details. Additionally, due to significant intra-class variation and small inter-class differences in remote sensing images, CNNs may experience class confusion. To address these issues, we propose a novel Category-Guided Feature Collaborative Learning Network (CG-FCLNet), which enables fine-grained feature extraction and adaptive fusion. Specifically, we design a Feature Collaborative Learning Module (FCLM) to facilitate the tight interaction of multi-scale features. We also introduce a Scale-Aware Fusion Module (SAFM), which iteratively fuses features from different layers using a spatial attention mechanism, enabling deeper feature fusion. Furthermore, we design a Category-Guided Module (CGM) to extract category-aware information that guides feature fusion, ensuring that the fused features more accurately reflect the semantic information of each category, thereby improving detailed segmentation. The experimental results show that CG-FCLNet achieves a Mean Intersection over Union (mIoU) of 83.46%, an mF1 of 90.87%, and an Overall Accuracy (OA) of 91.34% on the Vaihingen dataset. On the Potsdam dataset, it achieves a mIoU of 86.54%, an mF1 of 92.65%, and an OA of 91.29%. These results highlight the superior performance of CG-FCLNet compared to existing state-of-the-art methods.

KEYWORDS: Semantic segmentation; remote sensing; feature context interaction; attention module; category-guided module

1 Introduction

Semantic segmentation of remote sensing images [1] is a critical task in the field of computer vision. Unlike traditional segmentation, it focuses on both geometric shapes and semantic information for each pixel. Each pixel must be categorized (e.g., human, vehicle, building) for deeper content understanding. It's widely used in fields like land cover mapping [2], urban planning [3], and road and building extraction [4,5]. However, due to the complex scenes, small target sizes, and uneven category distribution typical of remote-sensing images, achieving highly accurate semantic segmentation remains an extremely challenging task.

In recent years, Convolutional Neural Networks (CNNs) have demonstrated outstanding performance in many fields and have become the mainstream method for semantic segmentation. Fully Convolutional Networks (FCNs) [6] enable networks to process inputs of any size and perform pixel-level segmentation by removing fully connected layers from traditional CNNs. However, their simple structure often leads to coarse segmentation results, lacking fine details. To address this issue, researchers have proposed various

improved models. For example, U-shaped Convolutional Neural Network (U-Net) [7] builds upon FCNs by introducing skip connections, which effectively recover lost detail information and improve segmentation accuracy. However, skip connections fail to adequately distinguish and fuse low-level and high-level features, limiting the network's ability to represent features. To further enhance the integration of multi-scale contextual information, DeepLabv3+ [8] employs dilated convolutions, overcoming limitations in capturing multi-scale features. Subsequently, Pyramid Scene Parsing Network (PSPNet) [9] addresses the shortcomings in aggregating cross-region contextual information through a pyramid pooling module, further improving the model's semantic understanding. To strengthen feature representation, Multi-Level Feature Mining Network (MLFMNet) [10] introduces the Irregular Pyramid Receptive Field (IPRF) module, overcoming the limitations of traditional pyramid structures when processing detailed features in complex scenes. Additionally, Dual-Branch Hybrid Reinforcement Network (DHRNet) [11] adopts a dual-branch parallel structure, improving both global semantic and local detail extraction, while reducing redundancy and optimizing computational efficiency and memory usage. In contrast, Hidden Feature-Guided Network (HFGNet) [12] uses a multi-feature fusion module to enhance object discrimination and address this issue, although small-scale feature loss or inaccuracies may still occur. Multiscale Global Context Network (MSGCNet) [13] employs cross-attention to narrow the semantic gap and enrich information, achieving better segmentation. However, unclear edge features in occluded areas hinder feature interaction. Multiscale Feature Context Aggregation Network (MFCANet) [14] combines features across scales using a multi-scale feature context aggregation module, excelling in cross-scale information handling, but faces difficulties in detecting densely occluded small objects. Multi-scale Context-aware and Global Feature Fusion Network (MCGFF-Net) [15] introduces a Cross-Layer Feature Fusion Module (CFM), enhancing semantic information interaction, but still requires improvement in handling noise and redundancy. Global Extraction Local Exploration Network (GeletNet) [16] enhances object recognition by strengthening mid-level feature interactions through the Knowledge Transfer Module (KTM), but its performance is limited due to the lack of interactions between other layers. Context Exploration and Multi-level Interaction Network (CEMINet) [17] aggregates multi-level features through its Hierarchical Feature Hybrid Interaction (HFHI) module, refining the feature structure top-down, but still struggles with long-range dependencies.

Although existing models have made progress in segmentation tasks, they have not adequately explored the interaction between feature maps at different stages. In remote sensing images, where object distribution is often sparse, relying solely on local contextual information for pixel classification can blur category boundaries and result in incomplete segmentation. Additionally, when local features of different object types are highly similar, this similarity causes mutual interference, weakening the model's discriminative ability, and resulting in feature entanglement, unclear boundaries, and loss of structural details.

To address these issues, we propose a novel Category-Guided Feature Collaborative Learning Network (CG-FCLNet). This network enhances the interaction between multi-stage features and utilizes category-aware information to guide feature fusion during the decoding process. CG-FCLNet integrates feature information from different stages by employing an advanced Feature Collaborative Learning Module (FCLM), which reveals the inherent relationships and complementary advantages between them. Building on this, we introduce the Scale-Aware Fusion Module (SAFM), designed to mine multi-scale global contextual information outputted by FCLM. This module iteratively integrates features through a spatial attention mechanism, achieving fine fusion and optimization. Additionally, we introduce the Category-Guided Module (CGM), which uses coarse segmentation results to guide the fusion of adjacent feature maps, leading to more refined segmentation outcomes. The main contributions of our work are as follows:

(1) We propose the Feature Collaborative Learning Module (FCLM), which enhances the interaction between features at different stages and explores their intrinsic relationships, effectively improving the model's discriminative ability.

(2) We propose the Scale-Aware Fusion Module (SAFM), which uses spatial attention to iteratively integrate the global context of the decoder to enhance the discriminability of the final prediction.

(3) We propose the Category-Guided Module (CGM), which guides category information at each layer during the feature fusion process to optimize feature representation and integration.

The rest of this paper is organized as follows: Section 2 discusses related work; Section 3 describes the details of our proposed method; Section 4 presents the experimental setup, ablation studies, and comparative results, along with a detailed discussion; Section 5 elaborates on the conclusion. Abbreviations provide a comprehensive list of abbreviations and terms used throughout the paper.

2 Related Works

In this section, we review some of the research related to the proposed model, focusing on multi-scale feature Context extraction and category-aware mechanisms. As shown in Table 1, the main methods from these studies are presented.

2.1 Multi-Scale Feature Context Interaction

In semantic segmentation, multi-scale feature interaction provides rich information that enables models to better understand image semantics and achieve high-quality segmentation. For example, GeletNet [16] introduces the KTM, which extracts discriminative information for salient objects by emphasizing the interaction between mid-level features. However, it does not fully leverage the effectiveness of interactions between adjacent feature layers. CEMINet [17] facilitates interaction between high-level and low-level features in a top-down manner, but it struggles to effectively capture long-range dependencies, limiting its ability to extract more robust semantic information. CNN and Multiscale Local-Context Transformer Network (CMLFormer) [18] introduces a multi-scale local-context transform block (MLTB), which efficiently captures both local and global feature information at various scales with low complexity. Geometric Prior-guided Interactive Network (GPINet) [19] proposes local-global interaction modules (LGIMs) that allow the network to refine learned representations interactively and efficiently. However, the hybrid network structure may increase model complexity and require additional computational resources. Additionally, Bhatti et al. [20] present a detailed feature extraction approach that combines each feature with its cross-information, enhancing the model's ability to recognize complex objects and improving multi-feature fusion robustness.

Compared to these methods, we explore the intrinsic relationships between adjacent feature maps and strengthen their connections through mutual learning. Meanwhile, we focus on the distinct characteristics of each feature map, optimizing and fine-tuning them accordingly. This series of operations is primarily facilitated by the Feature Collaborative Learning Module (FCLM), which ensures that the feature maps can collaborate effectively while retaining their characteristics.

