Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.083877

ARTICLE

BFANet: Fine-Grained Boundary-Aware Semantic Segmentation Driven by Dynamic Feature Alignment

Wang Zhang[1], **Lanlan Li**[2], **Jiayi Xing**[1] **and Qiangqiang Yao**[1,*]

[1]School of Mechanical Engineering, Qinghai University, Xining, China
[2]School of Medicine, Qinghai University, Xining, China

*Corresponding Author: Qiangqiang Yao. Email: yaoqiang@bjtu.edu.cn

Received: 12 April 2026; Accepted: 23 June 2026; Published: 13 August 2026

ABSTRACT: Lightweight semantic segmentation remains challenging because compact backbones often weaken feature discriminability and lose fine-grained boundary details. In DeepLabV3+-style encoder-decoder architectures, the direct fusion of high-level semantic features and low-level spatial features may introduce semantic-spatial misalignment, resulting in blurred object contours and fragmented predictions. To address these issues, this paper proposes BFANet, a boundary-aware lightweight semantic segmentation framework based on DeepLabV3+ with a MobileNetV2 backbone. BFANet integrates parameter-free SimAM feature refinement, low-level-guided Dynamic Feature Alignment, and progressive decoder fusion to enhance discriminative feature responses, reduce cross-level feature inconsistency, and recover fine boundary structures. Experiments on a curated PASCAL VOC 2012 subset demonstrate that BFANet achieves consistent improvements over the DeepLabV3+ baseline with only a negligible increase in model complexity. Specifically, BFANet improves mIoU from 77.57% to 80.23%, mPA from 86.36% to 88.05%, and mDice from 87.52% to 88.78%. Moreover, the boundary-level evaluation further shows that BFANet improves Boundary F1 from 57.28% to 61.55% and Boundary IoU from 48.74% to 51.42%, indicating better preservation of fine object contours and more accurate boundary alignment. Meanwhile, the parameter count increases only from 5.82M to 5.84M, and GFLOPs increase from 53.03 to 53.75. These results show that BFANet provides a better balance among segmentation accuracy, boundary quality, and lightweight efficiency, especially for small objects, thin structures, and boundary-sensitive categories.

KEYWORDS: Semantic segmentation; boundary-aware; dynamic feature alignment; lightweight network; fine-grained segmentation; DeepLabV3+

1 Introduction

Semantic segmentation is a core task in computer vision, aiming to assign a semantic label to each pixel in an input image. With the increasing demand for deploying models on embedded and resource-constrained devices, lightweight architectures with low computational overhead and fast inference have become essential. Consequently, lightweight semantic segmentation has attracted substantial research attention in recent years. A series of lightweight segmentation architectures including LDMSNet and MobileNetV2-based segmentation schemes prove that convolutional neural networks can be efficiently compressed for mobile vision applications with competitive performance [1,2], later enabling compact dense prediction architectures. Based on these advances, DeepLabV3+ has been widely adopted as a representative segmentation framework due to its atrous spatial pyramid pooling for multi-scale context modeling and effective encoder–decoder structure with skip connections [3].

Despite these advances, high-quality semantic segmentation under strict computational constraints remains challenging. Lightweight backbones typically employ aggressive channel reduction and repeated downsampling, which reduces representational capacity. This design improves efficiency but leads to the loss of fine-grained spatial information, resulting in blurred object boundaries, fragmented predictions for thin structures, and jittered boundaries for small objects. Classical attention mechanisms, such as CBAM, enhance feature selectivity through spatial and channel attention but require additional parameters and computation [4]. In contrast, SimAM computes attention weights without additional parameters by inferring neuron importance based on energy functions in a 3D space, making it particularly suitable for compact networks [5].

Recent lightweight segmentation methods have explored efficient designs from multiple perspectives. BiSeNet and BiSeNetV2 employ separate spatial and context paths to balance efficiency and accuracy [6,7]. Efficient segmentation architectures have explored different strategies, including lightweight dual-branch designs and detail-preserving feature aggregation [8]. Transformer-inspired lightweight networks, such as TopFormer, demonstrate that efficient multi-scale semantic aggregation can be achieved in compact architectures through a token pyramid structure [9]. Nevertheless, boundary degradation and feature inconsistency remain unsolved, especially in encoder-decoder pipelines with lightweight backbones.

The PASCAL VOC 2012 benchmark is commonly used to evaluate segmentation models and measure pixel-level accuracy [10]. On this dataset, lightweight models often fail to segment fine boundaries and structurally complex categories. This issue is largely caused by semantic-spatial misalignment between feature levels. Low-level features retain texture and boundary cues but have weak semantics, while high-level features contain strong semantic abstraction but poor spatial localization. Direct fusion of these heterogeneous features in the decoder results in blurred boundaries and inaccurate predictions.

Several approaches have been proposed to address these limitations. STDC-based segmentation improves efficiency through short-term dense concatenation in the backbone design [11]. To address blurry boundaries in segmentation, Hu et al. [12] developed a plug-and-play boundary patch refinement (BPR) framework, which improves boundary quality by extracting and refining high-resolution local boundary patches. Alignment-aware methods address this issue: FaPN aligns features across pyramid levels before fusion [13], while Semantic Flow uses flow-guided propagation to warp and align multi-scale features [14]. These studies indicate that effective lightweight segmentation requires the preservation of efficiency, enhancement of discriminative responses, reduction of semantic-spatial inconsistency, and reinforcement of boundary-aware representations. Representative studies, such as BCINet, MMSMCNet, and EGFNet, as well as related RGB-D/RGB-T fusion methods, have explored progressive guided fusion, bilateral cross-modal interaction, modal memory sharing, morphological complementary learning, and edge-aware fusion [15–18]. Although these methods are mainly designed for RGB-D or RGB-T tasks, they offer useful insights for lightweight RGB semantic segmentation, particularly in enhancing complementary cues, reducing inconsistent responses, and preserving boundary-sensitive information during feature fusion. This observation further motivates the design of a lightweight decoder that can refine features, align heterogeneous representations, and strengthen boundary details without significantly increasing computational cost.

Motivated by these observations, this paper proposes BFANet, a boundary-aware lightweight semantic segmentation framework based on DeepLabV3+ with a MobileNetV2 backbone. BFANet explicitly targets the above failure modes—blurred boundaries, fragmented thin structures, and jittered small object contours—through three complementary components: (1) a SimAM-based feature refinement strategy to enhance both ASPP output features and projected low-level features in a parameter-free manner; (2) a Dynamic Feature Alignment (DFA) module that leverages low-level structural cues to recalibrate upsampled semantic features, reducing spatial misalignment; (3) an improved multi-stage decoder fusion strategy that

strengthens the integration of multi-scale context and boundary information. Extensive experiments on a curated subset of PASCAL VOC 2012 consisting of 2913 images show that BFANet improves the baseline mIoU (DeepLabV3+ with MobileNetV2 backbone) from 77.57% to 80.23% and mDice from 87.52% to 88.78%, particularly enhancing segmentation of thin structures, small objects, and structurally complex categories, without introducing significant computational overhead.

