2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Local-Semantic Attentive Bidirectional Bottleneck Network with Residual Feature Augmentation for Real-time Semantic Segmentation

ManYuan Gui[1], Jinlai Zhang[2(✉)], Yonghen Hu[1], Sheng Wu[4], Du Xu[2], Bo Ouyang[3], Shaosheng Fan[4], Zhenzhen Jin[5]

[1] School of Computer Science and Technology, Changsha University of Science and Technology, Changsha, 410114, Hunan, China
[2] College of Mechanical and Vehicle Engineering, Changsha University of Science and Technology, Changsha, 410114, Hunan, China
[3] Department of Computer Science and Technology, Tsinghua University, Beijing, 100000, Beijing, China
[4] School of Artificial Intelligence, Changsha University of Science and Technology, Tianxin District, Changsha, Hunan, 410114, China
[5]College of Mechanical Engineering, Guangxi University, Naning, 530004, Guangxi, China

Abstract. Real-time semantic segmentation is critical for applications such as autonomous driving, where the core challenge lies in achieving high segmentation accuracy while maintaining efficient inference. This paper proposes LSA-BiNet, a bidirectional bottleneck network via local-semantic attention and residual feature augmentation. The framework has three key innovations: (1) The Local Receptive Field Attention (LRFA) module achieves high-order feature interactions with 1st-order computational complexity through region-wise soft-weight computation and channel gating; (2) The Spatial Variance Fusion Module (SVFM) collaboratively models local and non-local features via low-frequency variance modulation and local detail enhancement; (3) The Residual Cross-level Attention Decoder (RCAD) enables precise pixel-level prediction using cross-level feature projection, dual gating mechanisms, and residual attention weighting. Extensive experiments on Cityscapes and CamVid benchmarks demonstrate that LSA-BiNet achieves state-of-the-art (SOTA) mean Intersection-over-Union (mIoU) of 72.74% and 68.53% without ImageNet pretraining, while maintaining low computational complexity (8.81 GFLOPs) and real-time inference speeds (51.08 FPS on Cityscapes, 79.62 FPS on CamVid). Ablation studies confirm significant contributions of each module, establishing LSA-BiNet's superiority over contemporary SOTA models.

Keywords: Real-time semantic segmentation, local-semantic attention, residual feature augmentation, bidirectional bottleneck network, computational efficiency.

1 Introduction

Real-time semantic segmentation plays an essential role in numerous applications, particularly in fields such as autonomous driving, robotics, and Intelligent Transportation Systems (ITS). Semantic segmentation involves classifying each pixel of an image into predefined categories, enabling the system to understand and interpret the scene. However, achieving high segmentation accuracy in real-time is a difficult challenge, as it requires balancing the trade-off between computational efficiency and performance quality [3,10]. While several solutions have been proposed in the literature, most of them either prioritize accuracy at the expense of speed or attempt to optimize for speed while sacrificing accuracy. The pursuit of an effective solution remains particularly important for dynamic, time-sensitive environments such as autonomous vehicles, where real-time decisions must be made based on precise visual data.

Recent advancements in deep learning, particularly in the field of convolutional neural networks (CNNs) [5, 8, 21, 29], have brought remarkable improvements in semantic segmentation performance. Convolutional architectures such as U-Net [20], DeepLab [4], and FCN [14] have set the foundation for modern approaches. The introduction of encoder-decoder architectures and dilated convolutions has helped to expand the receptive field and capture global context, while preserving spatial resolution for pixel-wise accuracy. Despite these advancements, achieving a model that performs well in real-time while still maintaining accuracy across various challenging datasets remains a formidable problem. Particularly, for applications like autonomous driving, where every millisecond of decision-making counts, it is imperative to develop models that not only deliver high performance but also run with minimal latency. This balance is where traditional models often fall short, leading to the need for more innovative solutions.

One of the key limitations of current semantic segmentation models is their inability to effectively model both local and global contextual information [29]. Local information refers to pixel-level features that represent fine-grained details, while global context captures the broader scene structure and long-range dependencies. Many SOTA models rely on either a global context model or a local feature extractor, without fully addressing the need for both types of information to be seamlessly integrated. The issue of aggregating such features efficiently, while avoiding excessive computational complexity, is an area where further progress is needed. In this context, attention mechanisms have been a powerful tool for enhancing the feature extraction process [7, 9]. Attention allows the model to focus on the most relevant parts of the input, thus improving both the segmentation accuracy and computational efficiency. However, designing an attention mechanism that can simultaneously capture both local and global contextual information remains a challenging task. The lack of an effective and efficient attention mechanism that caters to both local and non-local features has been a major bottleneck in semantic segmentation models.

In addition to the attention mechanism, feature aggregation and decoding strategies are crucial to enhancing the model's performance [12]. Traditional approaches often struggle to adequately combine multi-level features [30-33] from the encoder and decoder [7]. This is particularly true for real-time applications, where the combination of high-level features and low-level features must be done in a way that avoids computational overload. Feature fusion techniques that can balance the contributions of various feature levels are essential for improving segmentation accuracy without incurring a significant increase in computational cost. Furthermore, many SOTA models rely on pre-trained networks (such as ImageNet) to initialize their weights [17], which can lead to issues when dealing with specialized datasets or applications with limited labeled data. The dependency on ImageNet pretraining limits the model's adaptability to different domains, as the features learned from one domain may not generalize well to others [12].