2.2 Category-Aware Mechanism

Studies [21–24] demonstrate that category-aware mechanisms enhance segmentation accuracy by integrating information across categories, enabling adaptive feature highlighting. For example, the Class Guidance Block module [25] enhances the model's ability to handle fine-grained category differences by strengthening feature fusion and refinement. However, class imbalance remains a challenge, particularly

when category distributions are uneven. Semantic Category Balance-Aware Involved Anti-Interference Network (SCBANet) [26] introduces an Optional Decoder Module based on Semantic Category Balance (ODMSCB), which dynamically adjusts the decoder's output to address this issue. Category attention guided network (CAGNet) [27] uses the Category Attention Guided Module (CAGM) to capture key category differences. Similarly, Category-Based Interactive Attention and Perception Fusion Network (CIAPNet) [28] refines self-attention through the Category Grouped Attention (CGA) module to focus on distinct category features. However, they still have limitations in dynamically adjusting category weights. To address this issue, Uncertain Category-Aware Fusion Network (UCAFNet) [29] proposes an Uncertain Category-Aware Fusion Strategy (UCAFS), which dynamically adjusts category weights during fusion for better performance in complex scenarios. However, the lack of coordination between these modules may limit the effectiveness of information integration.

In contrast, we consider semantic differences between features at different stages. Instead of simply upsampling and adding them for fusion, we use a Scale-Aware Fusion Module (SAFM) to capture the importance of different image regions, achieving precise feature fusion. Furthermore, we introduce a Category-Guided Module (CGM), which guides adjacent feature map fusion using coarse segmentation results of multi-layer features, ensuring effective category information use and enhancing segmentation accuracy.

2.3 Self-Attention Mechanism

CNNs are proficient at extracting local features and achieving scale invariance, but they are limited in modeling global information. To address this, researchers have introduced self-attention mechanisms [30–32] to capture long-range dependencies and improve segmentation performance. For example, Attention Aggregation Feature Pyramid Network (A^2-FPN) [33] introduced a feature pyramid network with linear attention weighting to enhance the fusion of cross-scale features and improve segmentation accuracy. However, this method does not effectively mitigate the impact of low-resolution feature maps on segmentation results. To address this, Multistage Attention ResU-Net (MAResU-Net) [34] combines attention mechanisms with U-Net by weighting low-resolution feature maps to enhance segmentation performance. Despite these improvements, it still faces challenges in handling fine-grained features and multi-scale information. Hybrid Multiple Attention Network (HMANet) [35] integrates multiple attention mechanisms to strengthen class information and learn pixel-wise spatial correlations, effectively extracting detailed features. Additionally, Global-Local Self-Attention Network (GLSANet) [36] proposes a global-local self-attention mechanism that combines global semantic information with local details to enhance model robustness and accuracy in complex backgrounds. Based on these insights, we propose a new method that enhances multi-stage feature interactions and incorporates category-aware information to improve segmentation in complex scenes.

Table 1: Main methods in related studies

Category	Advantages	Disadvantages	Reference
Multi-scale feature context interaction	Capture local and global information, improve detail sensitivity, and address blurred boundaries and detail loss.	The computational complexity and memory consumption increase significantly, while information redundancy or inconsistency may occur.	[16–19]

(Continued)

Table 1 (continued)

Category	Advantages	Disadvantages	Reference
Category-aware mechanism	Enhance the model's ability to distinguish between categories and improve category differentiation.	Limited by the quality and diversity of category label data, inaccurate labels or imbalanced category distribution may affect the model's performance.	[26–29]
Self-attention mechanism	Focus on relevant features within the image and capture long-range dependencies.	Lacking explicit modeling of spatial structure, it may also fail to effectively distinguish redundant information.	[33–36]

3 Methodology

Fig. 1 illustrates the overall framework of CG-FCLNet, which consists of four main modules. First, an encoder based on convolutional neural networks extracts multi-scale semantic features from the image. Second, an innovative Feature Collaborative Learning Module (FCLM) is introduced. This module enhances key features in the feature map through deep interaction, while suppressing irrelevant information. It ensures that the network focuses on the most important parts of the image. Next, the Scale-Aware Fusion Module (SAFM) integrates features from different levels, resulting in a more comprehensive and enriched feature representation. Meanwhile, the Category-Guided Module (CGM) uses category information to guide feature fusion, leading to more accurate segmentation results.

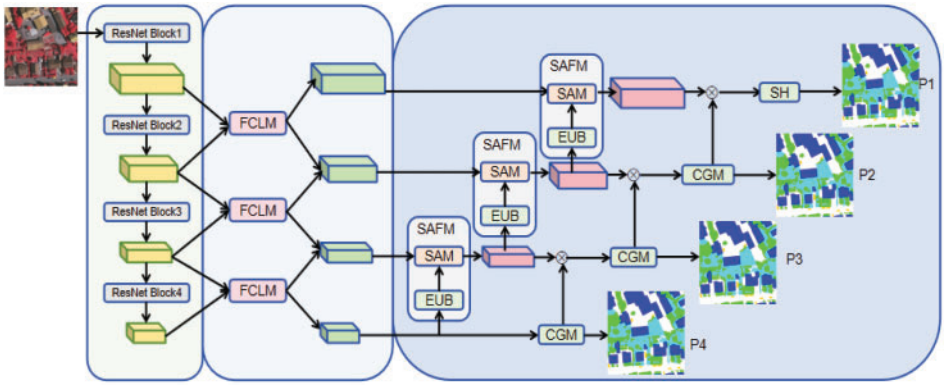

Figure 1: The architecture of our proposed Category-Guided Feature Collaborative Learning Network (CG-FCLNet). It utilizes the backbone to extract features, which are then enhanced by the FCLM. The category information extracted by the CGM guides the SAFM in integrating feature information

3.1 Feature Collaborative Learning Module

Cross-scale feature interaction significantly enhances the model's ability to understand both details and the global context by integrating features of different resolutions. This approach bridges the gap in correlation between features at different network layers, enabling effective cross-layer feature transfer. Based on this, we propose a novel Feature Collaborative Learning Module (FCLM), as illustrated in Fig. 2. The module is primarily composed of Fig. 2b and c, with the overall architecture shown in Fig. 2a. This module

explores the latent relationships between features at different stages and performs targeted integration based on the strengths and characteristics of each stage, capturing finer semantic information and extracting more discriminative features.

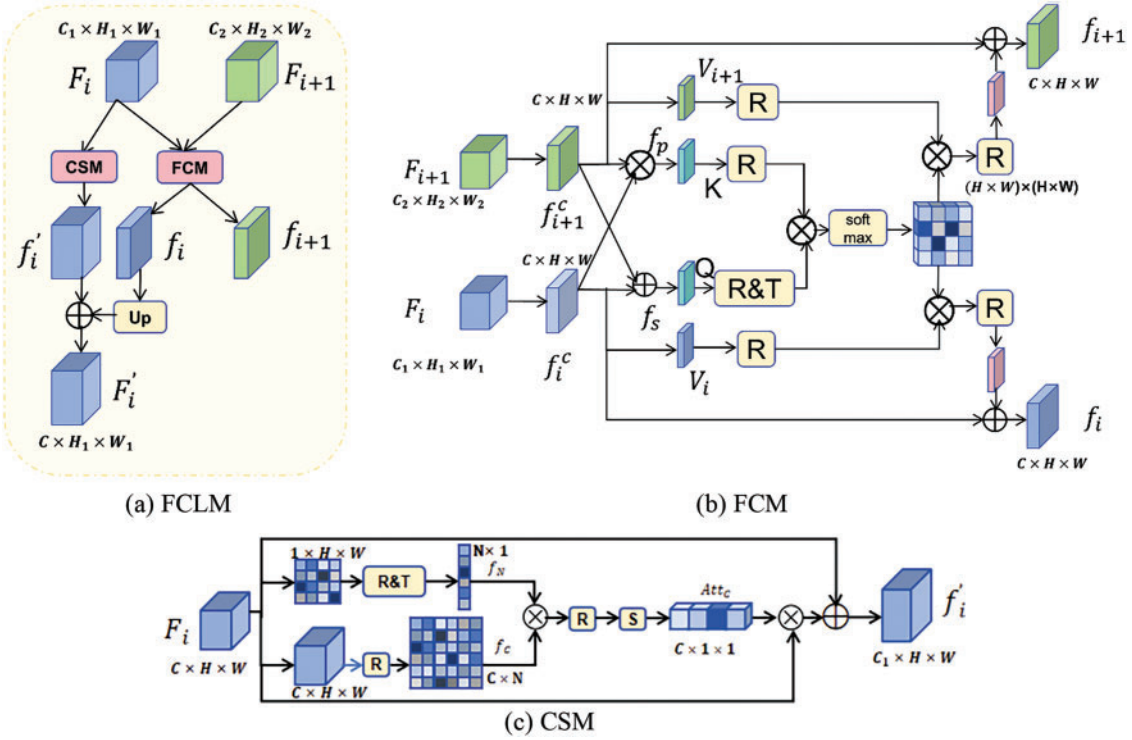

Figure 2: Structure of the feature collaborative learning module (FCLM). (a) The implementation details of FCLM, consist of (b) and (c). (b) Feature Collaborative Module (FCM) introduces multi-scale information to the network. (c) Channel Selection Module (CSM) selects the most suitable feature maps and eliminates redundant information, where R, T, and S represent Reshape, Transpose, and Sigmoid, respectively

Specifically, the FCLM shown in Fig. 2a consists of two parts: a Feature Collaborative Module (FCM) and a Channel Selection Module (CSM). The FCM models the contextual correlation of features across different stages. It efficiently integrates feature information from various stages, delving into the potential connections and complementary strengths. Through effective fusion and processing, FCM can generate discriminative features. The CSM is designed to process each stage's feature maps based on their unique characteristics. It selects the most suitable feature map channels from a multitude of options, eliminating redundant information to prevent interference with subsequent processing. This optimizes the feature representation. Finally, the feature maps produced by both the FCM and CSM are summed to obtain the final feature map. This approach achieves effective integration of cross-stage feature information and generates discriminative features. It also ensures that each stage's feature maps are refined and optimized with precision. Next, we will provide the design principles and detailed explanations for these two modules.