The main contributions of this paper are summarized as follows:

(1) We propose a boundary-aware feature recalibration principle for lightweight semantic segmentation, which addresses the semantic–spatial inconsistency by adaptively refining high-level semantic features under the guidance of low-level structural information.
(2) Based on this principle, we design a unified segmentation framework that integrates content-adaptive alignment and feature refinement mechanisms, enabling consistent feature representation across different scales while maintaining low computational overhead.
(3) Extensive experiments on the PASCAL VOC 2012 benchmark demonstrate that the proposed method achieves a +2.66% mIoU improvement over the baseline with negligible increase in parameters and computational cost, validating the effectiveness and efficiency of the proposed design.

The rest of this paper is organized as follows: In Section 2, we review related work on lightweight segmentation, attention mechanisms, boundary-aware segmentation, and feature alignment. Section 3 presents the proposed BFANet, detailing the overall architecture, the MobileNetV2 encoder, SimAM-based feature refinement, Dynamic Feature Alignment, decoder fusion, and classification head. Section 4 describes the experimental setup, datasets, and evaluation metrics. Section 5 provides experimental results, including ablation studies, attention mechanism comparisons, and performance evaluation against state-of-the-art methods. Finally, Section 6 concludes the paper and discusses potential future directions.

2 Related Work

2.1 Multi-Scale Structural Deblurring Methods

Building on these architectures, MobileNet-based backbones, originally designed for efficient image classification, have been widely adopted and adapted for lightweight segmentation tasks. MobileNetV3 further refines lightweight backbone design for mobile applications [19], while PIDNet introduces a three-branch architecture that separately processes detail, context, and boundary information through parallel pathways, enabling real-time semantic segmentation [20].

Transformer-inspired methods have also influenced lightweight segmentation. Segformer adopts hierarchical Transformer representations with an efficient decoder, achieving effective multi-scale semantic feature extraction without positional encoding [21]. Other efficient architectures have explored various strategies for real-time segmentation: ICNet uses image cascades and multi-resolution branches [22], ERFNet employs factorized convolutions [23], ESPNet introduces efficient spatial pyramids [24], and CGNet leverages context-guided blocks [25]. RTFormer demonstrates that hybrid architectures combining efficient transformers with convolutional layers can achieve real-time semantic segmentation with competitive accuracy [26]. Despite these advances, several limitations remain. Lightweight models often fail to accurately segment boundary-sensitive or structurally complex categories. Moreover, encoder-decoder pipelines commonly fuse high-level and low-level features without explicit alignment, leading to cross-level ambiguity. Finally, lightweight backbones may not preserve sufficient discriminative capacity under strict computational budgets. Unlike approaches that redesign entire backbones or introduce complex multi-branch structures, BFANet improves feature refinement and decoder fusion within the widely used MobileNetV2-DeepLabV3+ framework, facilitating integration into existing lightweight segmentation systems.

2.2 Attention Mechanisms and Boundary-Aware Segmentation

Boundary-aware segmentation is another critical line of research. Attention mechanisms have been extensively applied to improve feature representation in dense prediction tasks. Attention-based context modeling methods, such as DANet, CCNet, OCRNet, and position-aware attention networks, improve long-range dependency representation for semantic segmentation [27–30]. Lightweight attention modules, including ECA-Net and Coordinate Attention, improve feature representation through efficient attention recalibration while maintaining low computational overhead, and have demonstrated effectiveness in recognition and segmentation tasks [31,32]. BASeg introduces boundary refinement and boundary-guided context aggregation to improve contour-aware representation [33]. EPS proposes a plug-and-play edge-aware supervision strategy that can be integrated into different segmentation networks [34]. However, these methods often rely on explicit boundary supervision, auxiliary edge branches, or additional training constraints. In contrast, BFANet focuses on implicit boundary-guided feature alignment within a lightweight decoder, avoiding extra edge labels and heavy auxiliary structures.

More recent transformer-based segmentation approaches, including Segmenter, MaskFormer, and kMaX-DeepLab, demonstrate that effective decoder design and feature integration strategies are critical for achieving accurate fine-grained segmentation [35–37]. Ruan et al. [38] propose a context-aware feature enhancement network that leverages multi-scale contextual information to improve semantic segmentation accuracy for complex remote sensing images.

Although effective, many boundary-aware approaches rely on auxiliary branches, heavy decoding structures, or specialized loss functions, which increase computational cost. In contrast, BFANet enhances boundary-sensitive representations economically: SimAM strengthens discriminative low-level and contextual features, while the DFA module transfers boundary cues to the aligned high-level feature stream. This design preserves the simplicity of the original encoder-decoder pipeline while improving contour quality in a lightweight manner.

2.3 Feature Alignment and Decoder Fusion

Feature fusion is central in encoder-decoder segmentation, as it directly affects pixel-level prediction quality. In DeepLabV3+, high-level features are bilinearly upsampled and concatenated with low-level features, which is efficient but does not explicitly resolve semantic-spatial inconsistencies [3]. Studies have shown that this misalignment causes misclassification around object boundaries and fine structures. FaPN addresses this issue by learning feature alignment in dense prediction, highlighting the importance of alignment for boundary accuracy [13]. Semantic Flow argues that direct cross-scale pixel addition or concatenation often produces contextually misaligned features, and proposes a principled alignment strategy [14]. These studies indicate that effective decoder fusion requires not only semantic complementarity but also spatial consistency across feature levels. Real-time segmentation works, such as Fast-SCNN and ContextNet, emphasize the importance of efficient detail recovery and context-detail interaction [39,40].

Despite their effectiveness, many alignment-based methods rely on offset learning, flow estimation, or heavy feature pyramids, which are unsuitable for lightweight settings. The DFA module in BFANet addresses these limitations by generating content-adaptive alignment weights from low-level structural features to recalibrate upsampled semantic features in feature space. Combined with multi-stage decoder refinement and SimAM, it allows recovery of sharp boundaries and consistent spatial layouts without significant additional complexity.

2.4 Summary

Significant progress has been made in lightweight semantic segmentation, attention modeling, boundary refinement, and feature alignment. However, few methods simultaneously achieve boundary-aware refinement, lightweight alignment, and efficient feature fusion within a compact encoder-decoder framework. BFANet is proposed to fill this gap, integrating SimAM, DFA, and multi-stage decoder fusion to improve boundary-sensitive segmentation efficiently.

