LMSNet for Real-Time Semantic Segmentation on Urban Road Images

Yi Cao, Hongbo Qu

***Abstract*—Semantic segmentation algorithms are essential for technologies in autonomous systems. The development of lightweight convolutional neural networks has facilitated the deployment of deep learning-driven semantic segmentation approaches on energy-efficient mobile devices. However, these lightweight networks often overlook feature fusion relationships, employing linear fusion strategies that limit segmentation accuracy. To tackle this challenge, this study introduces an innovative lightweight network design intended to serve as the core structure for the encoder. This method aims to decrease both the parameter count and the computational demands, thereby substantially boosting the network's efficiency in real-time applications. Additionally, a new Multi-Feature Extraction Module (MFM) is introduced before the decoder to capture multi-scale object information from images, thereby enhancing contextual understanding and improving segmentation accuracy and robustness. Furthermore, to further boost network performance, we extract high-frequency features from images and form a new stream. These features are then fused with existing feature maps through a High-Frequency Feature Fusion (HFF) module. This strategy captures finer details, thereby enhancing segmentation precision. Extensive experiments conducted on the publicly available Cityscapes and CamVid datasets using an NVIDIA 4090 GPU demonstrate average Intersection over Union (IoU) scores of 75.8% and 70.3%, respectively. Our approach not only surpasses existing network architectures in terms of accuracy but also significantly reduces computational resource consumption.**

***Index Terms*—autonomous driving; real-time semantic segmentation; lightweight networks; spatial attention mechanism**

I. Introduction

DRIVERLESS technology represents the convergence of artificial intelligence and transportation [1]. In recent years, rapid advancements in industrial technology have led to automobiles becoming ubiquitous in many households. The surge in private vehicle ownership in China, combined with frequent accidents caused by factors such as drunk or fatigued driving, has highlighted the importance and necessity of researching autonomous vehicles. This promising industry has attracted significant academic interest in unmanned vehicle research. Autonomous driving, also known as intelligent or driverless driving, involves vehicles autonomously planning travel routes and executing control actions via computer systems. Technologies such as radar, ultrasound, GPS, and computer vision are utilized to interpret various signals and perceive road conditions, enabling safe navigation.

The advent of driverless cars is poised to revolutionize transportation by alleviating human driving responsibilities, minimizing accidents attributable to human error, and addressing issues such as traffic congestion and environmental pollution. Despite these potential benefits, occasional reports of accidents involving casualties and property damage have sparked safety concerns. Ensuring accurate and rapid detection of road conditions, traffic signs, and other critical information is paramount for the advancement of autonomous vehicles. Real-time semantic scene segmentation stands out as a pivotal task in driverless technology, posing significant challenges to traditional computer vision methods. Autonomous driving environments necessitate precise object recognition and segmentation under stringent real-time constraints. Advances in deep learning have introduced methodologies that leverage convolutional neural networks (CNNs) for robust feature extraction and pixel-level classification, significantly improving real-time road segmentation performance. However, existing deep learning models frequently suffer from a large number of parameters and high computational complexity, which can impede their practical deployment in real-time applications within unmanned scenarios. Therefore, there is a pressing need for lightweight yet powerful models that can deliver high accuracy while meeting the strict latency requirements of autonomous driving systems.

Moreover, the diverse types of objects encountered in autonomous driving environments—such as vehicles, pedestrians, and lane markings—exhibit considerable variation in scale, shape, and appearance, significantly complicating the road segmentation task. The requirement for high-resolution images to ensure accurate semantic segmentation further intensifies computational demands. To tackle these challenges, we have developed LMSNet, a novel real-time semantic segmentation network specifically designed for driverless applications. LMSNet is characterized by its minimal parameter count and low computational complexity, enabling it to meet the stringent real-time performance requirements of autonomous driving systems. Our approach incorporates a lightweight backbone network optimized for reduced parameters and enhanced computational efficiency. Additionally, we introduce an innovative Multi-Scale Feature Extraction Module (MFM) that effectively fuses features across multiple scales, thereby improving segmentation performance. Extensive experiments on benchmark datasets have demonstrated LMSNet's superior segmentation capabilities, particularly in complex road scenarios.