3.1.1 Feature Collaborative Module

Feature Collaborative Module (FCM): When two feature maps are multiplied, the result emphasizes areas where both exhibit significant responses. This highlights the salient information shared between the features and guides the model to focus on these regions, facilitating the collaborative identification of objects. In contrast, adding two feature maps combines the information each contains. Therefore, in the FCM, we first

standardize the feature maps from different scales. Then, we apply addition and multiplication operations to generate new features. Finally, we use self-attention mechanisms to model the cross-stage feature context interaction. This strategy enables the model to capture fine-grained features more effectively, making the resulting features more discriminative and thereby improving the model's ability to recognize objects in complex scenes and diverse contexts.

Fig. 2b illustrates the detailed design of the FCM. It starts with the cross-stage feature maps F_i and F_{i+1} as inputs (where F_i represents the feature map generated at the i-th stage). Due to the scale differences between feature maps from various stages, a 3 × 3 convolution is first applied to standardize the scale and the number of channels, resulting in f_i^c and f_{i+1}^c. By performing product and sum operations on f_i^c and f_{i+1}^c to highlight their correlation. This approach also considers all the information from these two features, enhancing their expressiveness and adaptability. The obtained feature maps f_p and f_s are input into a convolutional layer to generate the corresponding query matrix Q and key matrix K, where $\{Q, K\} \in R^{C\times H\times W}$. They are then reshaped into $R^{C\times N}$, where $N = H \times W$ represents the number of pixels. A matrix multiplication operation is performed between the transpose of Q and K, and a softmax layer is used to calculate the weights $S \in R^{N\times N}$. This process can be formulated as:

$$S_{ij} = \frac{exp(Q_i \cdot K_j)}{\sum_{i=1}^{N}(exp(Q_i \cdot K_j))} \tag{1}$$

where S_{ij} measures the i-th position's impact on the j-th position.

Meanwhile, we input f_i^c and f_{i+1}^c into a convolutional layer to generate the corresponding value matrices V_i and V_{i+1}, which are then reshaped into $R^{C\times N}$. Finally, V_i and V_{i+1} are each multiplied with the transpose of S, performing matrix multiplication operations. The resulting two matrices, which are of shape $R^{C\times N}$, are reshaped back into $R^{C\times H\times W}$ and the corresponding f_i^c and f_{i+1}^c are added together to obtain the final output.

3.1.2 Channel Selection Module

Channel Selection Module (CSM): While the FCM primarily focuses on feature fusion and correlation, the CSM is introduced as a supplementary component to better capture the inherent information of the features themselves. The CSM directly calculates channel attention scores from the feature map and optimizes them by training the mapping matrix. This approach integrates a channel attention mechanism, enabling the model to focus more accurately on important feature information. It can capture key channel information and retain a certain proportion of the original feature data, thereby enhancing the effectiveness of feature selection.

The specific implementation of CSM is shown in Fig. 2c. A 1 × 1 convolutional layer is applied, followed by reshaping and transposing operations on F_i to obtain a matrix $f_N \in R^{N\times 1}$. Meanwhile, a matrix $f_c \in R^{C\times N}$ is obtained by applying a 1 × 1 convolutional layer and reshaping the result. These two matrices are then multiplied and reshaped, and the sigmoid function is applied to obtain the channel attention map Att_c. The process Att_c can be described as:

$$Att_c = \delta\left(Conv_{1\times 1}\left(f_c \times f_N\right)\right) \tag{2}$$

where δ represents the Sigmoid function. The feature map F_i is multiplied by the channel attention map Att_c. A residual connection is then performed between the result and F_i to obtain the final feature map $f_i^{'}$.

3.2 Scale-Aware Fusion Module

In the decoder, directly combining features from different stages may reduce feature expressiveness due to semantic differences and information redundancy [37]. Inspired by relevant studies [38,39], we designed a Scale-Aware Fusion Module (SAFM). As shown in SAFM in Fig. 1, to address the scale and resolution differences of feature maps at various stages. The components of this module are illustrated in Fig. 3, we first apply an Efficient Up-convolution Block (EUB) to gradually upsample the feature map at the current stage, ensuring it matches the size and resolution of the feature map at the next stage. This approach preserves image details without significantly increasing the computational burden. Next, we introduce the Scale-Aware Model (SAM), which performs fine-grained fusion of features from different stages. This fusion strategy leverages the potential of features at each stage and dynamically selects the most valuable feature information for integration. This provides a more comprehensive and accurate feature representation for subsequent segmentation tasks. The specific operations are as follows.

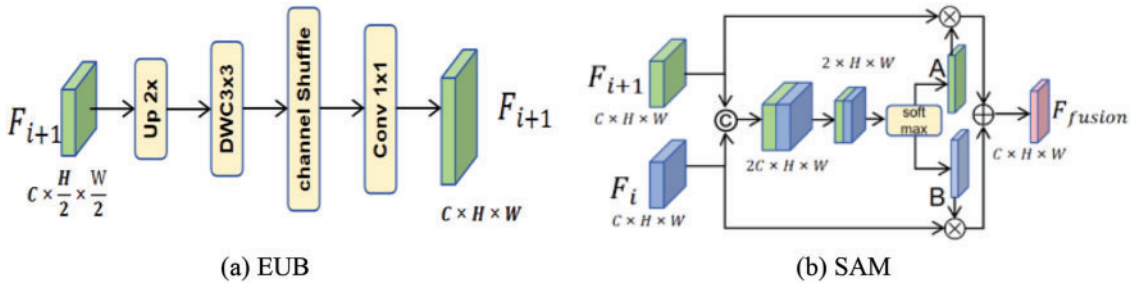

Figure 3: Structure of the scale-aware fusion module (SAFM). The SAFM comprises two main components: (a) Efficient Up-convolution Block (EUB) is used to restore the feature scale. (b) Scale-Aware Model (SAM) integrates feature information

Firstly, we use the EUB to unify the scales of the feature maps. As shown in Fig. 3a, the size of the feature map is expanded through an upsampling method with a ratio of 2, which increases its spatial dimension. Then, the upsampled feature map is enhanced using a 3×3 depth-wise convolution to explore additional feature information. In addition, we adopt the channel shuffle technique [40] to promote communication between different channels, enabling better fusion and cooperation of features across channels. Finally, a 1×1 convolution layer integrates the feature information obtained from the previous operations.

The high-level feature map with adjusted scales and the adjacent low-level feature map are sent to the SAM for feature fusion. The operations of the SAM module are shown in Fig. 3b. Firstly, the two feature maps are concatenated to preliminarily integrate the information of different feature maps. Next, a convolution layer is used to perform dimension reduction on the concatenated feature map. This step not only decreases computational complexity but also refines the feature representation. Subsequently, a softmax layer to obtain two spatial attention maps A and $B \in R^{H \times W}$. These weights are dynamically adjusted according to the importance of the feature maps, enabling the model to focus on the key information and improving the attention and utilization efficiency of important features. Finally, the attention weights are multiplied by the original feature maps and summed to obtain the fused features. This process ensures the efficiency and accuracy of the model in dealing with multi-scale features. The process can be described as:

$$F_{fusion} = A \times F_{i+1} + B \times F_i \tag{3}$$

3.3 Category-Guided Module

To address the challenge of large intra-class variations and small inter-class differences in high-resolution remote sensing imagery, we propose a Category-Guided Module (CGM). The module first utilizes deep-level features for category prediction, generating a preliminary segmentation map. Next, channel separation is applied to the rough segmentation results, which are then processed using the Sigmoid function. This step assigns a probability value to each pixel, indicating its likelihood of belonging to a specific category, thereby enhancing the precision of category information at the pixel level. Additionally, the output from the CGM is used to guide the Scale-Aware Fusion Module (SAFM) process layer by layer. This approach enables the model to capture semantic information at different levels, ensuring that the feature fusion process emphasizes the semantic relationships between categories. Consequently, the model can accurately identify objects and delineate their boundaries, leading to more precise segmentation results. This enhances the model's performance and accuracy, especially for complex remote-sensing images.