3 Proposed Method

3.1 Overall Architecture

Semantic segmentation networks based on encoder–decoder architectures, such as DeepLabV3+, have demonstrated strong performance across various benchmarks. However, when deployed with lightweight backbones like MobileNetV2, these models often suffer from boundary ambiguity and loss of fine-grained spatial details. This degradation primarily stems from two factors: (1) the limited representational capacity of lightweight encoders, and (2) the semantic–spatial inconsistency introduced during naive feature fusion in the decoder. To address these challenges, we propose a boundary-aware feature recalibration framework, which aims to improve the consistency between high-level semantic features and low-level structural details in lightweight segmentation networks. Instead of treating feature refinement, alignment, and decoding as independent operations, the proposed framework follows a unified principle: high-level semantic features should be adaptively recalibrated under the guidance of low-level structural cues to achieve spatially consistent representation. Under this principle, the proposed method realizes a coordinated design that includes: SimAM-based feature refinement to enhance discriminative responses without introducing additional parameters; a Dynamic Feature Alignment (DFA) mechanism to perform content-adaptive alignment guided by structural information; and an improved decoder that facilitates progressive fusion of multi-scale context and boundary details. These components function as complementary implementations of the same underlying principle rather than independent modifications. The overall architecture is illustrated in Fig. 1.

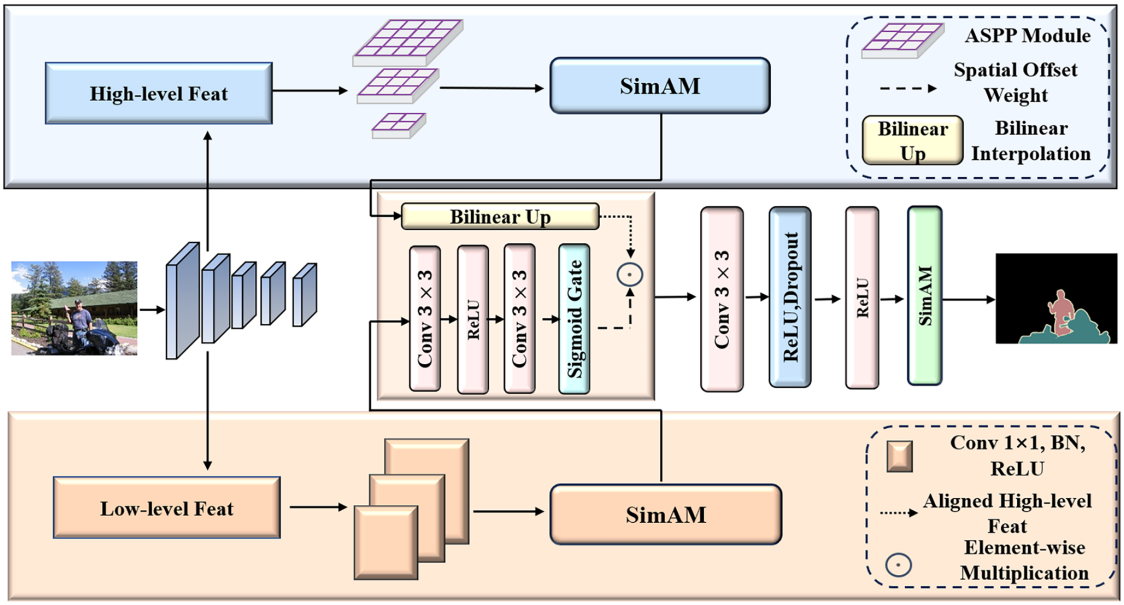

Figure 1: Overall architecture of the proposed BFANet.

3.2 Lightweight Encoder with MobileNetV2

We adopt MobileNetV2 as the feature extraction backbone due to its favorable balance between computational efficiency and representational capacity. MobileNetV2 employs depthwise separable convolutions and inverted residual blocks, which substantially reduce the number of parameters and multiply-accumulate operations compared to standard convolutional networks.

The backbone extracts features at two hierarchical levels: Low-level features from an early stage (after the 4th block), which preserve spatial detail and edge structures; High-level features from a deeper stage, which encode abstract semantic information. In our implementation, the spatial resolution is controlled by the output stride parameter. To maintain a reasonable receptive field without excessive downsampling, atrous convolution is applied in the later stages of the backbone, following the standard practice in DeepLab-like architectures.

3.3 SimAM: Parameter-Free Spatial Attention

Attention mechanisms have been widely adopted to improve feature selectivity in deep networks. However, many existing attention modules introduce substantial parameter overhead, which conflicts with the lightweight design principle. To address this issue, we incorporate SimAM (Simple, Parameter-Free Attention Module), which derives attention weights based on an energy formulation without requiring additional learnable parameters. Given a feature tensor $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, SimAM computes the importance of each spatial location by measuring its deviation from the channel-wise mean. For each element $x_{c,i,j}$, the energy-based response is computed as:

$$e_{c,i,j} = \frac{\left(x_{c,i,j} - \mu_c\right)^2}{4\left(\sigma_c^2 + \lambda\right)} + \frac{1}{2} \tag{1}$$

where μ_c and σ_c^2 denote the mean and variance over the spatial domain for channel c, and λ is a small constant for numerical stability (set to 10^{-4} in our implementation). The attention weight is obtained by applying a sigmoid activation:

$$\alpha_{c,i,j} = \sigma\left(e_{c,i,j}\right) \tag{2}$$

and the refined feature is given by element-wise multiplication:

$$\hat{\mathbf{X}} = \mathbf{X} \odot \sigma\left(\mathbf{E}\right) \tag{3}$$

where $\odot$ denotes element-wise multiplication.

In the proposed network, SimAM is applied at two critical locations: (1) after the ASPP module to selectively enhance multi-scale contextual responses; (2) after the 1×1 projection of low-level features to emphasize boundary-relevant structural cues before fusion with upsampled high-level features.

3.4 Dynamic Feature Alignment Module

In the standard DeepLabV3+ decoder, the ASPP-enhanced high-level feature is bilinearly upsampled and directly concatenated with low-level features. While efficient, this often leads to blurred boundaries and fragmented predictions due to semantic-spatial inconsistency: high-level features encode abstract semantics but lack precise localization, whereas low-level features contain rich boundary details but weak semantic relevance.

To address this, we propose a Dynamic Feature Alignment (DFA) module that adaptively recalibrates upsampled high-level features using low-level structural guidance. Unlike geometric alignment methods that estimate explicit spatial offsets, DFA performs implicit feature-space alignment through content-adaptive modulation. The low-level feature is transformed to generate spatially adaptive alignment weights, which modulate the high-level feature channel-wise and spatially, encouraging local consistency with structural details.

DFA can be viewed as a boundary-guided feature recalibration mechanism: regions with strong boundary responses in low-level features produce distinctive modulation patterns that guide high-level semantics to better align with object contours.

The high-level feature is first upsampled to match the spatial resolution of the low-level feature:

$$\mathbf{F}_{up} = \text{Upsample}\left(\mathbf{F}_c, H_l, W_l\right) \tag{4}$$

where bilinear interpolation is used, and the upsampled feature is resized to match the spatial resolution of the low-level feature.