- We developed a streamlined backbone network that features reduced parameters and lower computational demands, rendering it ideal for real-time semantic seg-

Manuscript received August 3, 2024; revised February 6, 2025. This work was supported in part by the Development and Research of Intelligent Gauge Detection System for Rail Transit Based on Dynamic Ranging under Grant 24B460027.

Yi Cao is a lecturer of Zhengzhou University of Science and Technology, Zhengzhou, 450064 China (corresponding author e-mail: caoyi153624@126.com).

Hongbo Qu is a lecturer of Zhengzhou University of Science and Technology, Zhengzhou, 450064 China (e-mail: quhongbo0330@163.com)

mentation applications.
- We designed an innovative MFM module that efficiently extracts multi scale features and enhances the integration of contextual information.
- We extracted high-frequency features from the images, formed a new stream, and fused these feature maps with the existing ones through a High-Frequency Feature Fusion (HFF) module.

II. Related work

In 2015, Long et al. [2] proposed a fully convolutional network by substituting the conventional CNN's fully connected layers with convolutional layers, which facilitated end-to-end learning for semantic segmentation. While this model had advantages in capturing global relationships and context, it suffered from the loss of significant spatial information during downsampling, resulting in coarse segmentation outputs. To overcome this constraint, Ronneberger et al. [3] introduced U-Net, a framework combining an encoder and decoder through extended skip connections that integrate spatial data from the encoder with semantic information from the decoder. preserving finer segmentation details. In an effort to mitigate spatial information loss during downsampling, Chen et al. [4], [5], [6], [7] introduced the DeepLab series of networks. These architectures utilized dilated convolutional layers to increase the receptive field while maintaining the feature map resolution, and they employed a spatial pyramid pooling mechanism to integrate multi-scale semantic information, improving the efficiency of feature exploitation. In 2017, Lin et al. [8] proposed RefineNet, which gradually merged low-resolution semantic features with high-resolution spatial features to leverage feature information across different resolutions, producing high-resolution output feature maps containing diverse scale features. However, the large parameter count and redundant structures in these networks hindered operational speed, limiting their applicability in real-time semantic segmentation scenarios where both speed and fine segmentation quality are crucial. As a response, lightweight convolutional neural networks have emerged as a key research area, focusing on reducing parameters and enhancing operational speed. For instance, ENet by Paszke [9] introduced a sparse convolutional layer and minimized channels in residual modules to significantly cut down network parameters and computations. The ShuffleNet and MobileNet series by Zhang et al. [10], [11], Howard et al. [12], and Sandler et al. [13] incorporated depth-wise separable convolution layers to lower computational requirements with minimal accuracy loss. Further innovations include the introduction of separable convolution layers in the MobileNet family, as seen in ERFNet proposed by Romera et al. [14], where a standard 3x3 convolutional layer is decomposed into two 3x1 and 1x3 convolutional layers, reducing computational load while maintaining receptive fields. Lightweighting techniques have also been applied to semantic segmentation tasks, enabling a better balance between segmentation accuracy, network size, and operational speed as showcased in various studies [15], [16], [17], [18]. For instance, LCANet proposed by Dong et al. [19] utilizes nonlinear fusion to enhance segmentation by learning relationships between different feature maps.

While existing networks have made strides in accuracy and real-time performance, they often neglect semantic context relationships, potentially leading to information loss and reduced precision. To tackle these challenges, this work introduces LMSNet, a streamlined convolutional architecture designed for real-time semantic segmentation by incorporating multi-scale feature extraction.

III. Method

The network structure of LMSNet is designed to optimize real-time semantic segmentation performance while maintaining a lightweight architecture. It comprises several key components aimed at enhancing segmentation accuracy and operational efficiency.