Specifically, as shown in Fig. 4, a coarse-grained segmentation result is obtained through a 3 × 3 and a 1 × 1 convolutional layer, denoted as $P \in R^{B\times K\times H\times W}$ (where K represents the total number of segmentation categories). A channel-separation operation is then applied to this rough segmentation result P to obtain P_i, where $P_i \in R^{B\times 1\times H\times W}$, and $i = 1, ..., K$. Next, a Sigmoid operation is performed on each P_i, so that the probability value of each pixel in P_i represents the likelihood that the pixel belongs to the corresponding category.

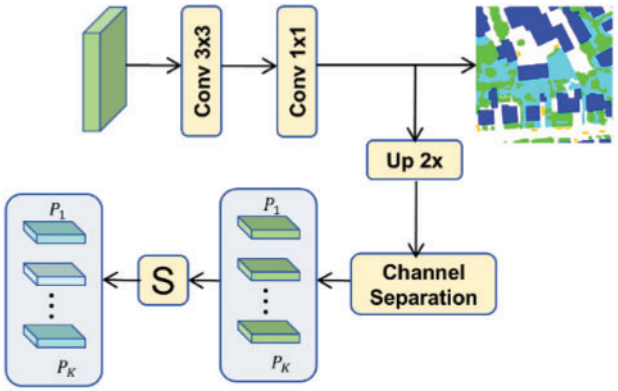

Figure 4: Structure of the category-guided module (CGM). *S* represents Sigmoid operation

The fusion features obtained in SAFM are multiplied by P_i to generate enhanced features for each category. Finally, the refined segmentation result is obtained through the segmentation head (containing 3 × 3 and 1 × 1 convolutional layers).

3.4 Loss Function

Like most prior works, we employ the cross-entropy loss L_{ce} and Dice loss L_{dice} as the total loss function. Specifically, the expressions for L_{ce} and L_{dice} are as follows, respectively:

$$L_{ce} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_k^n(\log \hat{y}_k^n) \tag{4}$$

$$L_{dice} = 1 - \frac{2|y \cap \hat{y}|}{|y| + |\hat{y}|} \tag{5}$$

$$L = L_{ce} + L_{dice} \tag{6}$$

where $\hat{y}$ denotes the network prediction segmentation result and y denotes the true value. N and K denote the number of samples and the number of categories, respectively.

Our decoder's four segmentation heads produce four prediction maps $P1$, $P2$, $P3$, and $P4$ across its stages. Thus, the total loss can be expressed as:

$$L_{total} = L_{p1} + L_{p2} + L_{p3} + L_{p4} \tag{7}$$

where L_{p1}, L_{p2}, L_{p3} and L_{p4} are the losses of each prediction map.

Overall, through the integrated use of FCLM, SAFM, and CGM, our approach is capable of sensitively detecting the distinguishing features among various objects and leveraging category information to guide the segmentation process. Consequently, this model exhibits superior performance in detail processing and edge positioning compared to alternative methods. This capability is particularly crucial in practical applications, especially in critical areas such as land use analysis, environmental change detection, and urban planning.

4 Experiment

In this section, we first provide a detailed introduction to the dataset used. Next, we discuss the specific parameter settings for model training and the quantitative assessment indicators of the network architecture. Additionally, we compare our proposed method with other advanced semantic segmentation networks and conduct ablation studies to further analyze and verify our approach. Through these steps, we comprehensively present the performance and advantages of our method from various perspectives.

4.1 Datasets

ISPRS Vaihingen dataset consists of 33 high-resolution images with an average size of 2494 × 2064 pixels per image. It contains five foreground classes (impervious surface, building, low vegetation, tree, and car) and one background class (clutter). According to the official dataset segmentation settings, we used image ID 1, 3, 5, 7, 11, 13, 15, 17, 21, 23, 26, 28, 32, 34, and 37 for training, image ID 30 for validation, and the remaining 17 images for testing.

ISPRS Potsdam dataset contains 38 high-resolution images, each with a size of 6000 × 6000 pixels. Each image contains the same category information as the Vaihingen dataset. According to the formal dataset segmentation setting, we used image ID 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 and 7_12 for training, image ID 2_10 for validation, and the remaining 14 images for testing (excluding the incorrectly labeled 7_10 image).

4.2 Implementation Details

All models were implemented using PyTorch on the NVIDIA GTX 3090 GPU in our lab device. In our experiments, we used the AdamW optimizer with an initial learning rate of 6×10^{-4} and a batch size of 8. The images were randomly cropped into 512 × 512 image patches. Augmentation techniques, such as random scaling ([0.5,0.75,1.0,1.25,1.5]), random vertical flips, and random horizontal flips, were used during the training process. The maximum number of epochs was set to 105 for both the Potsdam and Vaihingen datasets. Additionally, Test-Time Augmentation (TTA) techniques, including horizontal and vertical flips, were applied during the testing phase. This study uses Overall Accuracy (OA), Mean Intersection over Union

(mIoU), and Mean F1 Score (mF1) to evaluate the model's performance. The specific formulas are as follows:

$$OA = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} TP_k + FP_k + TN_k + FN_k} \tag{8}$$

$$mIoU = \frac{1}{K} \sum_{k=1}^{K} \frac{TP_k}{TP_k + FP_k + FN_k} \tag{9}$$

$$precision = \frac{1}{K} \sum_{k=1}^{K} \frac{TP_k}{TP_k + FP_k} \tag{10}$$

$$recall = \frac{1}{K} \sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k} \tag{11}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{12}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively, for a particular object indexed as category k. Precision refers to the proportion of actual positive instances among those that the model predicts as positive. Recall refers to the proportion of positive instances that the model predicts correctly out of all actual positive instances.

4.3 Ablation Study

We performed several types of ablation studies on the Vaihingen and Potsdam datasets with the aim of evaluating the contribution of the CG-FCLNet components. To ensure the fairness of the experiments, the test-time augmentation strategy was not adopted during all ablation experiments. We carried out ablation experiments according to the different settings listed in Table 2. The highest values in each column are in bold.

Table 2: Results of the ablation study on two datasets, the highest values in each column are in bold

Model	Vaihingen			Potsdam		
	mF1	mIoU	OA	mF1	mIoU	OA
Baseline	88.26	79.27	89.59	90.33	82.58	88.88
Baseline+CSM	89.02	80.47	90.12	91.21	84.06	89.89
Baseline+FCM	89.33	80.97	90.16	91.39	84.38	89.97
Baseline+FCLM	89.63	81.43	90.33	91.59	84.71	90.20
Baseline+CGM	89.13	80.65	90.12	91.26	84.16	89.71
Baseline+SAFM	89.50	81.25	90.17	91.39	84.36	89.78
Baseline+FCLM+SAFM	89.83	81.78	90.45	91.85	85.16	90.47
Baseline+FCLM+CGM	89.75	81.65	90.54	91.78	85.05	90.41
Baseline+FCLM+SAFM+CGM	**90.12**	**82.25**	**90.72**	**92.10**	**85.62**	**90.66**

In this research, we adopted a baseline structure similar to U-Net, incorporating a pre-trained ResNet18 encoder. In the decoder, we used bilinear upsampling and addition operations to fuse features, effectively

transferring them through skip connections. Based on this, we separately tested different modules or combinations of them to verify the effectiveness of each.

Superiority of the FCLM: The FCLM consists of FCM and CSM. To verify the effectiveness of these two components, we compared the FCLM with FCM and CSM. As shown in Table 2, after introducing FCM, the mF1, OA, and mIoU of the network on the Vaihingen dataset are increased by 1.07%, 0.57%, and 1.70%, which verifies the effectiveness of FCM. After introducing CSM, the mF1, OA, and mIoU of the network on the Vaihingen dataset are increased by 0.76%, 0.53%, and 1.20% respectively, which verifies the effectiveness of CSM. The FCLM composed of the combination of the two increases the mF1, OA, and mIoU on the Vaihingen dataset by 1.37%, 0.74%, and 2.16%, respectively, which confirms the complementarity of FCM and CSM in features. Similar trends can be observed on the Potsdam dataset as well, where mF1, OA, and mIoU increased by 1.26%, 1.32%, and 2.13%, respectively. This complementarity shows that the FCLM can effectively generate richer feature representations through cross-stage feature interaction, promoting efficient information exchange and integration between features at different levels. As a result, the model can capture more critical information and feature variations. The FCM improves feature discriminability and semantic detail, boosting performance but increasing computational overhead and risking information loss. The CSM optimizes feature representation by removing redundancies, and reducing computational costs, though it may discard valuable features. The FCLM integrates FCM and CSM, balancing discriminability, efficiency, and representation.