To address these challenges, we propose the Local-Semantic Attentive Bidirectional Bottleneck Network (LSA-BiNet), a novel real-time semantic segmentation model that introduces three key innovations to improve segmentation performance while maintaining computational efficiency. LSA-BiNet is designed with a bidirectional bottleneck architecture, employing local-semantic attention mechanisms and residual feature augmentation to enhance both local and global feature interactions. Our model aims to provide a solution that addresses the limitations of existing methods by efficiently integrating local and non-local features while ensuring real-time performance. Specifically, the LRFA module facilitates high-order feature interactions with first-order computational complexity. It does this by computing soft-weight regions and employing channel gating mechanisms. This enables LSA-BiNet to capture fine-grained local information while remaining computationally efficient. The key idea is to enhance the model's ability to focus on the most relevant features in the input image, allowing it to achieve higher segmentation accuracy without a significant increase in computational cost. The SVFM provides a novel way to model both local and non-local features. It does so by employing low-frequency variance modulation and local detail enhancement. This collaborative feature aggregation helps to strengthen the representation of both types of features, allowing LSA-BiNet to effectively combine local and global context for better pixel-level predictions. The low-frequency modulation is designed to capture long-range dependencies, while the local detail enhancement ensures fine-grained accuracy [27]. The RCAD is a key component that enables precise pixel-level predictions by using cross-level feature projection, dual gating mechanisms, and residual attention weighting. This decoder improves the segmentation output by fusing features from multiple levels and attention gates, allowing the model to refine the final output while maintaining a low computational footprint. The use of residual attention mechanisms helps to ensure that the model remains effective even when trained with limited resources or specialized datasets.

The paper is organized as follows: Section 2 reviews the classic semantic segmentation architecture and related technologies for real-time semantic segmentation. Section 3 proposes our new semantic segmentation model and its related methods. Section 4 verifies the effectiveness of the method through experimental comparison and ablation study. Finally, in Section 5, we summarize the research and discuss future directions.

2 Methodology

The primary objective of this paper is to achieve high-precision image semantic segmentation while maintaining efficient inference. Given an input image, the foremost goal is to extract semantic information through a multi-scale feature pyramid, concurrently implement Local Receptive Field Attention (LRFA) modulation and Spatial Variance Fusion Module (SVFM) within the network, and ultimately generate pixel-level predictions via a Residual Cross-level Attention Decoder (RCAD).

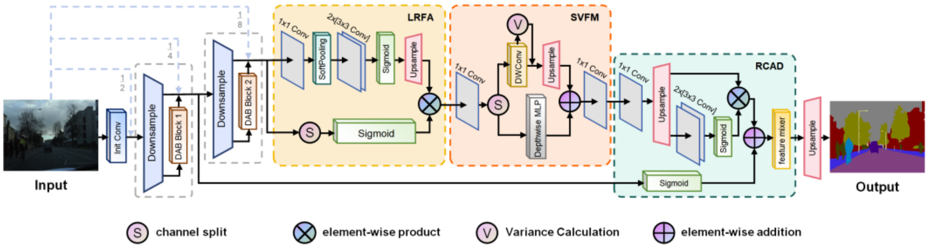

Fig. 1. The overall workflow of LSA-BiNet. After initial convolutions and multi-scale feature extraction, the input is processed through encoder stages containing: Local Receptive Field Attention (LRFA) module generating regional importance heatmaps via SoftPooling and modulating features using first-channel gating; Spatial Variance Fusion Module (SVFM) employing dual-branch architecture (EASA branch for non-local dependencies, LDE branch for local details) for feature fusion; Residual Cross-level Attention Decoder (RCAD) module fusing cross-level features through dual gating mechanisms and residual calibration to produce pixel-wise predictions.

As depicted in Fig. 1, the proposed LSA-BiNet first extracts multi-scale features through initial convolutional layers and downsampling operations. Subsequently, these features are fed into an encoder composed of multiple DAB modules (Depthwise Asymmetric Bottleneck) to capture rich contextual information. During the encoding process, we introduce the Local Receptive Field Attention (LRFA) module and Spatial Variance Fusion Module (SVFM) to enable efficient high-order feature interactions and cooperatively leverage both local and non-local feature interactions, respectively. Finally, the Residual Cross-level Attention Decoder (RCAD) generates the segmentation results by fusing shallow detail features with deep semantic features, while utilizing a residual attention mechanism for feature calibration.

2.1 Local Receptive Field Attention

The Local Receptive Field Attention (LRFA) is an efficient attention mechanism designed to achieve high-order feature interactions through a local importance map and channel gating mechanism, while maintaining low computational complexity. As illustrated in Fig. 2, this module is constructed via a five-step processing flow: first, channel compression is performed on the input features using 1×1 convolution; then, regional importance aggregation is implemented using a 7×7 SoftPooling operation with stride

3; next, spatial downsampling is processed through 3×3 convolution with stride 2; then, feature transformation is applied via another 3×3 convolution; finally, a normalized importance heatmap is generated via Sigmoid activation followed by bilinear upsampling to the input resolution.

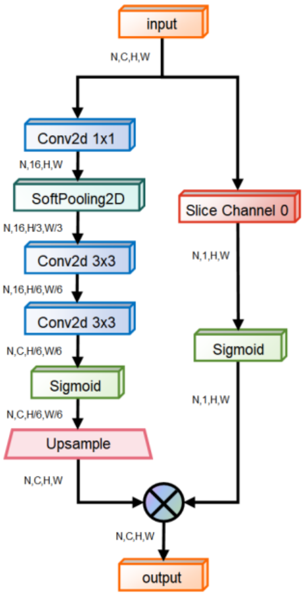

Fig. 2. The structure of LRFA module. The proposed LRFA generates regional heatmaps via SoftPooling and modulates features using first-channel gating.

The core formulation of the SoftPooling operation is defined as:

$$\mathcal{P}(X)|_x = \frac{\sum_{i\in R} e^{x_i}\cdot x_i}{\sum_{j\in R} e^{x_j}} \tag{1}$$

where denotes the local pooling window centered at position x. This operation computes a weighted average where the numerator contains the regional sum of products between feature values and exponentially weighted features, while the denominator represents the regional sum of exponential features. This differs fundamentally from the original description by: (1) omitting learnable weights since no convolutional kernels are used in the pooling implementation, (2) taking the product of and rather than just e^{x_i}, and (3) performing averaging within each region rather than weighted summation.