A. Structure of Network

The architecture of our proposed LMSNet, as illustrated in Figure 1, adheres to a conventional encoder-decoder framework while emphasizing a streamlined structure to facilitate efficient end-to-end training. The encoder is built upon a three-stage backbone network, where blue segments denote downsampling modules integrating convolutional layers with a stride of 2 and a 3x3 kernel, alongside pooling layers with a 2x2 kernel, effectively reducing information loss during encoding. Within this backbone, excluding stage 4, the first unit of each stage serves as a downsampling element, resulting in a cumulative downsampling ratio of only 8, which preserves spatial details. These blue areas represent Enhanced Convolution (EConv) modules, thoroughly described in Section III-B. For the decoder, we employ a gradual upsampling strategy using three transposed convolutional layers to incrementally reconstruct feature map resolution and mitigate spatial data loss. Each upsampling unit includes a 1x1 convolutional layer, batch normalization (BN) [20], a rectified linear unit (ReLU), and twofold bilinear interpolation, culminating in a pixel-level classifier at the final stage. To enhance feature representation, a Multi-scale Feature Extraction Module (MFM) is integrated between the encoder and decoder stages, fusing features from multiple scales to improve segmentation accuracy (Section III-C). Additionally, we introduce a high-frequency feature extraction stream from RGB images, which are merged with existing feature maps via a High-Frequency Feature Fusion (HFF) module, capturing nuanced details and enhancing segmentation precision (Section III-D).

B. EConv Module

He et al. introduced an extended foundational network architecture epitomized by ResNet. This architecture enables the network to capture intricate contexts and semantic nuances within images by progressively deepening the layers and diminishing the resolution of feature maps. On the other hand, MobileNetV3 achieves innovation through substituting conventional 3x3 convolutional layers with depthwise separable 3x3 convolutions. This substitution notably reduces the network size and computational burden, thereby enhancing efficiency while preserving essential representational capacity. This makes it especially apt for deployment in environments with limited computational resources, such as mobile and embedded systems. The integration of the SE (Squeeze-and-Excitation) attention mechanism further boosts network performance by recalibrating channel-wise feature responses. However, the presence of two fully-connected