Superiority of the CGM and FCLM+CGM: To verify the effectiveness of CGM, it was integrated alone into the Baseline model, creating Baseline + CGM. As shown in Table 2, the Baseline model with CGM achieved an mF1 score of 0.87, an OA of 0.53, and a mIoU of 1.38% on the Vaihingen dataset, while the corresponding improvements on the Potsdam dataset were 0.93%, 0.83%, and 1.58%, respectively. These results demonstrate the effectiveness of CGM in guiding feature fusion by progressively incorporating category-specific information. CGM ensures that category information is utilized at each stage of feature fusion, enhancing the model's ability to identify and integrate features from different categories in the image. However, while CGM improves the model's recognition capability, it has limitations in capturing multi-scale contextual information. To address this, the introduction of FCLM further enhances the network's performance metrics. This indicates that the combination of FCLM and CGM complements each other, improving the model's ability to integrate multi-scale information.

The superiority of the SAFM and FCLM+SAFM: After integrating SAFM with the Baseline model, the mF1 score, OA, and mIoU on the Vaihingen dataset showed improvements of at least 1.24%, 0.58%, and 1.98%, respectively. On the Potsdam dataset, the improvements were 1.06%, 0.9%, and 1.78%, respectively. The SAFM employs a spatially adaptive feature aggregation technique that utilizes spatial information more effectively for feature fusion compared to traditional methods. This approach significantly enhances segmentation by enabling the model to better capture local image details. However, SAFM still faces limitations in capturing multi-scale background information, which can affect the model's understanding of the overall image structure. To address this issue, the introduction of FCLM further improved the model's performance, demonstrating the effectiveness of combining FCLM with SAFM. The feature interaction mechanism in FCLM enables the network to better integrate multi-level feature information, thereby enhancing the model's ability to interpret complex background details.

The superiority of the FCLM+SAFM+CGM: After integrating the three modules, higher accuracy was successfully achieved on both datasets. Specifically, compared to the baseline model, on the Vaihingen dataset, the mF1, OA, and mIoU were improved by at least 1.86%, 1.13%, and 2.98%, respectively; while on the Potsdam dataset, the corresponding improvements were 1.77%, 1.78%, and 3.04%, respectively. These results validate that the integration of FCLM, SAFM, and CGM enables the model to capture more comprehensive

and valuable features, while also utilizing class-specific information during the feature fusion process. This improves the quality of feature fusion and boosts the model's overall performance in semantic segmentation. By integrating multi-scale features, the strategy enhances the model's understanding of both local image details and global structure, significantly optimizing segmentation results and demonstrating the method's substantial value in analyzing complex remote-sensing images.

4.4 Comparison with State-of-the-Art Methods

To verify our method, we compare it with several popular approaches on the Vaihingen and Potsdam datasets. The algorithms compared include a variety of CNN models, such as the standard FCN [6] and DeepLabV3+ [8]; models that incorporate attention mechanisms, such as A^2-FPN [33] and MAResU-Net [34]; LoG-GAN [24], which utilizes class information; and advanced methods that integrate transformer architecture, including MLFMNet [10], MSGCNet [13], SegFormer [41], and UnetFormer [42].

Comparison with other methods on the Vaihingen dataset: Table 3 presents the quantitative analysis results of our semantic segmentation task on the Vaihingen dataset. The highest values in each column are bolded. The data in Table 3 shows that our proposed CG-FCLNet method achieves the best performance in the three key indicators of mF1, mIoU, and OA. Compared with the suboptimal method MSGCNet, mF1 improved by 0.23%, mIoU by 0.37%, and OA by 0.17%. In particular, in the small-size target Car category, our F1 and IoU reach 89.79% and 81.46%, respectively, which are 0.34% and 0.54% higher than the second-ranked MSGCNet. These results demonstrate that our method excels not only in the recognition and segmentation of large-scale objects but also in capturing and segmenting small-scale objects.

Table 3: Results on the Vaihingen dataset, the highest values in each column are bolded

Model	F1 Score (%)/IoU Score (%)							
	Imp. surf	Building	Lowveg	Tree	Car	mF1	mIoU	OA
FCN	92.18/85.50	94.80/90.11	88.04/72.48	89.82/81.52	84.93/73.81	89.15	80.68	90.36
Deeplabv3+	92.82/86.61	95.51/91.40	84.91/73.78	90.23/82.19	87.94/78.48	90.28	82.49	91.03
MLFMNet	92.88/86.71	94.87/90.23	**85.21/74.23**	90.12/82.01	88.46/79.30	90.31	82.50	90.88
MSGCNet	93.07/87.05	95.74/91.83	84.69/73.44	90.23/82.19	89.45/80.92	90.64	83.09	91.17
LoG-CAN	93.06/87.01	95.65/91.67	84.67/73.42	90.28/82.29	89.32/80.70	90.60	83.02	91.14
A^2-FPN	**93.08/87.06**	95.45/91.29	84.66/73.40	90.07/81.93	88.62/79.57	90.38	82.65	91.02
MAResU-Net	92.99/86.89	95.59/91.55	84.66/73.40	90.35/82.39	87.80/78.26	90.28	82.50	91.11
Segformer	92.31/85.71	94.54/89.64	84.69/73.44	90.34/82.39	87.24/77.36	89.82	81.71	90.56
Unetformer	93.07/87.03	95.66/91.69	84.73/73.50	90.31/82.33	88.25/78.97	90.40	82.71	91.14
Ours	93.06/87.02	**95.87/92.07**	85.16/74.16	**90.46/82.58**	**89.79/81.46**	**90.87**	**83.46**	**91.34**

To visually compare the segmentation results of different algorithms, we show the segmentation results on the Vaihingen dataset, as shown in Fig. 5. In the first case, our method demonstrated excellent performance in recognizing and segmenting small objects, such as cars and buildings, accurately identifying and delineating their edges. In contrast, other methods exhibited varying degrees of omission and errors during the recognition process. Specifically, in the small building area marked by the middle red box, except for our method and MLFMNet, other methods failed to fully recognize certain building structures. For the small cars in the other two red boxes, our method not only accurately identified the cars but also clearly outlined their edge contours, while other methods showed blurred boundaries and misidentification issues. In the second case, the building in the left red box had significant appearance differences compared to the other buildings, with its color and texture resembling those of Lowveg and cars. This similarity led other

models to incorrectly classify the building as Lowveg and the car as a building. However, by incorporating context information processing and fine-grained feature learning mechanisms into the feature interaction module and integrating category information, our method was able to capture subtle differences in the image more precisely, successfully distinguishing between these similar categories and avoiding misclassification.