The low-level feature is passed through a lightweight convolutional transformation to generate spatially adaptive alignment weights:

$$\mathbf{M} = \sigma\left(\phi_{align}\left(\hat{\mathbf{F}}_l\right)\right) \tag{5}$$

which $\phi_{align}(\cdot)$ consists of two convolutional layers (a 3×3 conv followed by a 1×1 conv), and $\sigma(\cdot)$ is the sigmoid activation. The resulting alignment map $\mathbf{M} \in \mathbb{R}^{256 \times H_l \times W_l}$ has the same spatial and channel dimensions as $\mathbf{F}_{up}$. The parameters of $\phi_{align}(\cdot)$ are learned end-to-end together with the rest of the network via standard backpropagation; they are not frozen during training. This design allows the alignment module to adaptively learn spatially varying weights that guide high-level semantic features to better align with low-level structural cues.

The aligned high-level feature is obtained through channel-wise and spatially adaptive modulation:

$$\tilde{\mathbf{F}}_h = \mathbf{F}_{up} \odot \mathbf{M} \tag{6}$$

where $\odot$ denotes element-wise multiplication.

Compared with a single-channel spatial alignment map, which applies identical modulation across all channels, the proposed channel-wise alignment map provides both spatial and channel-adaptive recalibration. This is particularly important for high-level semantic features, where different channels encode distinct semantic responses and object parts. A single-channel map may be insufficient to capture such diversity, while the proposed design enables more fine-grained feature modulation guided by low-level structural cues.

Although the channel-wise design introduces slightly higher computational cost, the overhead is negligible in the overall network, while yielding consistent performance improvements. This demonstrates a favorable trade-off between accuracy and efficiency.

3.5 Low-Level Feature Projection and Decoder Fusion

Low-level features extracted from early backbone stages contain rich spatial detail but also include noise and semantically irrelevant responses. Direct fusion with high-level features may introduce redundancy and impair decoder performance. Therefore, we first project the low-level feature $\mathbf{F}_l$ into a compact representation:

$$\hat{\mathbf{F}}_l = \text{SimAM}\left(\text{ReLU}\left(\text{BN}\left(\phi_{proj}\left(\mathbf{F}_l\right)\right)\right)\right) \tag{7}$$

where $\phi_{proj}(\cdot)$ is a 1×1 convolution that expands the channel dimension from 24 to 48. This expansion is chosen as a balance between improving representational capacity and limiting computational overhead: doubling the channels provides sufficient feature richness to capture boundary and structural cues necessary for high-quality alignment and fusion, while maintaining low parameter and computational costs consistent with the lightweight design of BFANet.

After dynamic alignment, the aligned high-level feature $\tilde{\mathbf{F}}_h$ is concatenated with the refined low-level feature:

$$\mathbf{F}_{cat} = \left[\tilde{\mathbf{F}}_h; \hat{\mathbf{F}}_l\right] \in \mathbb{R}^{304 \times H_l \times W_l} \tag{8}$$

The concatenated feature is then processed by a decoder fusion block consisting of two successive 3×3 convolutional layers with batch normalization and ReLU activation. A SimAM module is applied after the second convolution to further enhance the fused representation:

$$\mathbf{F}_f = \text{SimAM}\left(\phi_{fuse}\left(\mathbf{F}_{cat}\right)\right) \tag{9}$$

where $\mathbf{F}_f \in \mathbb{R}^{256 \times H_l \times W_l}$ denotes the final decoder output.

This multi-stage fusion strategy strengthens the integration of semantic context and boundary detail, yielding a more discriminative feature space for pixel-wise classification.

3.6 Classification Head and Output

To improve generalization and prevent overfitting, a dropout layer with a rate of 0.1 is applied after the feature fusion stage in the decoder. The relatively low dropout rate is selected to balance regularization and information preservation, which is particularly important for lightweight models with limited capacity. The fused feature $\mathbf{F}_f$ is passed through a 1×1 convolutional classifier to generate class logits:

$$\mathbf{Y} = \phi_{cls}\left(\mathbf{F}_f\right) \in \mathbb{R}^{K \times H_l \times W_l} \tag{10}$$

where K is the number of semantic classes.

The logits are then bilinearly upsampled to the original image resolution:

$$\mathbf{P} = \text{Upsample}\left(\mathbf{Y}, H, W\right) \in \mathbb{R}^{K \times H \times W} \tag{11}$$

For each pixel location p, the predicted class label is obtained by:

$$\hat{c}\left(p\right) = \arg \max_{k \in \{1,\ldots,K\}} \mathrm{P}_k\left(p\right) \tag{12}$$

4 Experimental Setup and Dataset Evaluation

4.1 Datasets

To evaluate BFANet, we constructed a curated subset of PASCAL VOC 2012 consisting of 2913 images, focusing on challenging scenarios such as small or overlapping objects, fine-grained structures, and ambiguous boundaries. Instead of purely manual selection, the subset was generated using a semi-automated selection algorithm. First, boundary complexity scoring was performed, where the complexity score of each object instance was calculated as the ratio of contour length to object area. Higher scores indicate more structurally complex objects. Second, threshold filtering was applied, and only instances exceeding a predefined complexity threshold were included to emphasize boundary-sensitive and structurally intricate

objects. Third, category coverage was considered to ensure that representative boundary-sensitive categories (e.g., chair, bottle, and motorbike) were included, providing diverse and challenging evaluation scenarios. The resulting subset was divided into training, validation, and test sets in a 9:1:1 ratio with pixel-level annotations. Representative examples (Fig. 2a), statistical analyses (Fig. 2b), and t-SNE visualizations (Fig. 2c) further demonstrate the structural diversity and boundary complexity of the selected samples. Therefore, we emphasize that this subset serves as a focused evaluation set rather than a replacement for the complete PASCAL VOC benchmark. Evaluation on the full PASCAL VOC split and additional datasets would provide stronger evidence of generalization capability, which will be investigated in future work.