The final attention output incorporates a dual gating mechanism:

$$\text{Output}=X \odot W \odot G \tag{2}$$

where is the upsampled importance map, is the channel gate computed by applying sigmoid to the first channel of input features X, and denotes element-wise multiplication. This channel gating mechanism, implemented via the gate path, was absent in the original description.

Unlike mainstream attention mechanisms, LRFA innovatively employs the first channel of input features as a dynamic gating signal. After activation through a sigmoid function, this signal is element-wise multiplied with an upsampled importance map restored to the original dimensions, ultimately achieving adaptive calibration of the original features. This module design maintains 1st-order attention latency while achieving feature interaction capabilities approaching 2nd-order attention mechanisms.

2.2 Spatial Variance Fusion Module

The Spatial Variance Fusion Module (SVFM) implements a dual-branch architecture that synergistically integrates local detail enhancement and global structural modeling, addressing the low-pass filtering limitations of conventional transformers that cause over-smooth reconstructions. As illustrated in Fig. 3, The computational workflow begins with feature decomposition:

$$\{X, Y\} = \mathrm{split}\left(\mathcal{C}_{1\times1}^{(2C)}(f)\right) \tag{3}$$

where denotes a convolution doubling channel dimensionality, partitioning input into structural component and detail component for specialized branch processing.

The EASA branch captures long-range dependencies through variance modulation and adaptive pooling. First, spatial reduction extracts low-frequency structural cues:

$$X_s = \mathcal{DW}_{3\times3}\left(\mathrm{AdaptiveMaxPool}_{1/8}(X)\right) \tag{4}$$

where applies depthwise convolution to 8× downsampled features. Global statistical descriptors are then computed via spatial variance:

$$X_v = \mathrm{Var}\big(X, \mathrm{dim} = (-2, -1)\big) \tag{5}$$

The modulation mechanism dynamically balances these components:

$$X_l = X \odot \mathrm{Upsample}\left(\mathrm{GELU}\left(\mathcal{C}_{1\times1}^{(C)}(\alpha \cdot X_s + \beta \cdot X_v)\right)\right) \tag{6}$$

where learnable parameters weight low-frequency structures (X_s) and global statistics (X_v), with GELU activation enabling nonlinear interaction, *followed by upsampling to the original spatial resolution*, before feature recalibration via Hadamard product.

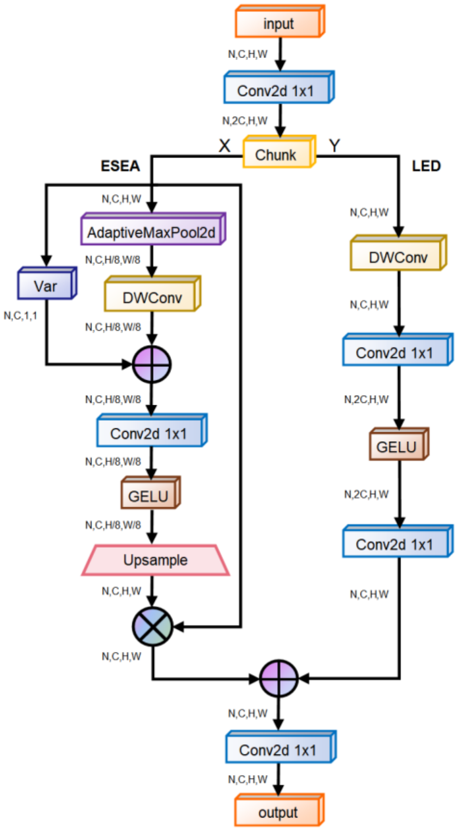

Fig. 3. The structure of SVFM. It employs dual-branch architecture (non-local dependency modeling and local detail enhancement) for feature fusion through variance modulation and adaptive pooling.

Concurrently, the LDE branch processes detail component using bottleneck transformations:

$$Y_d = \mathcal{C}_{1\times1}^{(C)}\left(\mathrm{GELU}\left(\mathcal{C}_{1\times1}^{(2C)}\left(\mathcal{DW}_{3\times3}(Y)\right)\right)\right) \tag{7}$$

Here extracts fine-grained local patterns, while consecutive convolutions with 2× channel expansion/reduction and GELU activation form an inverted bottleneck structure that amplifies high-frequency components.

Finally, feature aggregation combines branch outputs:

$$F_p = \mathcal{C}_{1\times1}^{(C)}(X_l + Y_d) \tag{8}$$

This additive fusion leverages structural priors from EASA and detail features from LDE, with the convolution projecting aggregated features to original channel dimensionality to complete the adaptive synthesis process.

In the LSA-BiNet architecture, the SVFM employs a dual-branch cooperative mechanism. For its EASA branch, it adopts downsampling and variance modulation to replace standard self-attention, significantly reducing computational complexity while effectively modeling long-range dependencies. Concurrently, the LDE branch enhances high-frequency details through a lightweight convolutional structure.

2.3 Residual Cross-level Attention Decoder

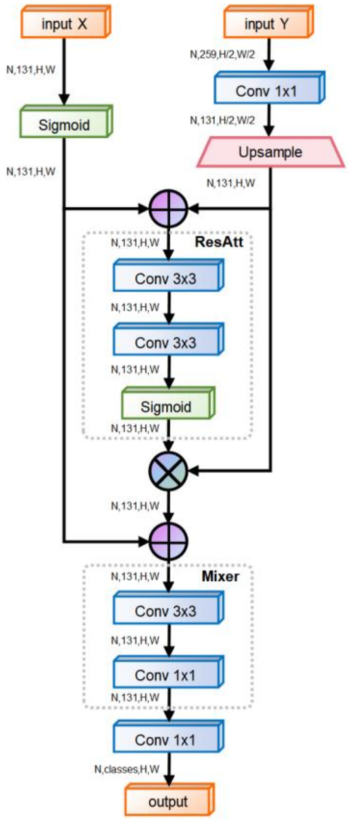

Fig. 4. The structure of RCAD module. It fuses cross-level features via dual gating mechanisms with residual calibration.