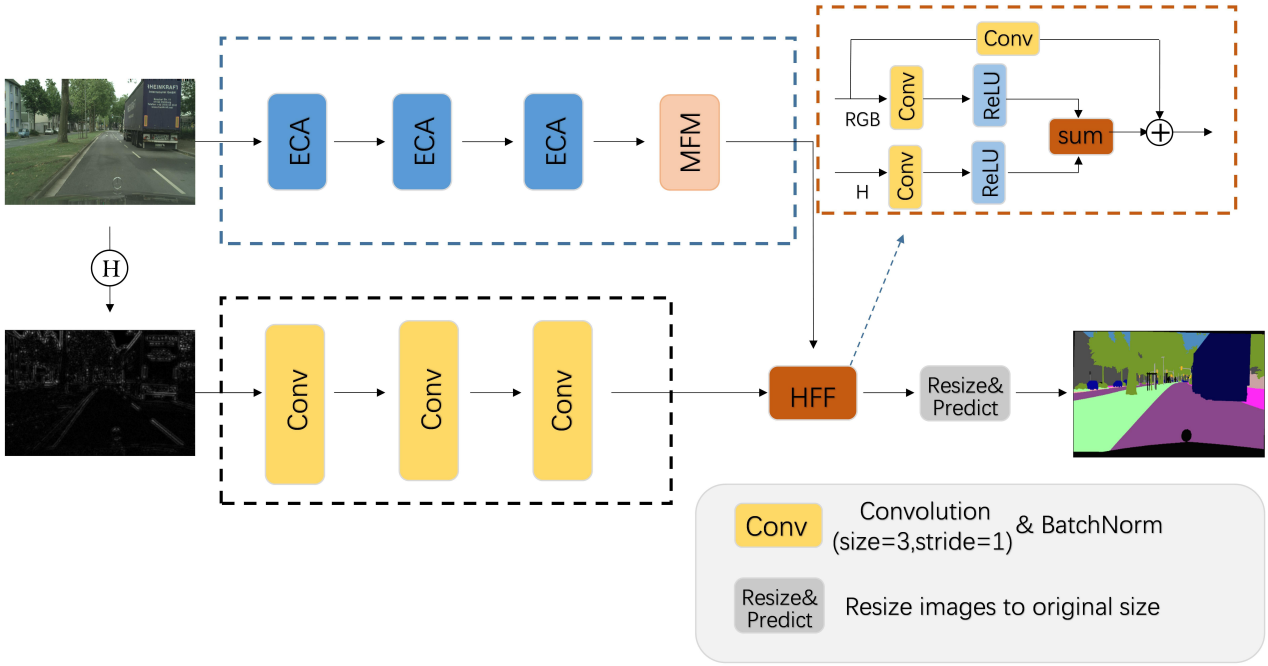

Fig. 1. An overview of the LMSNet architecture designed for real-time semantic segmentation. This architecture is carefully crafted to achieve a balance between precision and computational efficiency, making it ideal for use in environments with limited resources. herein, ECA refers to the Efficient Channel Attention Module, MFM represents the Multi-scale Feature Aggregation Module, and HFF stands for the High-Frequency Feature Integration Module.

layers within the SE module introduces additional parameters, which can be a drawback for lightweight applications. Building upon the advancements of MobileNetV3, we have replaced the SE module with the more parameter-efficient ECA (Efficient Channel Attention) module [23], thereby reducing the number of parameters while maintaining or even enhancing performance. Additionally, we have introduced the Enhanced Convolution (EConv) module, as shown in Figure 2. These modifications aim to refine the model's parameters further and enhance the overall lightweight design of our architecture, ensuring efficient real-time processing capabilities without compromising on accuracy.

The ECA module presents an effective channel attention mechanism that utilizes a localized cross-channel interaction approach without reducing dimensionality, effectively mitigating any negative impacts of dimensionality reduction on channel attention learning effectiveness. In this research, the ECA module is employed for residual propagation, strengthening channel characteristics and improving network segmentation performance. After two consecutive operations with 3 × 3 convolutional kernels, we obtain integrated features of dimensions 1 × 1 × C through global average pooling (GAP). The ECA module governs channel weights by swiftly convolving in one dimension with a size k, where k dynamically adapts based on the mapping of channel dimensions C, with σ denoting the sigmoid activation function.

C. Multi-scale Feature Extraction Module

Prior real-time semantic segmentation techniques have overlooked image context to maintain a lightweight design, often resulting in the loss of intricate features and a decline in network segmentation precision. Drawing inspiration from this limitation, we present the multi-scale feature extraction module. In this module, we use dilated convolutions with dilation rates set at 2, 4, and 6, along with 3 × 3 convolution kernels. This combination enables local features in the deeper layers to connect with a wider receptive field., preventing the loss of minute target features during data transmission. As shown, starting from the top and moving downward, the first branch utilizes a 1 × 1 convolution to maintain the original receptive field. The following three branches (second to fourth) use dilated convolutions with different dilation rates, each designed to capture features at unique receptive fields. The fifth branch applies global average pooling to the input to gather broad, overarching features. Finally, the feature maps from all five branches are concatenated along the channel dimension, and multi-scale information is integrated via a 1 × 1 convolution, resulting in the creation of a new feature map, F.