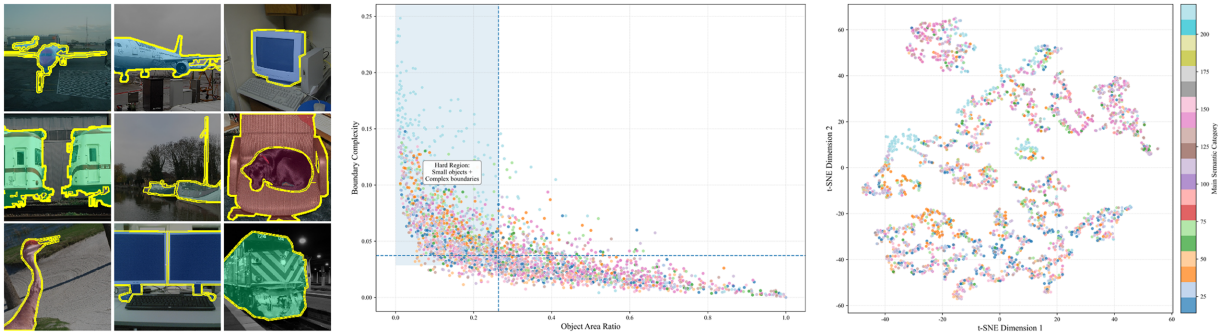

Figure 2: (**a**) Representative examples from the curated PASCAL VOC 2012 subset with overlaid segmentation masks, showing thin structures, complex foreground interactions, and ambiguous object boundaries. (**b**) Quadrant-based distribution of object area ratio vs. boundary complexity; most samples correspond to small objects with high boundary complexity. (**c**) t-SNE visualization of geometric and structural mask features, showing multiple localized clusters and confirming substantial structural diversity in the curated challenging samples.

4.2 Environment Configuration

All experiments, including model training, validation, and inference, were conducted on a unified platform to ensure fair comparison. The hardware consisted of an NVIDIA GeForce RTX 5060 Ti GPU with 16 GB memory, and the software environment included Windows 11, Python 3.12, and PyTorch, with CUDA and cuDNN used to accelerate training. Training was performed in two stages: a frozen backbone stage (epochs 0–15, batch size = 6), during which the MobileNetV2 backbone was frozen to stabilize initial training, followed by an unfrozen stage (epochs 16–120, batch size = 6) allowing the backbone and added modules to be fully optimized. Mixed-precision training (fp16 = True) was employed to reduce memory usage, and the MobileNetV2 backbone was initialized with pretrained weights. The optimizer was SGD with cosine learning rate decay, momentum of 0.9, weight decay of 1e−4, and validation was performed every 5 epochs. The input image size was set to [512, 512].

5 Results and Analysis

5.1 Comparison of Different Attention Mechanisms

As shown in Table 1, segmentation performance was influenced by the choice of attention mechanism. Among the compared modules—Triplet Attention (TA), Convolutional Block Attention Module (CBAM), and SimAM—SimAM achieved the highest precision at 95.14%. SimAM obtained an mIoU of 79.39%, outperforming TA (79.14%) and falling slightly below CBAM (79.56%). Although SimAM's mPA (88.01%) was lower than both TA (88.99%) and CBAM (88.30%), the difference remained within an acceptable

range. Overall, these results demonstrate that SimAM effectively refines feature responses and enhances the representation of informative regions while maintaining competitive segmentation accuracy.

Table 1: Performance comparison of different attention mechanisms.

Attention	mIoU (%)	mPA (%)	Precision (%)
TA	79.14	88.99	94.69
CBAM	79.56	88.30	94.96
SimAM	79.39	88.01	95.14

Classical attention mechanisms such as CBAM improve feature selectivity but introduce additional parameters and computational cost, which is particularly undesirable for lightweight models where efficiency is critical. In contrast, SimAM is completely parameter-free, employing an energy-based formulation to compute three-dimensional attention weights without adding any learnable parameters, while introducing only negligible computational overhead. This lightweight design not only preserves model compactness but also facilitates better gradient flow during backpropagation, leading to more stable and efficient training. Unlike CBAM which requires spatial and channel attention sub-modules with their associated convolutions, SimAM derives attention purely through spatial similarity calculations, making it exceptionally well-suited for resource-constrained deployment scenarios. The combination of zero-parameter overhead, efficient computation, and improved feature discriminability makes SimAM the optimal choice for lightweight semantic segmentation networks where every parameter and operation counts.

5.2 Ablation Experiments

From Table 2, we observe that incorporating DFA improves performance over the baseline, increasing mIoU from 77.57% to 79.15% and mDice from 87.52% to 88.11%. This demonstrates the effectiveness of dynamic feature alignment in mitigating semantic-spatial inconsistencies. Adding SimAM also improves mIoU to 79.39% and mDice to 87.89%, indicating its contribution to feature refinement and discriminability enhancement. When both proposed components are combined in BFANet, the full model achieves 80.23% mIoU and 88.78% mDice, showing that the modules are complementary and jointly improve segmentation performance. These improvements are obtained with only a small increase in model complexity, from 5.82M to 5.84M parameters and from 53.03 to 53.75 GFLOPs. In addition to GFLOPs and parameters, we measured the inference speed of BFANet on a single NVIDIA RTX 5060 Ti GPU (input size 512 × 512, batch size 1), achieving ~0.0161 s per image (~62 FPS).

Table 2: Ablation study of individual modules on PASCAL VOC 2012.

Model	mIoU (%)	mPA (%)	mDice	GFLOPs	Params
Baseline	77.57	86.36	87.52	53.03	5.82
DFA	79.15	88.76	88.11	53.75	5.84
SimAM	79.39	88.01	87.89	53.75	5.84
BFANet	80.23	88.05	88.78	53.75	5.84

Table 3 further investigates the impact of SimAM placement. Applying SimAM only after ASPP yields the highest mIoU (78.88%) and mDice (87.91%), indicating that high-level context enhancement is more

effective than applying SimAM solely to low-level features or decoder stages. These results confirm that careful placement of attention modules can maximize feature refinement without adding substantial complexity.

Table 3: Ablation study of SimAM placement variants.

Variant	mIoU (%)	mPA (%)	mDice
Baseline	77.57	86.36	87.52
SimAM-ASPP-only	78.88	87.73	87.91
SimAM-Low-only	78.73	87.89	87.51
SimAM-Decoder-only	78.30	87.11	87.32

5.3 Performance Comparison with Other Segmentation Models

As shown in Table 4, the proposed BFANet achieved better segmentation performance than the baseline DeepLabV3+. DeepLabV3+ obtained an mIoU of 77.57% and an mPA of 86.36%, with 5.82M parameters and 53.03 GFLOPs. In comparison, BFANet improved the mIoU to 80.23% and the mPA to 88.05%, corresponding to gains of 2.66 and 1.69 percentage points, respectively. Meanwhile, the number of parameters increased only slightly from 5.82M to 5.84M, and the computational cost increased marginally from 53.03 to 53.75 GFLOPs. We additionally report Boundary F1 and Boundary IoU. These metrics were computed from the raw predicted class-label masks and the corresponding ground-truth masks, rather than from colorized or blended visualization results. As reported in Table 4, BFANet achieved a Boundary F1 of 61.55% and a Boundary IoU of 51.42%, outperforming DeepLabV3+ by 4.27 and 2.68 percentage points, respectively. These results indicate that BFANet improves not only region-level segmentation accuracy but also boundary-level alignment between predicted object contours and ground-truth annotations. Overall, BFANet achieves a more favorable trade-off between segmentation accuracy, boundary quality, and computational complexity. To verify the stability and reliability of BFANet, three independent experimental runs were conducted with different random seeds (11, 42, and 123). The obtained mIoU values for the three runs are 80.29%, 80.18%, and 80.22%, respectively. Statistical analysis shows that the mean value and standard deviation of the results are 80.23 ± 0.05%. Overall, BFANet yields consistent performance improvements across the majority of categories in the dataset.