The Residual Cross-level Attention Decoder (RCAD) module serves as the core decoding component in the LSA-BiNet architecture, designed to enhance semantic segmentation accuracy through cross-level feature fusion and a residual attention mechanism. This module innovatively integrates complementary information from low-resolution deep features and high-resolution shallow features, adaptively reinforcing critical features while suppressing noise via an attention gating mechanism, as illustrated in Fig. 4.The computational workflow of RCAD can be decomposed into four key steps:

$$Y_c = \mathrm{Conv}_{1\times 1}(Y) \tag{9}$$

The channel projection operation resolves the dimensional mismatch between deep and shallow features by compressing the channel dimension of deep feature from to C_1. This transformation enables direct fusion of semantic information from the deep pathway with spatial details from the shallow pathway through a simple convolution layer, establishing the foundational alignment for cross-level attention.

$$\mathrm{Att}_{\mathrm{map}} = \sigma(\sigma(X) + Y_c) \tag{10}$$

Dual gating mechanism generates the attention heatmap through sequential sigmoid activations. The first activation creates a spatial attention prior from shallow features, emphasizing locally salient regions. The subsequent activation after element-wise addition with projected deep features refines the attention weights to capture cross-level feature correlations, resulting in a final attention map that optimally weights relevant regions across both feature hierarchies.

$$\mathrm{Output}_{\mathrm{att}} = X \odot \mathrm{Att}_{\mathrm{map}} + \mathrm{ResAtt}\left(X \odot \mathrm{Att}_{\mathrm{map}}\right) \odot Y_c \tag{11}$$

Residual attention weighting performs two-stage feature refinement. Primary fusion applies attention filtering to shallow features to suppress noise. Residual enhancement processes this filtered representation through a residual block (comprising two convolutional layers), then modulates the deep features with the refined attention signals. The element-wise sum combines the spatially-preserved shallow features with semantically-enriched deep features, maintaining feature integrity through residual connections.

$$\mathrm{Pred}{=}\mathrm{Conv}_{1\times 1}\left(\mathrm{Mixer}\left(\mathrm{Output}_{\mathrm{att}}\right)\right) \tag{12}$$

Final classification involves feature mixing and projection. The module employs depthwise separable convolution to efficiently integrate channel-wise relationships while preserving spatial structures. The convolution layer then projects the refined features to prediction space, generating the segmentation mask while maintaining computational efficiency.

3 Experiments

In this section, we comprehensively evaluate the performance of the LSA-BiNet network based on two benchmark datasets widely used for semantic segmentation in urban scenes: Cityscapes and CamVid. We conducted a series of ablation studies on the CamVid test set to deeply analyze the contribution of each module. Finally, we compare the LSA-BiNet network with SOTA real-time semantic segmentation models.

3.1 Datasets

Cityscapes [6] is a benchmark dataset focused on semantic segmentation of urban street scenes. It provides 5,000 high-resolution (2048×1024) finely annotated street-view images, including 2,975 training images, 500 validation images, and 1,525 test images (with non-public annotations). The dataset covers 30 semantic categories, with 19 common categories used for model evaluation. These images are captured from real urban environments, featuring diverse street scenes and complex elements.

CamVid [2] is another widely used dataset for semantic segmentation in autonomous driving research, released by the University of Cambridge Engineering Department. It contains 701 high-resolution images (960×720 pixels), divided into 367 training, 101 validation, and 233 test sets. All images are extracted from continuous driving videos and annotated frame-by-frame. The dataset covers 32 semantic categories, though 11 categories are typically used in experimental evaluations.

3.2 Implementation Protocol

The experimental training was conducted on four Tesla P40 GPUs using the PyTorch 1.12.0 framework, CUDA 11.3, and cuDNN v8. During training, multi-GPU data parallelism was employed, while evaluation was performed on a single GPU. For the Cityscapes dataset, a batch size of 8 was utilized with the Stochastic Gradient Descent (SGD) optimizer configured with a momentum of 0.9, weight decay of 1×10^{-4}, and an initial learning rate of 4.5×10^{-2}. For the CamVid dataset, a batch size of 16 was applied alongside the Adam optimizer with parameters $\beta = (0.9,0.999)$, weight decay of 2×10^{-4}, and an initial learning rate of 1×10^{-3}. The learning rate policy adopted polynomial decay with linear warmup. Training proceeded for a maximum of 1,000 epochs without pretrained weights. Data augmentation included random horizontal mirroring and random scaling (scale range: 0.5 to 2.0). Input resolutions were set to 512×1024 pixels for Cityscapes and 360×480 pixels for CamVid.

3.3 Experimental Performance Metrics

The primary performance metric employed in this experiment is the mean Intersection over Union (mIoU), which serves as a standard evaluation measure for semantic segmentation tasks. This metric effectively quantifies the spatial overlap between predicted segmentation masks and their corresponding ground truth labels. The calculation process involves a multi-step approach that ensures a comprehensive assessment of segmentation quality.

A critical foundation for mIoU calculation is the construction of a confusion matrix with dimensions $C \times C$, where represents the total number of semantic classes. Each element in this matrix records the count of pixels where the ground truth belongs to class while being predicted as class j. Pixels labeled with the special ignore_label value (default 255) are systematically excluded from this computation to maintain metric integrity.

The core calculation involves determining the Intersection over Union (IoU) for each individual class i. This class-specific metric is derived using the formula:

$$\mathrm{IoU}_i = \frac{M_{ii}}{\sum_{j=1}^{C} M_{ij} + \sum_{j=1}^{C} M_{ji} - M_{ii}} \tag{13}$$

where represents the true positives (correctly classified pixels for class i), corresponds to the total ground truth pixels for class (sum of row i), and indicates the total predicted pixels for class (sum of column i). The denominator effectively computes the union of ground truth and predicted pixels for class by combining the total occurrences in both sets while subtracting the double-counted intersection.