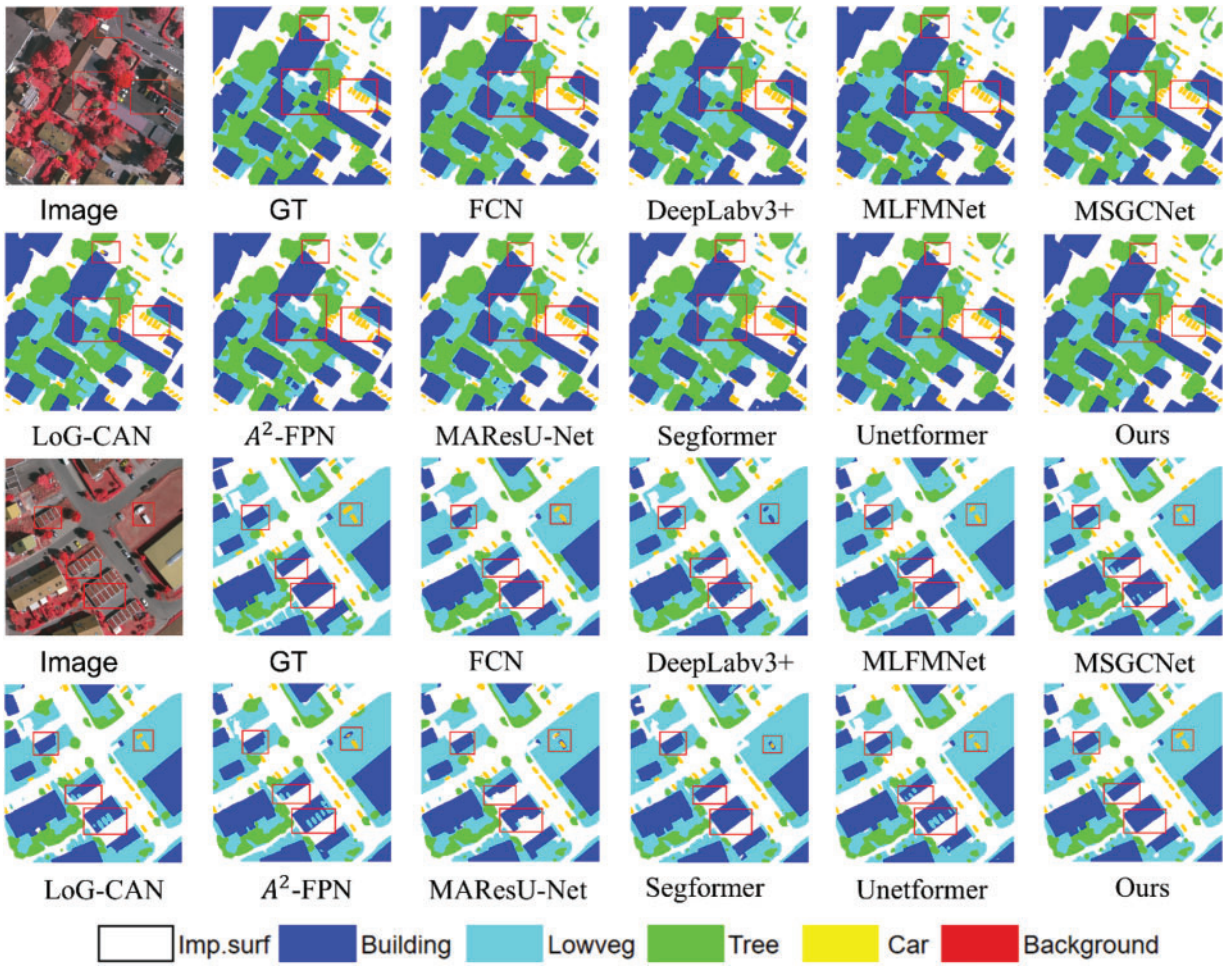

Figure 5: Visualization comparisons on the Vaihingen dataset. The "Image" represents the input RGB images, and the "GT" represents ground truth

Comparison with other methods on the Potsdam dataset: As with the Vaihingen dataset, we conducted experiments on the Potsdam dataset. The results presented in Table 4 demonstrate that our proposed CG-FCLNet method has achieved the highest scores in mF1, mIoU, and OA (the highest values in each column are bolded). Compared to the second-best performing MSGCNet method, our mF1 increased by 0.24%, mIoU by 0.45%, and OA by 0.22%. Additionally, our method has attained the highest values in F1 score and IoU for each category, except Lowveg and Tree. Particularly, in the small target Car category, our F1 and IoU reached 96.50% and 93.23%, respectively, which is 0.37% and 0.69% higher than the MLFMNet.

Table 4: Results on the Potsdam dataset, the highest values in each column are bolded

Model	F1 Score (%)/IoU Score (%)							
	Imp. surf	Building	Lowveg	Tree	Car	mF1	mIoU	OA
FCN	92.13/85.41	94.47/89.52	86.46/76.15	87.87/78.36	94.45/89.49	91.08	83.79	89.69
Deeplabv3+	93.43/87.68	95.99/92.30	87.02/77.02	88.35/79.13	95.02/90.52	91.96	85.33	90.72
MLFMNet	93.44/87.69	96.15/92.59	87.59/77.93	88.56/79.46	96.13/92.54	92.37	86.04	90.98
MSGCNet	93.53/87.84	95.95/92.21	**87.78/78.22**	88.71/79.71	96.09/92.48	92.41	86.09	91.07
LoG-CAN	93.34/87.52	95.80/91.93	87.19/77.29	88.51/79.38	95.62/91.60	92.09	85.54	90.82
A^2-FPN	93.67/88.09	96.27/92.82	87.23/77.35	**88.76/79.79**	95.16/90.76	92.22	85.76	91.01
MAResU-Net	93.48/87.75	96.33/92.92	87.47/77.73	88.58/79.51	95.88/92.09	92.35	86.00	91.02
Segformer	93.51/87.80	95.96/92.23	87.28/77.44	88.15/78.81	94.96/90.40	91.97	85.34	90.82
Unetformer	93.52/87.83	96.37/92.99	87.08/77.12	88.16/78.83	96.08/92.45	92.24	85.85	90.93
Ours	**93.90/88.49**	**96.50/93.24**	87.73/78.15	88.64/79.59	**96.50/93.23**	**92.65**	**86.54**	**91.29**

To visually compare the segmentation results of different algorithms, we show the segmentation results on the Potsdam dataset, as shown in Fig. 6. In the first case, although the texture of Lowveg in the red box is similar to other categories like Imp. surf and Building, which increases the difficulty of distinguishing between these categories, our model still performs accurate segmentation. This demonstrates the model's ability to handle categories with similar texture features and capture subtle differences. In the second case, the trees within the red box on the left have complex shapes and blurred edges. Our model successfully identifies and accurately segments the trees through fine-grained feature learning and contextual information processing, preventing the merging of segmentation areas and ensuring the independence of the tree regions. In the background area within the red box at the top right, the color and texture resemble those of buildings, which other models may mistakenly identify as buildings. In contrast, our model effectively distinguishes between the background and the buildings.

In addition, we evaluated our model on low-resolution images. Low-resolution images were generated from high-resolution ones through a 4× downsampling factor. The results of comparison with other methods (namely MLFMNet [10], LoG-GAN [24], and SegFormer [41]) are shown in Table 5. The highest values in each column are in bold. It can be seen that under the 4× degradation factors on both datasets, the model achieved the best performance in terms of mF1, mIoU, and OA metrics (except for the OA on the Potsdam dataset), demonstrating its ability to effectively capture important features and maintain high segmentation accuracy under low-resolution conditions. However, the loss of details and contextual information in low-resolution images may limit the capability of the feature interaction module, leading to performance degradation compared to high-resolution images.

Fig. 7 illustrates the trend of OA changes concerning epochs for training and testing data on the Vaihingen and Potsdam datasets. In both datasets, the training accuracy (represented by the blue solid line) rises rapidly in the early stages of training, then slows down, eventually stabilizing in the later stages. Notably, the training accuracy remains consistently above 0.90, reflecting the model's ability to effectively extract features from the training set and perform excellently on both datasets. Similarly, the testing accuracy (represented by the red dashed line) follows a similar growth pattern. The gap between training and testing accuracy is relatively small in both datasets, indicating that the model not only performs well on the training data but also maintains high accuracy on unseen data, demonstrating strong generalization capability.