Table 4: Comparison with the DeepLabv3+ baseline using boundary metrics.

Model	mIoU (%)	mPA (%)	Boundary F1	Boundary IoU	Params	GFLOPs
DeepLabv3+	77.57	86.36	57.28	48.74	5.82	53.03
BFANet	80.23	88.05	61.55	51.42	5.84	53.75

As illustrated in Fig. 3, the proposed method produced more accurate and visually coherent segmentation results across multiple representative scenes. Compared with DeepLabv3+, the target regions were segmented more completely, and clearer object boundaries were obtained. A higher consistency with the ground-truth annotations was also observed. For objects with slender structures or complex shapes, such as the bottle, bicycle, duck, and train, incomplete regions, blurred contours, and background misclassification were still present in the results produced by DeepLabv3+. These challenging examples also serve as failure-detail cases, indicating that conventional segmentation methods may still suffer from boundary ambiguity and structural discontinuity when dealing with thin parts, irregular contours, or cluttered backgrounds. By contrast, the proposed method preserved the overall object shape more effectively and recovered finer

structural details. In addition, in scenes containing both foreground objects and complex backgrounds, a more accurate discrimination of the target regions was achieved, while irrelevant background responses were suppressed more effectively. These visual comparisons demonstrate that the proposed method was more capable of feature representation, boundary refinement, and detailed region modeling, thereby yielding superior qualitative segmentation performance. Meanwhile, the remaining minor boundary deviations in some complex regions suggest that segmentation of extremely slender or highly ambiguous structures remains a challenging issue for future improvement.

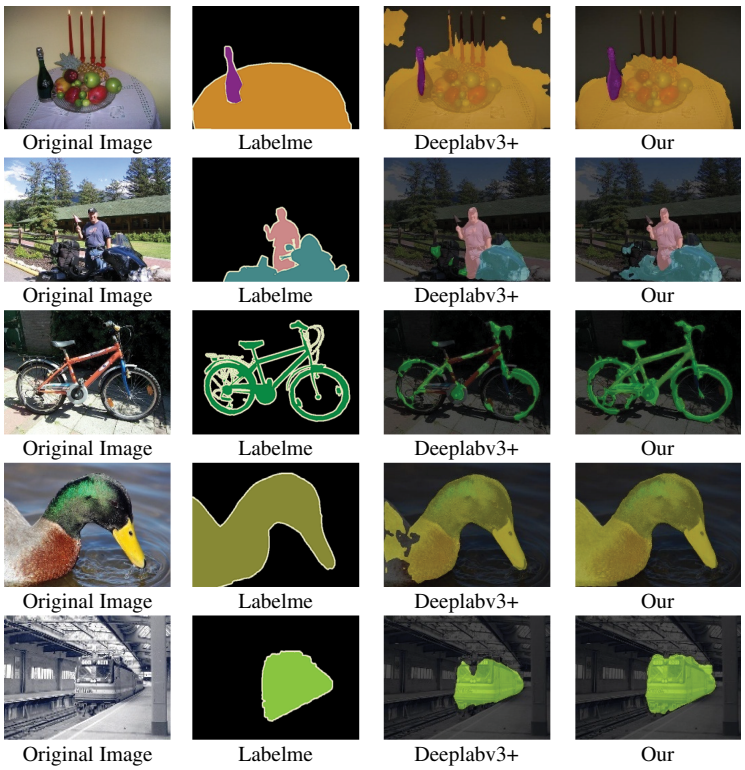

Figure 3: Qualitative comparison of the original image, ground truth, DeepLabv3+, and the proposed method.

Fig. 4 presents the per-class IoU comparison between the baseline DeepLabV3+ and the proposed BFANet on the curated PASCAL VOC 2012 subset used in this study. It should be noted that all results are obtained using the same subset rather than the full augmented PASCAL VOC 2012 benchmark. Therefore, the absolute performance values may differ from those reported in standard benchmark evaluations. However, all compared methods are trained and evaluated under identical experimental settings, ensuring a fair and consistent comparison. Overall, BFANet achieves consistent performance improvements across the majority of categories. Notably, substantial gains are observed for boundary-sensitive and structurally complex categories, including dining table (from 54.61% to 68.80%), bottle (from 74.96% to 81.38%), motorbike (from 72.58% to 79.74%), bird (from 79.24% to 84.80%), sofa (from 68.19% to 73.85%), and chair (from 42.60% to 46.61%). These results demonstrate the effectiveness of the proposed framework in enhancing segmentation accuracy for categories with intricate structures and strict boundary requirements. Such categories typically involve thin structures, irregular shapes, or complex part compositions, making them more susceptible to boundary ambiguity and background interference. The observed improvements indicate that BFANet is more effective in refining object boundaries and recovering fine-grained details, which is consistent with the design of the SimAM and Dynamic Feature Alignment (DFA) modules. In contrast, categories with relatively

large and regular shapes, such as bus, car, and sheep, show only marginal changes, suggesting that they are less sensitive to boundary-aware refinement. Overall, this analysis provides quantitative evidence that BFANet is particularly effective for boundary-sensitive and structurally complex categories under consistent evaluation settings.