The final mean IoU (mIoU) is computed as the arithmetic mean of all class-specific IoU values:

$$\mathrm{mIoU}=\frac{1}{C}\sum_{i=1}^{C} \mathrm{IoU}_i \tag{14}$$

To prevent skewed results from classes with no true positives, any class where is automatically excluded from this average calculation. This exclusion ensures the metric accurately reflects segmentation performance on meaningful, detectable classes.

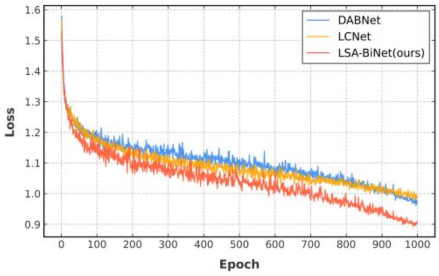

Fig. 5. Comparison of loss changes during training of different models on the Cityscapes dataset.

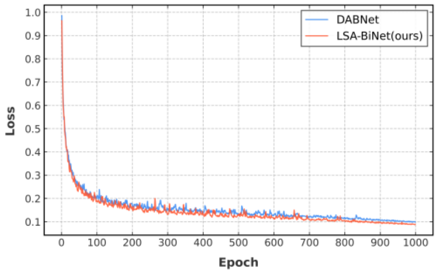

Fig. 6. Comparison of loss changes during training of different models on the CamVid dataset.

3.4 Main Results

As shown in Table 1, LSA-BiNet delivers superior segmentation accuracy on the Cityscapes dataset compared to other lightweight networks, achieving a leading mIoU of 72.74% on the test set. This performance significantly exceeds that of prior lightweight models such as LEDNet (69.2%), ESNet (70.7%), and LCNet (67.40%). Although LSA-BiNet has 1.84M parameters—only marginally more than minimal models like ENet—it remains highly efficient, requiring just 8.81G FLOPs, which is substantially lower than the FLOPs of models with inferior accuracy, for example, LEDNet: 23.0G;

LCNet: 15.9G. Furthermore, an inference speed of 51.08 fps demonstrates that LSA-BiNet effectively balances accuracy and real-time performance, and notably, this high performance is achieved without any ImageNet pertaining.

Table 1. Performance Comparison of LSA-BiNet Against Existing Semantic Segmentation Networks Estimated on the Cityscapes Dataset.

Method	Source	Pretrain	Input Size	mIoU (%) ↑		Params (M) ↓	Speed (fps) ↑	FLOPs (G) ↓
				val.	**test**			
SegNet [1]	TRAMI2017	ImageNet	360×640	57.8	56.1	29.5	74.9	652.5
ENet [18]	ICCV2015	No	360×640	59.0	58.3	**0.36**	83.8	8.7
SQNet [23]	NeurIPS2016	ImageNet	1024×2048	59.9	59.8	16.3	18.7	288.2
ESPNet [15]	ECCV2016	No	512×1024	60.0	60.3	**0.36**	**292.0**	**6.9**
ESPNet V2 [16]	CVPR2019	No	512×1024	66.2	66.4	1.3	134.8	7.4
CGNet [26]	TIP2016	No	1024×2048	63.5	64.8	0.49	53.0	14.0
ICNet [28]	ECCV2018	ImageNet	1024×2048	-	7.8	7.8	24.4	14.2
EDANet [13]	MMAsia2019	No	512×1024	68.1	67.3	0.68	161.0	17.9
LEDNet [24]	ICIP2019	No	512×1024	70.6	69.2	0.95	86.5	23.0
DABNet [11]	BMVC2019	No	1024×2048	69.34	69.66	0.76	43.31	20.9
ESNet [25]	PVCR2019	No	512×1024	70.4	70.7	1.66	112.3	48.7
LCNet [22]	TITS2024	No	512×1024	67.05	67.40	0.51	46.62	15.9
LSA-BiNet(ours)	This work	No	512×1024	**72.65**	**72.74**	1.84	51.08	8.81

Table 2. Per-Class Results of Different Segmentation Models on the Cityscapes Test Set.

Method	Roa	Sid	Bui	Wal	Fen	Pol	TLi	TSi	Veg	Ter	Sky	Ped	Rid	Car	Tru	Bus	Tra	Mot	Bic	class
ENet [18] (ICCV2015)	96.3	74.2	75.0	32.2	33.2	43.4	34.1	44.0	88.6	61.4	90.6	65.5	38.4	90.6	36.9	50.5	48.1	38.8	55.4	96.3
SQNet [23] (NeurIPS2016)	96.9	75.4	87.9	31.6	35.7	50.9	52.0	61.7	90.9	65.8	93.0	73.8	42.6	91.5	18.8	41.2	33.3	34.0	59.9	96.9
ESPNet [15] (ECCV2016)	97.0	77.5	76.2	35.0	36.1	45.0	35.6	46.3	90.8	63.2	92.6	67.0	40.9	92.3	38.1	52.5	50.1	41.8	57.2	97.0
CGNet [26] (TIP2016)	95.5	78.7	88.1	40.0	43.0	54.1	59.8	63.9	89.6	67.6	92.9	74.9	54.9	90.2	44.1	59.5	25.2	47.3	60.2	95.5
EDANet [13] (MMAsia2019)	97.8	80.6	89.5	42.0	46.0	52.3	59.8	65.0	91.4	68.7	93.6	75.7	54.3	92.4	40.9	58.7	56.0	50.2	64.0	97.8
ERFNet [19] (TITS2017)	97.2	80.0	89.5	41.6	45.3	56.4	60.5	64.6	91.4	68.7	94.2	76.1	56.4	92.4	45.7	60.6	27.0	48.7	61.8	97.2
ICNet [28] (ECCV2018)	97.1	79.2	89.7	43.2	48.9	61.5	60.4	63.4	91.5	68.3	93.5	74.6	56.1	92.6	51.3	72.7	51.3	53.6	70.5	97.1
LEDNet [24] (ICIP2019)	98.1	79.5	91.6	47.7	49.9	62.8	61.3	72.8	92.6	61.2	94.9	76.2	53.7	90.9	64.4	64.0	52.7	44.4	71.6	98.1
DABNet [11] (BMVC2019)	97.8	82.2	90.8	51.4	55.3	59.0	61.1	71.9	91.3	61.2	93.6	77.6	53.2	93.1	52.6	68.6	36.1	54.7	72.3	97.8
LCNet [22] (TITS2024)	97.4	80.1	90.2	52.3	52.5	57.1	52.8	67.0	91.2	59.3	93.1	74.3	48.8	92.4	44.0	62.8	51.9	44.0	69.6	97.4
LSA-BiNet(ours)	97.8	83.3	91.3	55.5	56.1	64.3	66.2	75.9	92.0	60.5	94.4	79.6	55.5	93.6	68.4	72.9	45.5	55.1	74.2	97.8