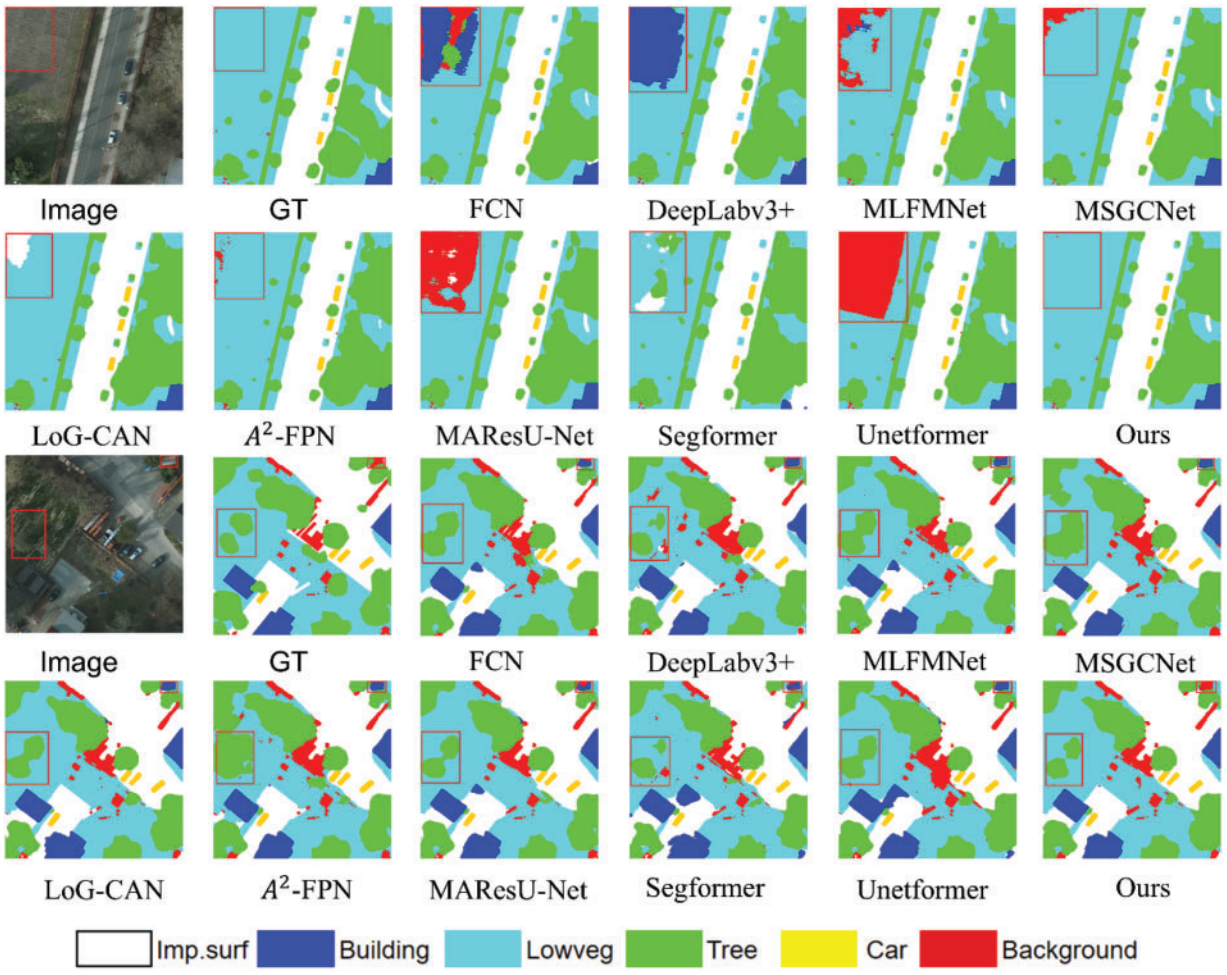

Figure 6: Visualization comparisons on the Potsdam dataset. The "Image" represents the input RGB images, and the "GT" represents ground truth

Table 5: Comparison of our model with other methods on low-resolution images the highest values in each column are in bold

Model	Degradation factors	Vaihingen			Potsdam		
		mF1	mIoU	OA	mF1	mIoU	OA
MLFMNet	4×	82.62	71.62	87.44	88.00	78.95	87.35
LoG-CAN	4×	81.94	70.77	87.04	87.46	78.10	86.81
Segformer	4×	82.35	71.25	87.17	87.90	78.67	**87.67**
Ours	4×	**83.20**	**72.44**	**87.87**	**88.29**	**79.33**	87.54

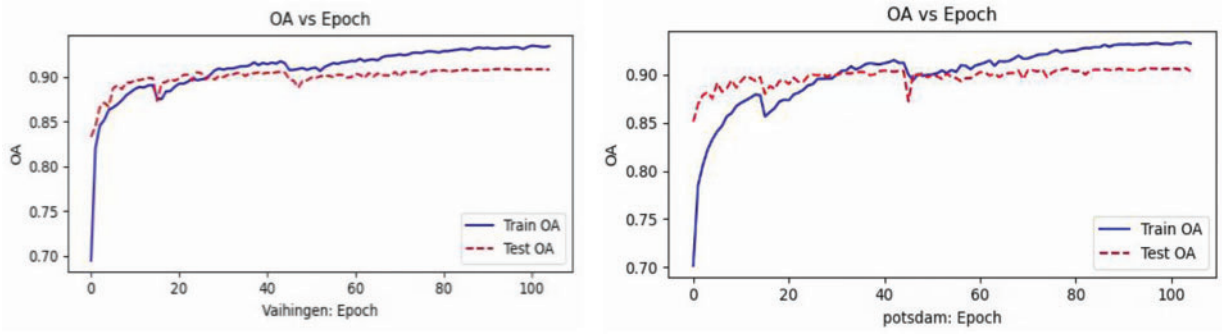

Figure 7: Comparison of OA between training and testing data

4.5 Efficiency Analysis

The efficiency of deep learning methods can be evaluated using metrics such as model parameters, floating-point operations (FLOPs), and frames per second (FPS). FLOPs assess complexity, FPS evaluate speed, and model parameters (M) evaluate storage space consumption. We input data of size $3 \times 1024 \times 1024$ into all models and evaluated their efficiency under the same running environment. The evaluation results are shown in Table 6, the best values in each column are in bold. As seen in the table, CG-FCLNet's model parameters and computational complexity are moderate compared to other methods. However, the inclusion of multiple collaborative modules in feature extraction increases inference complexity, leading to a lower FPS.

Table 6: Comparison of efficiency with state-of-the-art methods the best values in each column are in bold

Model	Params (Mb)	FLOPs (Gbps)	FPS
FCN	**11.42**	**39.87**	**136.13**
Deeplabv3+	21.32	184.66	52.64
MLFMNet	14.55	74.69	68.97
MSGCNet	27.61	114.54	36.47
LoG-CAN	12.48	61.68	71.55
A^2-FPN	12.16	167.33	76.99
MAResU-Net	16.17	101.66	33.99
Segformer	13.68	53.04	26.45
Unetformer	11.72	46.97	88.03
Ours	13.00	62.86	37.92

5 Conclusion

In this paper, we propose a novel network framework for the semantic segmentation of remote sensing images, called the Category-Guided Feature Collaborative Learning Network (CG-FCLNet). It incorporates the Feature Collaborative Learning Module (FCLM), which facilitates interactions between features at different levels and uncovers their intrinsic relationships. This improves the model's understanding of complex scenes. To further enhance performance, we design a Scale-Aware Fusion Module (SAFM) that uses a spatial attention mechanism to iteratively fuse multi-level features, significantly boosting segmentation

accuracy. Additionally, the Category-Guided Module (CGM) optimizes feature fusion by guiding category-aware information, ensuring that the fused features more accurately reflect the semantic information of the categories. Experimental results show that CG-FCLNet outperforms existing methods on the Potsdam and Vaihingen datasets, achieving higher segmentation accuracy and better detail preservation.

In future work, we are committed to enhancing the precision of our model, particularly in addressing the issue of unclear boundaries in low-resolution images. We will focus on optimizing the model's ability to learn and process edge features, with the aim of improving the accuracy of edge segmentation.

Acknowledgement: The authors would like to express their sincere gratitude to all those who contributed to this research. Their support and efforts were key factors in the success of this study. Additionally, special thanks to the editors and reviewers for their critical and constructive comments and suggestions.

Funding Statement: This research was funded by National Natural Science Foundation of China (61603245).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Min Yao, Guangjie Hu; data collection: Guangjie Hu, Yaozu Zhang; analysis and interpretation of results: Min Yao, Guangjie Hu, Yaozu Zhang; draft manuscript preparation: Min Yao, Guangjie Hu, Yaozu Zhang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data provided in this study are available upon request from the corresponding author.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Abbreviations

CG-FCLNet	Category-Guided Feature Collaborative Learning Network
FCLM	Feature Collaborative Learning Module
SAFM	Scale-Aware Fusion Module
CGM	Category-Guided Module
FCM	Feature Collaborative Module
CSM	Channel Selection Module
EUB	Efficient Up-convolution Block
SAM	Scale-Aware Model
mIoU	Mean Intersection over Union
mF1	Mean F1 Score
OA	Overall Accuracy
FLOPs	Floating-Point Operations
FPS	Frames Per Second

References

1. Kotaridis I, Lazaridou M. Remote sensing image segmentation advances: a meta-analysis. ISPRS J Photogramm Remote Sens. 2021;173(3):309–22. doi:10.1016/j.isprsjprs.2021.01.020.
2. Li R, Zheng S, Duan C, Wang L, Zhang C. Land cover classification from remote sensing images based on multi-scale fully convolutional network. Geo Spatial Inf Sci. 2022;25(2):278–94. doi:10.1080/10095020.2021.2017237.
3. Lv Z, Yang T, Lei T, Zhou W, Zhang Z, You Z. Spatial-spectral similarity based on adaptive region for landslide inventory mapping with remote-sensed images. IEEE Trans Geosci Remote Sens. 2024;62:4405111. doi:10.1109/TGRS.2024.3380199.