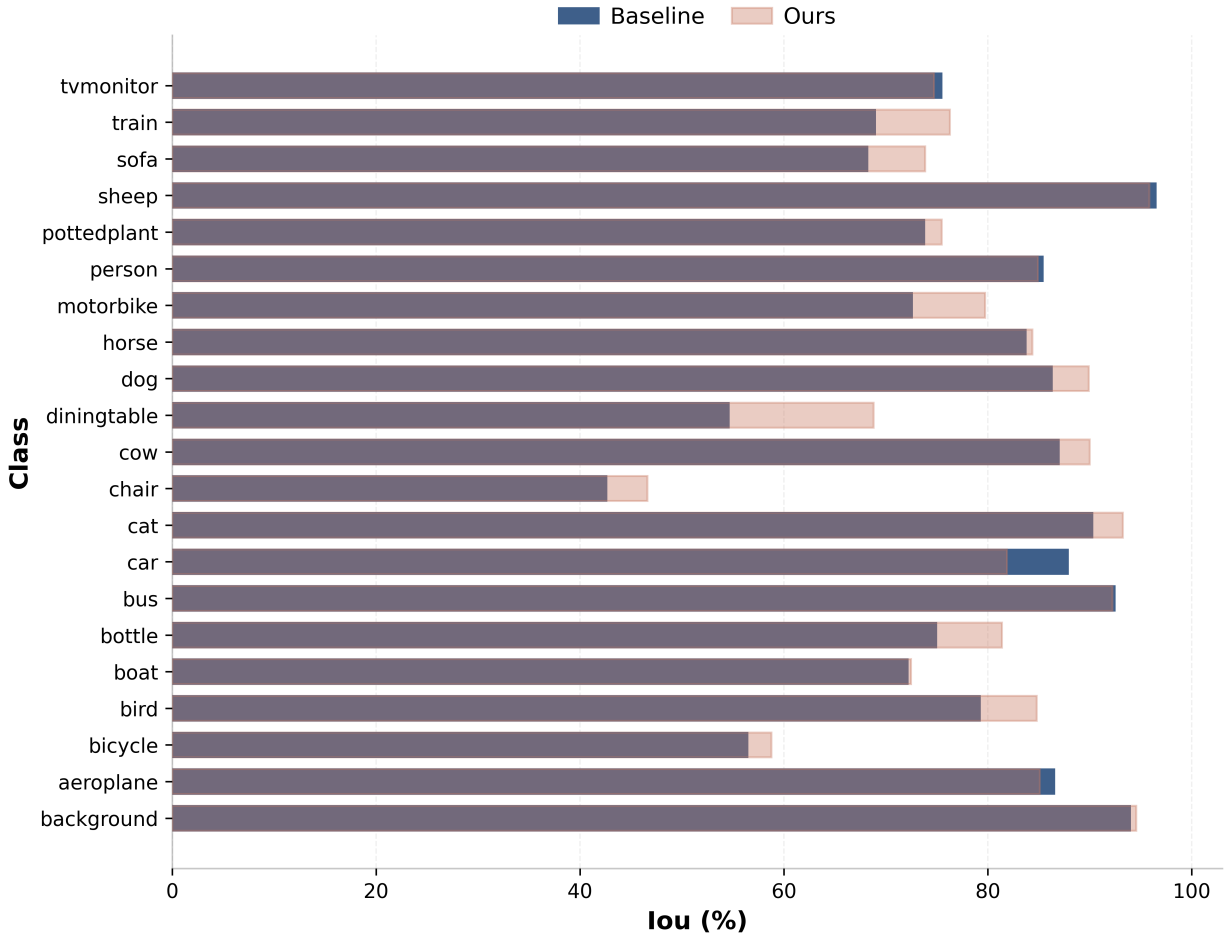

Figure 4: Per-class Iou comparison between the baseline model and the proposed method.

6 Conclusion

In this paper, BFANet, a boundary-aware lightweight semantic segmentation framework, was proposed to address semantic-spatial misalignment and insufficient feature discriminability in encoder-decoder pipelines. A parameter-free SimAM attention mechanism was integrated to enhance discriminative responses without adding computational overhead. A Dynamic Feature Alignment (DFA) module was designed to adaptively align high-level semantic features with low-level structural cues, reducing cross-level feature inconsistency. Furthermore, a multi-stage decoder fusion strategy was employed to progressively refine multi-scale context and boundary information. Extensive experiments on a curated subset of PASCAL VOC 2012 (2913 training images) demonstrated that BFANet outperforms the baseline DeepLabV3+ with MobileNetV2 backbone, improving the mIoU from 77.57% to 80.23% with only a marginal increase in computational cost (from 53.03 to 53.75 GFLOPs and 5.82M to 5.84M parameters). The results confirmed that explicit boundary-aware refinement and lightweight feature alignment are effective in capturing fine-grained details and preserving object boundaries. Overall, the proposed framework successfully combines

efficiency and accuracy in lightweight semantic segmentation. It provides a practical solution for edge devices and resource-constrained applications while maintaining high-quality boundary-aware predictions. Future work will explore further integration with transformer-based lightweight backbones and dynamic context modeling to enhance segmentation performance in complex scenes.

Acknowledgement: This work was supported by Qinghai University Postgraduate Research and Practice Innovation Program (Grant No. 2025-GMKY-42) and the Innovation and Entrepreneurship Training Program of Qinghai University (Grant No. 2025-DC80). The authors are grateful to all participants who contributed to this research.

Funding Statement: This research was funded by Qinghai University Postgraduate Research and Practice Innovation Program of Funder, grant number 2025-GMKY-42. This research was funded by the Innovation and Entrepreneurship Training Program of Qinghai University (2025-DC80).

Author Contributions: Conceptualization, Wang Zhang and Qiangqiang Yao; methodology, Wang Zhang and Jiayi Xing; software, Wang Zhang; validation, Lanlan Li and Qiangqiang Yao; formal analysis, Lanlan Li and Wang Zhang; resources, Wang Zhang; writing—original draft preparation, Wang Zhang and Lanlan Li; writing—review and editing, Qiangqiang Yao and Lanlan Li; funding acquisition, Wang Zhang and Qiangqiang Yao. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data available on request from the authors. The data that support the findings of this study are available from the Corresponding Author, Qiangqiang Yao, upon reasonable request.

Ethics Approval: Not applicable.

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

References

1. Yang H, Zhang D, Liu J, Cao Z, Wang N. LDMSNet: Lightweight dual-branch multi-scale network for real-time semantic segmentation of autonomous driving. Int J Automot Technol. 2025;26(2):577–91.
2. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 4510–20. doi:10.1109/cvpr.2018.00474.
3. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision; 2018 Sep 8–14; Munich, Germany. p. 833–51. doi:10.1007/978-3-030-01234-2_49.
4. Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision; 2018 Sep 8–14. Munich, Germany. p. 3–19. doi:10.1007/978-3-030-01234-2_1.
5. Yang L, Zhang RY, Li L, Xie X. A simple, parameter-free attention module for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning (ICML); 2021 Jul 18–24; Virtual. p. 11863–74. doi:10.48550/arXiv.2109.08608.
6. Yu C, Wang J, Peng C, Gao C, Yu G, Sang N. BiSeNet: bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European Conference on Computer Vision; 2018 Sep 8–14; Munich, Germany. p. 334–49. doi:10.1007/978-3-030-01261-8_20.
7. Yu C, Gao C, Wang J, Yu G, Shen C, Sang N. BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation. Int J Comput Vis. 2021;129(11):3051–68. doi:10.1007/s11263-021-01515-2.
8. Sheng C, Hu B, Meng F, Yin D. Lightweight dual-branch network for vehicle exhausts segmentation. Multimed Tools Appl. 2021;80(12):17785–806. doi:10.1007/s11042-021-10601-z.
9. Zhang W, Huang Z, Luo G, Chen T, Wang X, Liu W, et al. TopFormer: token pyramid transformer for mobile semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 12083–93.

10. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. Int J Comput Vis. 2010;88(2):303–38. doi:10.1007/s11263-009-0275-4.
11. Fan M, Lai S, Huang J, Wei X, Chai Z, Luo J, et al. Rethinking BiSeNet for real-time semantic segmentation. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 9711–20. doi:10.1109/cvpr46437.2021.00959.
12. Hu X, Tang C, Chen H, Li X, Li J, Zhang Z. Improving image segmentation with boundary patch refinement. Int J Comput Vis. 2022;130(11):2571–89. doi:10.1007/s11263-022-01662-0.
13. Huang S, Lu Z, Cheng R, FaPN H C. Feature-aligned pyramid network for dense image prediction. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 844–53. doi:10.1109/iccv48922.2021.00090.
14. Li X, You A, Zhu Z, Zhao H, Yang M, Yang K, et al. Semantic flow for fast and accurate scene parsing. In: Proceedings of the European Conference on Computer Vision; 2020 Aug 23–28; Virtual. p. 775–93. doi:10.1007/978-3-030-58452-8_45.
15. Zhou W, Yue Y, Fang M, Qian X, Yang R, Yu L. BCINet: Bilateral cross-modal interaction network for indoor scene understanding in RGB-D images. Inf Fusion. 2023;94:32–42. doi:10.1016/j.inffus.2023.01.016.
16. Zhou W, Zhang H, Yan W, Lin W. MMSMCNet: modal memory sharing and morphological complementary networks for RGB-T urban scene semantic segmentation. IEEE Trans Circuits Syst Video Technol. 2023;33(12):7096–108. doi:10.1109/TCSVT.2023.3275314.
17. Zhou W, Dong S, Xu C, Qian Y. Edge-aware guidance fusion network for RGB-thermal scene parsing. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI); 2022 Feb 22–Mar 1; Virtual. p. 3571–9. doi:10.1609/aaai.v36i3.20269.
18. Zhou W, Yang E, Lei J, Wan J, Yu L. PGDENet: progressive guided fusion and depth enhancement network for RGB-D indoor scene parsing. IEEE Trans Multimed. 2023;25:3483–94. doi:10.1109/TMM.2022.3161852.
19. Howard A, Sandler M, Chen B, Wang W, Chen LC, Tan M, et al. Searching for MobileNetV3. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–No 2; Seoul, Republic of Korea. p. 1314–24. doi:10.1109/iccv.2019.00140.
20. Xu J, Xiong Z, Bhattacharyya SP. PIDNet: A real-time semantic segmentation network inspired by PID controllers. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 19529–39. doi:10.1109/cvpr52729.2023.01871.
21. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and efficient design for semantic segmentation with transformers. In: Advances in neural information processing systems; 2021 Dec 6–14; Virtual. p. 12077–90.
22. Zhao H, Qi X, Shen X, Shi J, Jia J. ICNet for real-time semantic segmentation on high-resolution images. In: Proceedings of the European Conference on Computer Vision; 2018 Sep 8–14; Munich, Germany. p. 418–34. doi:10.1007/978-3-030-01219-9_25.
23. Romera E, Álvarez JM, Bergasa LM, Arroyo R. ERFNet: efficient residual factorized ConvNet for real-time semantic segmentation. IEEE Trans Intell Transp Syst. 2018;19(1):263–72. doi:10.1109/TITS.2017.2750080.
24. Mehta S, Rastegari M, Caspi A, Shapiro L, Hajishirzi H. ESPNet: efficient spatial pyramid of dilated convolutions for semantic segmentation. In: Proceedings of the European Conference on Computer Vision; 2018 Sep 8–14; Munich, Germany. p. 561–80. doi:10.1007/978-3-030-01249-6_34.
25. Wu T, Tang S, Zhang R, Cao J, Zhang Y. CGNet: a light-weight context guided network for semantic segmentation. IEEE Trans Image Process. 2021;30:1169–79. doi:10.1109/TIP.2020.3042065.
26. Wang J, Gou C, Wu Q, Feng H, Han J, Ding E, et al. RTFormer: efficient design for real-time semantic segmentation with transformer. In: Advances in Neural Information Processing Systems; 2022 Nov 28–Dec 9; New Orleans, LA, USA. p. 7423–36. doi:10.52202/068431-0539.
27. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, et al. Dual attention network for scene segmentation. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 3141–9. doi:10.1109/cvpr.2019.00326.

28. Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W. CCNet: criss-cross attention for semantic segmentation. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 603–12. doi:10.1109/iccv.2019.00069.
29. Yuan Y, Chen X, Wang J. Object-contextual representations for semantic segmentation. In: Proceedings of the European Conference on Computer Vision; 2020 Aug 23–28; Virtual. p. 173–90. doi:10.1007/978-3-030-58539-6_11.
30. Feng J, Zheng W, Gu Z, Guo D, Qin R. A position-aware attention network with progressive detailing for land use semantic segmentation of remote sensing images. Int J Remote Sens. 2023;44(21):6762–801. doi:10.1080/01431161.2023.2274820.
31. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 11531–9. doi:10.1109/cvpr42600.2020.01155.
32. Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 13708–17. doi:10.1109/cvpr46437.2021.01350.
33. Xiao X, Zhao Y, Zhang F, Luo B, Yu L, Chen B, et al. BASeg: boundary aware semantic segmentation for autonomous driving. Neural Netw. 2023;157(12):460–70. doi:10.1016/j.neunet.2022.10.034.
34. Yi J, Zhong X, Liu W, Wu Z, Deng Y. Polar edge distance loss in edge-aware plug-and-play scheme for semantic segmentation. Eng Appl Artif Intell. 2025;154(4):110810. doi:10.1016/j.engappai.2025.110810.
35. Strudel R, Garcia R, Laptev I, Schmid C. Segmenter: transformer for semantic segmentation. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 7242–52. doi:10.1109/iccv48922.2021.00717.
36. Cheng B, Schwing AG, Kirillov A. Per-pixel classification is not all you need for semantic segmentation. In: Advances in Neural Information Processing Systems; 2021 Dec 6–14; Virtual. p. 17864–75. doi:10.5555/3540261.3541628.
37. Yu Q, Wang H, Qiao S, Collins M, Zhu Y, Adam H, et al. kMaX-DeepLab: k-means mask transformer. In: Proceedings of the European Conference on Computer Vision (ECCV); 2022 Oct 23–27; Tel Aviv, Israel. p. 288–307. doi:10.1007/978-3-031-19818-2_17.
38. Ruan S, Wan Q, Chen R, Hu M, Guo X, Song K. Context-aware feature enhancement network for remote sensing image semantic segmentation. Remote Sens. 2026;18(4):543. doi:10.3390/rs18040543.
39. Poudel RPK, Liwicki S, Cipolla R. Fast-SCNN: fast semantic segmentation network. In: Proceedings of the British Machine Vision Conference (BMVC); 2019 Sep 9–12; Cardiff, UK.
40. Poudel RPK, Bonde U, Liwicki S, Zach C. ContextNet: exploring context and detail for semantic segmentation in real-time. In: Proceedings of the British Machine Vision Conference (BMVC); 2018 Sep 3–6; Newcastle, UK.