Table 3. Performance Comparison of LSA-BiNet Against Existing Semantic Segmentation Networks Estimated on the Camvid Test Dataset.

Method	Source	Pretrain	Input Size	mIoU (%) ↑	Params (M) ↓	Speed (fps) ↑	FLOPs (G) ↓
ENet [18]	ICCV2015	No	360×480	51.3	0.36	79.8	8.7
SegNet [1]	TPAMI2017	ImageNet	360×480	55.6	29.5	92.7	652.5
ESPNet [15]	ECCV2016	No	360×480	55.6	0.36	296.8	6.9
SwiftNet [17]	CVPR2019	No	720×960	63.3	11.8	-	52.0
CGNet [26]	TIP2020	No	360×480	65.6	0.5	101.5	14.0
EDANet [13]	MMAAsia2019	No	360×480	66.4	0.68	161.6	17.9
DABNet [11]	BMVC2019	No	360×480	67.04	0.76	108.1	20.9
ICNet [28]	ECCV2018	ImageNet	720×960	67.1	7.8	52.9	14.2
LCNet [22]	TITS2024	No	360×480	66.77	0.51	91.96	14.2
LSA-BiNet(ours)	This work	No	360×480	68.53	1.84	79.62	8.81

As summarized in Table 2, across the 19 object classes of Cityscapes, LSA-BiNet achieves the highest IoU in 12 categories: Sidewalk (Sid), Wall (Wal), Fence (Fen), Pole (Pol), Traffic Light (TLi), Traffic Sign (TSi), Pedestrian(Ped), Car (Car), Truck (Tru),Bus (Bus), Motorcycle (Mot) and Bicycle(Bic). Notable IoU improvements are observed in several of these challenging classes, including Wall (+3.2% over LCNet), Fence (+0.8% over DABNet), Pole (+1.5% over LEDNet),

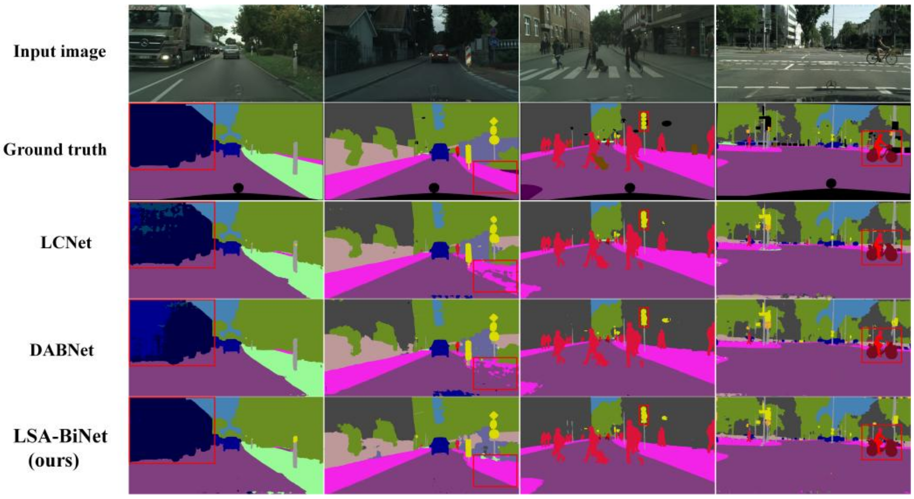

Fig. 7. Qualitative segmentation examples on the Cityscapes validation set. From top to bottom: Input images, Ground truth, the predictions of LCNet, DABNet and ours.

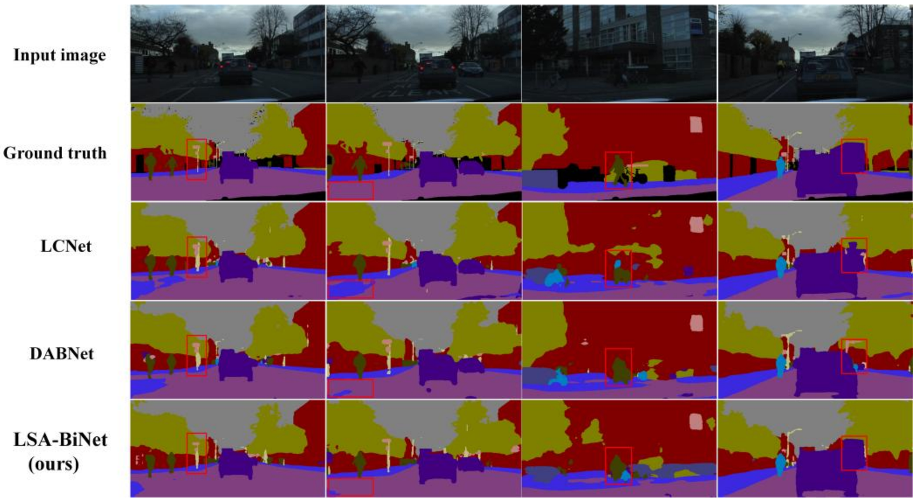

Fig. 8. Qualitative segmentation examples on the CamVid test set. From top to bottom: Input images, Ground truth, the predictions of LCNet, DABNet and ours.