4. Xu Q, Long C, Yu L, Zhang C. Road extraction with satellite images and partial road maps. IEEE Trans Geosci Remote Sens. 2023;61:4501214. doi:10.1109/TGRS.2023.3261332.
5. Sun Z, Zhou W, Ding C, Xia M. Multi-resolution transformer network for building and road segmentation of remote sensing image. ISPRS Int J Geo Inf. 2022;11(3):165. doi:10.3390/ijgi11030165.
6. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(4):640–51. doi:10.1109/TPAMI.2016.2572683.
7. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. Medical image computing and computer-assisted intervention–MICCAI 2015. Lecture notes in computer science. Vol. 9351. Cham: Springer; 2015. doi:10.1007/978-3-319-24574-4_28.
8. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer vision—ECCV 2018. Lecture notes in computer science. Vol. 11211. Cham: Springer; 2018. doi:10.1007/978-3-030-01234-2_49.
9. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26. Honolulu, HI, USA: IEEE; 2017. p. 6230–9. doi:10.1109/CVPR.2017.660.
10. Wei X, Rao L, Fan G, Chen N. MLFMNet: a multilevel feature mining network for semantic segmentation on aerial images. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;17:16165–79. doi:10.1109/JSTARS.2024.3452250.
11. Bai Q, Luo X, Wang Y, Wei T. DHRNet: a dual-branch hybrid reinforcement network for semantic segmentation of remote sensing images. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;17:4176–93. doi:10.1109/JSTARS.2024.3357216.
12. Wang Z, Zhang S, Zhang C, Wang B. Hidden feature-guided semantic segmentation network for remote sensing images. IEEE Trans Geosci Remote Sens. 2023;61:5603417. doi:10.1109/TGRS.2023.3244273.
13. Zeng Q, Zhou J, Tao J, Chen L, Niu X, Zhang Y. Multiscale global context network for semantic segmentation of high-resolution remote sensing images. IEEE Trans Geosci Remote Sens. 2024;62:5622913. doi:10.1109/TGRS.2024.3393489.
14. Jiang H, Luo T, Peng H, Zhang G. MFCANet: multiscale feature context aggregation network for oriented object detection in remote-sensing images. IEEE Access. 2024;12(10):45986–6001. doi:10.1109/ACCESS.2024.3381539.
15. Li Y, Meng W, Ma D, Xu S, Zhu X. MCGFF-Net: a multi-scale context-aware and global feature fusion network for enhanced polyp and skin lesion segmentation. Vis Comput. 2024;30(2):584. doi:10.1007/s00371-024-03720-9.
16. Li G, Bai Z, Liu Z, Zhang X, Ling H. Salient object detection in optical remote sensing images driven by transformer. IEEE Trans Image Process. 2023;32:5257–69. doi:10.1109/TIP.2023.3314285.
17. Xia C, Chen X, Sun Y, Ge B, Fang X, Gao X, et al. CEMINet: context exploration and multi-level interaction network for salient object detection. Digit Signal Process. 2024;147(4):104403. doi:10.1016/j.dsp.2024.104403.
18. Wu H, Zhang M, Huang P, Tang W. CMLFormer: CNN and multiscale local-context transformer network for remote sensing images semantic segmentation. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;17:7233–41. doi:10.1109/JSTARS.2024.3375313.
19. Li X, Xu F, Liu F, Tong Y, Lyu X, Zhou J. Semantic segmentation of remote sensing images by representation refinement and geometric prior-guided inference. IEEE Trans Geosci Remote Sens. 2023;62:5400318. doi:10.1109/TGRS.2023.3339291.
20. Bhatti UA, Yu Z, Chanussot J, Zeeshan Z, Yuan L, Luo W, et al. Local similarity-based spatial-spectral fusion hyperspectral image classification with deep CNN and Gabor filtering. IEEE Trans Geosci Remote Sensing. 2022;60:5514215. doi:10.1109/TGRS.2021.3090410.
21. Zhang F, Chen Y, Li Z, Hong Z, Liu J, Ma F, et al. ACFNet: attentional class feature network for semantic segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2. Seoul, Republic of Korea: IEEE; 2019. p. 6797–806. doi:10.1109/iccv.2019.00690.
22. Yuan Y, Chen X, Wang J. Object-contextual representations for semantic segmentation. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Computer vision—ECCV 2020. ECCV 2020. Lecture notes in computer science. Vol. 12351. Cham: Springer; 2020. doi:10.1007/978-3-030-58539-6_11.

23. Ma X, Che R, Wang X, Ma M, Wu S, Feng T, et al. DOCNet: dual-domain optimized class-aware network for remote sensing image segmentation. IEEE Geosci Remote Sensing Lett. 2024;21:2500905. doi:10.1109/LGRS.2024.3350211.
24. Ma X, Ma M, Hu C, Song Z, Zhao Z, Feng T, et al. LoG-CAN: local-global class-aware network for semantic segmentation of remote sensing images. In: ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 Jun 4–10. Rhodes Island, Greece: IEEE; 2023. p. 1–5. doi:10.1109/ICASSP49357.2023.10095835.
25. Du S, Liu M. Class-guidance network based on the pyramid vision transformer for efficient semantic segmentation of high-resolution remote sensing images. IEEE J Sel Top Appl Earth Obs Remote Sens. 2023;16:5578–89. doi:10.1109/JSTARS.2023.3285632.
26. Nie J, Wang Z, Liang X, Yang C, Zheng C, Wei Z. Semantic category balance-aware involved anti-interference network for remote sensing semantic segmentation. IEEE Trans Geosci Remote Sens. 2023;61:4409712. doi:10.1109/TGRS.2023.3325327.
27. Wang S, Hu Q, Wang S, Zhao P, Li J, Ai M. Category attention guided network for semantic segmentation of Fine-Resolution remote sensing images. Int J Appl Earth Obs Geoinf. 2024;127:103661. doi:10.1016/j.jag.2024.103661.
28. Liu T, Cheng S, Yuan J. Category-based interactive attention and perception fusion network for semantic segmentation of remote sensing images. Remote Sens. 2024;16(20):3864. doi:10.3390/rs16203864.
29. Meng X, Zhang S, Liu Q, Yang G, Sun W. Uncertain category-aware fusion network for hyperspectral and LiDAR joint classification. IEEE Trans Geosci Remote Sens. 2024;62:5523015. doi:10.1109/TGRS.2024.3424829.
30. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23. Salt Lake City, UT, USA: IEEE; 2018. p. 7132–41. doi:10.1109/CVPR.2018.00745.
31. Woo J, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Ferrari V, Hebert M, Sminchisescu, C, Weiss Y, editors. Computer vision—ECCV 2018. Lecture notes in computer science. Vol. 11211. Cham: Springer; 2018. doi:10.1007/978-3-030-01234-2_1.
32. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, et al. Dual attention network for scene segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20. Long Beach, CA, USA: IEEE; 2019. p. 3141–9. doi:10.1109/cvpr.2019.00326.
33. Li R, Wang L, Zhang C, Duan C, Zheng S. A^2-FPN for semantic segmentation of fine-resolution remotely sensed images. Int J Remote Sens. 2022;43(3):1131–55. doi:10.1080/01431161.2022.2030071.
34. Li R, Zheng S, Duan C, Su J, Zhang C. Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. IEEE Geosci Remote Sens Lett. 2021;19:8009205. doi:10.1109/LGRS.2021.3063381.
35. Niu R, Sun X, Tian Y, Diao W, Chen K, Fu K. Hybrid multiple attention network for semantic segmentation in aerial images. IEEE Trans Geosci Remote Sensing. 2022;60:5603018. doi:10.1109/TGRS.2021.3065112.
36. Hu X, Zhang P, Zhang Q, Yuan F. GLSANet: global-local self-attention network for remote sensing image semantic segmentation. IEEE Geosci Remote Sens Lett. 2023;20:6000105. doi:10.1109/LGRS.2023.3235117.
37. Feng S, Zhao H, Shi F, Cheng X, Wang M, Ma Y, et al. CPFNet: context pyramid fusion network for medical image segmentation. IEEE Trans Med Imaging. 2020;39(10):3008–18. doi:10.1109/TMI.2020.2983721.
38. Gu R, Wang G, Song T, Huang R, Aertsen M, Deprest J, et al. CA-Net: comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans Med Imaging. 2021;40(2):699–711. doi:10.1109/TMI.2020.3035253.
39. Rahman MM, Munir M, Marculescu R. EMCAD: efficient multi-scale convolutional attention decoding for medical image segmentation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22. Seattle, WA, USA: IEEE; 2024. p. 11769–79. doi:10.1109/CVPR52733.2024.01118.
40. Zhang X, Zhou X, Lin M, Sun J. ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23. Salt Lake City, UT, USA: IEEE; 2018. p. 6848–56. doi:10.1109/CVPR.2018.00716.

41. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: simple and efficient design for semantic segmentation with transformers. Adv Neural Inf Process Syst. 2021;34:12077–90.
42. Wang L, Li R, Zhang C, Fang S, Duan C, Meng X, et al. UNetFormer: a UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J Photogramm Remote Sens. 2022;190:196–214. doi:10.1016/j.isprsjprs.2022.06.008.