Traffic Light (+4.9% over LEDNet), Truck (+4.0% over LEDNet), and Bus (+0.2% over ICNet). While LSA-BiNet does not lead in every category (for example, Road and Building), its consistently strong performance across diverse object types contributes significantly to its overall superior performance. On the CamVid benchmark (Table 3), LSA-BiNet establishes a new state-of-the-art performance with a mIoU of 68.53%, surpassing DABNet (67.04%), ICNet (67.1%), and LCNet (66.77%). Notably, it maintains the same computational efficiency (8.81G FLOPs) and a high inference speed (79.62

fps) as on Cityscapes. These results further attest to the architecture's robustness, especially given that they are achieved without any pertaining.

In addition, an analysis of training convergence, presented in Fig. 5 for Cityscapes and Fig. 6 for CamVid, reveals consistent optimization advantages of LSA-BiNet over the competing models. On Cityscapes, LSA-BiNet converges more rapidly in the initial epochs and maintains more stable loss descent trajectories than LCNet and DABNet. Likewise, on CamVid, it exhibits less loss oscillation and reaches a lower loss plateau earlier than its competitors. These characteristics suggest that LSA-BiNet has superior optimization properties, which enhance learning efficiency across diverse urban scenes. Qualitative visual comparisons in Fig. 7 (Cityscapes) and Fig. 8 (CamVid) further highlight LSA-BiNet's advantages in segmentation quality. The model consistently produces more precise segmentation boundaries than its counterparts, as evidenced by the cleaner delineation of objects such as traffic signs and vehicle contours (indicated by red boxes where competing models exhibit jagged edges or misalignments). LSA-BiNet also retains finer structural details that are crucial in complex urban scenes. For instance, it accurately captures thin poles—mitigating the breakages observed with LCNet and DABNet—and maintains intricate components like bicycle wheels, keeping the spokes and circular structures that other models often fail to preserve. This qualitative superiority translates into a noticeable reduction in visual artifacts and misclassification errors. In LSA-BiNet's outputs, there are fewer spurious speckles and blurred edges along object boundaries (especially around vegetation and road edges), and significantly fewer instances of mislabeling, for example, parts of the road being misclassified as sidewalk or fine structures being confused with the background. Many of these errors are corrected by LSA-BiNet, as highlighted by the red-boxed regions. These observable improvements are closely aligned with the model's leading quantitative performance metrics.

3.5 Ablation Experiments

Table 4. The results of ablation study on the CamVid dataset

Method	mIOU (%) ↑	Params (M) ↓	Speed (fps) ↑	Flops (G) ↓
-LRFA -SVFM -RCAD	67.04	0.75	108.10	20.9
-SVFM -RCAD	65.43	0.80	67.41	3.50
-LRFA -RCAD	65.32	1.44	76.39	5.14
-LRFA -SVFM	67.21	1.12	78.90	7.15
-RCAD	66.12	1.48	77.58	5.15
Full	68.53	1.84	79.62	8.81

Here, we analyze the impact of incorporating different attention modules on the model's segmentation performance. The ablation study involves evaluating the model on the validation set during training while systematically disabling certain modules. The evaluation results are summarized in Table 4. We conducted five ablation tests corresponding to various combinations of module removal:

- **-LRFA -SVFM -RCAD**: Removal of LRFA, SVFM and RCAD modules results in the most significant performance degradation compared to the Full

Method: ↓1.49% mIOU, ↑28.48 fps, ↑12.09 G FLOPs. The baseline model without attention mechanisms achieves only 67.04% mIOU, demonstrating the fundamental contribution of the proposed modules to segmentation accuracy: performance drop.

- **-SVFM -RCAD**: Simultaneous removal of SVFM and RCAD modules causes severe accuracy deterioration compared to the Full Method:(↓3.10% mIOU) and speed (↓12.21 fps). This indicates SVFM's critical role in feature aggregation ($\mathcal{F}_\rho = \mathcal{C}_{1\times1}^{(C)}(X_l + Y_d)$) and RCAD's importance in cross-level fusion ($\text{Output}_{\text{att}} = X \odot \text{Att}_{\text{map}} + \text{ResAtt}(\cdots) \odot Y_c$).
- **-LRFA -RCAD**: Ablation of LRFA and RCAD modules leads to substantial accuracy loss compared to the Full Method: (↓3.20% mIOU). The absence of LRFA's triple modulation ($\mathcal{A}(X) = \sigma\left(X_{[0]}\right) \odot \psi\left(\sigma\left(\mathcal{J}(X)\right)\right) \odot X$) and RCAD's residual attention significantly impacts feature representation capability.
- **-LRFA -SVFM**: Removal of both LRFA and SVFM modules preserves worser accuracy compared to the Full Method: (↓1.32% mIOU) but decreases computational cost (↓1.66G FLOPs). This suggests SVFM's EASA branch ($X_l = X \odot \text{Upsample}\left(\text{GELU}\left(\mathcal{C}_{1\times1}^{(C)}(\alpha X_s + \beta X_v)\right)\right)$) partially compensates for LRFA's absence through non-local interactions.
- **-RCAD**: Isolated removal of the RCAD reduces accuracy by ↓2.41% mIOU while decreasing parameters by *Δ*Params = 0.36M. This validates RCAD's efficiency in feature calibration via dual gating mechanism and residual weighting.

The ablation study confirms that each proposed module significantly contributes to the overall performance. The full model achieves optimal balance: ↑1.49% mIOU over baseline with minimal speed sacrifice (*Δ*28.48 fps). Notably, LRFA provides the most efficient accuracy gain per computation unit: $\frac{\Delta\text{mIOU}}{\Delta\text{FLOPs}} = \frac{1.21}{3.71} = 0.326\%/\text{GFLOP}$.

4 Conclusion

Through in-depth research on balancing accuracy and efficiency for real-time semantic segmentation, this paper proposes a novel real-time segmentation architecture named LSA-BiNet. The model achieves breakthroughs via three core innovative modules: the Local Receptive Field Attention (LRFA) module simulates high-order feature interactions with 1st-order computational complexity, significantly reducing computational overhead; the Spatial Variance Fusion Module (SVFM) integrates local details and non-local context through a dual-branch structure (EASA and LDE branches); the Residual Cross-layer Attention Decoder (RCAD) optimizes cross-scale feature fusion using dual-gating mechanisms and residual learning. On Cityscapes and CamVid benchmarks, LSA-BiNet achieves 72.74% and 68.53% mIoU respectively, requiring only 1.84M parameters and 8.81G FLOPs while reaching 51.08 FPS (Cityscapes) and 79.62

FPS (CamVid) at 512×1024, 360×480 resolution. Ablation studies confirm each module’s significant performance contribution (e.g., removing LRFA-SVFM-RCAD reduces mIoU by 1.49%) and demonstrates superiority over existing lightweight models without pretraining. LSA-BiNet achieves optimal balance among accuracy, speed, and computational resources for mobile devices. Future work will explore lightweight Vision Transformer (ViT) designs to further enhance semantic modeling capabilities.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (No. 62403076), the Humanities and Social Science Fund of Ministry of Education (No. 24YJCZH416) and Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province (New energy intelligent vehicle technology, 2024RC1029).

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence 39(12), 2481 – 2495 (2017)
2. Brostow, G.J., Fauqueur, J., Cipolla, R.: Semantic object classes in video: A high-definition ground truth database. Pattern recognition letters 30(2), 88 – 97 (2009)
3. Chen, L.C., 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 (ECCV). pp. 801 – 818 (2018)
4. Chen, L.C., 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 (ECCV). pp. 801 – 818 (2018)
5. Chen, X., Zhang, Y., Wang, Y.: Mtp: multi-task pruning for efficient semantic segmentation networks. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). pp. 1 – 6. IEEE (2022)
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213 – 3223 (2016)
7. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3146 – 3154 (2019)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770 – 778 (2016)
9. Hu, P., Perazzi, F., Heilbron, F.C., Wang, O., Lin, Z., Saenko, K., Sclaroff, S.: Real-time semantic segmentation with fast attention. IEEE Robotics and Automation Letters 6(1), 263 – 270 (2020)

10. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016)
11. Li, G., Yun, I., Kim, J., Kim, J.: Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. arXiv preprint arXiv:1907.11357 (2019)
12. Li, H., Xiong, P., Fan, H., Sun, J.: Dfanet: Deep feature aggregation for real-time semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9522 – 9531 (2019)
13. Lo, S.Y., Hang, H.M., Chan, S.W., Lin, J.J.: Efficient dense modules of asymmetric convolution for real-time semantic segmentation. In: Proceedings of the 1st ACM International Conference on Multimedia in Asia. pp. 1 – 6 (2019)
14. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431 – 3440 (2015)
15. 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 (ECCV). pp. 552 – 568 (2018)
16. Mehta, S., Rastegari, M., Shapiro, L., Hajishirzi, H.: Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9190 – 9200 (2019)
17. Orsic, M., Kreso, I., Bevandic, P., Segvic, S.: In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12607 – 12616 (2019)
18. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147 (2016)
19. Romera, E., Alvarez, J.M., Bergasa, L.M., Arroyo, R.: Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems 19(1), 263 – 272 (2017)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention – MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234 – 241. Springer (2015)
21. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510 – 4520 (2018)
22. Shi, M., Lin, S., Yi, Q., Weng, J., Luo, A., Zhou, Y.: Lightweight context-aware network using partial-channel transformation for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems 25(7), 7401 – 7416 (2024)
23. Treml, M., Arjona-Medina, J., Unterthiner, T., Durgesh, R., Friedmann, F., Schuberth, P., Mayr, A., Heusel, M., Hofmarcher, M., Widrich, M., et al.: Speeding up semantic segmentation for autonomous driving (2016)
24. Wang, Y., Zhou, Q., Liu, J., Xiong, J., Gao, G., Wu, X., Latecki, L.J.: Lednet: A lightweight encoder-decoder network for real-time semantic segmentation. In: 2019 IEEE international conference on image processing (ICIP). pp. 1860 – 1864. IEEE (2019)
25. Wang, Y., Zhou, Q., Xiong, J., Wu, X., Jin, X.: Esnet: An efficient symmetric network for real-time semantic segmentation. In: Pattern Recognition and Computer Vision: Second

Chinese Conference, PRCV 2019, Xi’an, China, November 8–11, 2019, Proceedings, Part II 2. pp. 41–52. Springer (2019)

26. Wu, T., Tang, S., Zhang, R., Cao, J., Zhang, Y.: Cgnet: A light-weight context guided network for semantic segmentation. IEEE Transactions on Image Processing 30, 1169–1179 (2020)
27. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems 34, 12077–12090 (2021)
28. 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 (ECCV). pp. 405–420 (2018)
29. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9308–9316 (2019)
30. Du, R., Feng, R., Gao, K., Zhang, J., Liu, L.: Self-supervised point cloud prediction for autonomous driving. IEEE Transactions on Intelligent Transportation Systems. (2024).
31. Duan, Y., Meng, L., Meng, Y., Zhu, J., Zhang, J., Zhang, J., & Liu, X. MFSA-Net: Semantic Segmentation With Camera-LiDAR Cross-Attention Fusion Based on Fast Neighbor Feature Aggregation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. (2024).
32. Gao, K., Li, X., Hu, L., Liu, X., Zhang, J., Du, R., & Li, Y. STMF-IE: A Spatial-Temporal Multi-Feature Fusion and Intention-Enlightened Decoding Model for Vehicle Trajectory Prediction. IEEE Transactions on Vehicular Technology. (2024).
33. Wu, J., Zhang, J., Zhu, J., Duan, Y., Fang, Y., Zhu, J., ... & Meng, Y. Multi-scale convolution and dynamic task interaction detection head for efficient lightweight plum detection. Food and Bioproducts Processing, 149, 353-367. (2025).