Subsequently, to derive spatial attention details, the combined feature map F is fed into the Spatial Attention Module (SAM) [24], resulting in the final feature map F_s. Furthermore, the detailed workflow is illustrated in Figure 3, while the formula for CAM is provided in Equation 1.

$$\mathbf{M(s)} = \sigma(f^{5\times5}([\mathbf{F_{avg}}; \mathbf{F_{max}}])), \tag{1}$$

Where σ denotes the sigmoid function, and f signifies the convolution operation utilizing a 5 x 5 kernel. Here, $F_{avg} \in \mathbb{R}^{1\times H\times W}$ and $F_{max} \in \mathbb{R}^{1\times H\times W}$.

In this equation, σ denotes sigmoid function, while f signifies a convolution operation utilizing with a 5 x 5 kernel.

The aforementioned Multi-scale Feature Module (MFM) expands the sensory field and enriches feature information

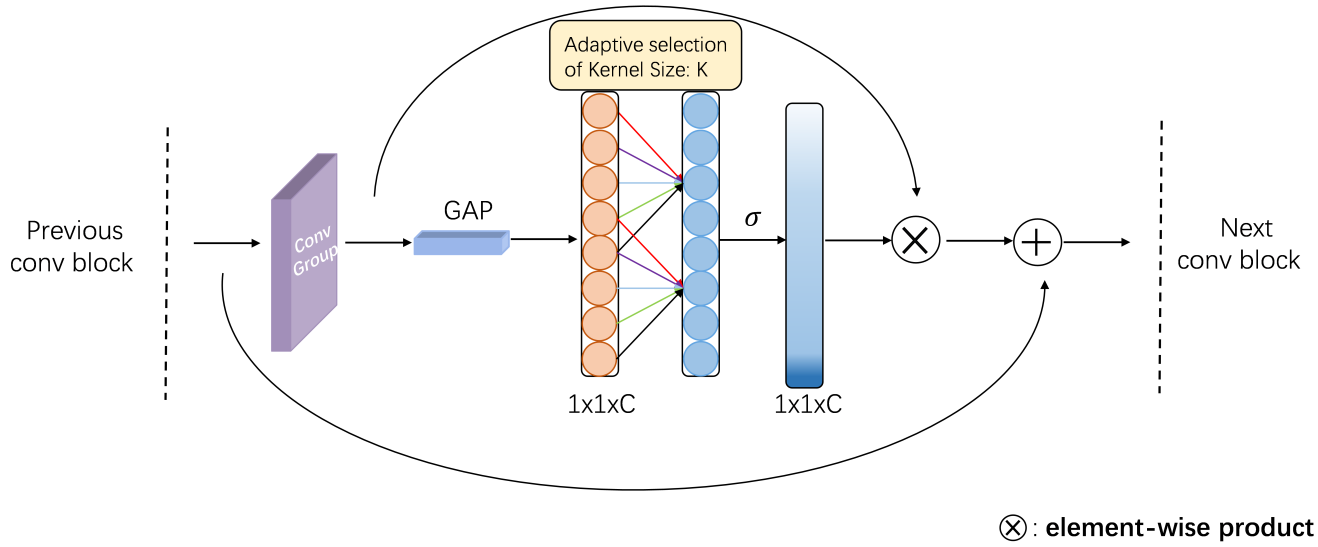

Fig. 2. The structure of the Econv. The EConv module integrates several innovative components to achieve superior performance while maintaining computational efficiency, making it particularly well-suited for real-time semantic segmentation tasks.

by amplifying the multilevel sampling rate of convolutional parallel sampling. Consequently, image-level features adeptly capture global contextual information. By considering the interplay between these contexts, the module mitigates segmentation errors stemming from fixation on local features, thereby enhancing image segmentation accuracy. Furthermore, the module consolidates contextual information across various scales, bolstering the network's capability to discern regions of diverse sizes on the road.

D. High-Frequency Feature Fusion

To further enhance the segmentation performance of the network, we introduced a high-frequency filter as a preprocessing step. The filter is designed to augment the edges and other high-frequency components of various objects within images. By emphasizing detailed features such as object boundaries and textures, the filter not only improves the model's ability to recognize target regions but also promotes finer semantic segmentation outcomes. High-frequency components are crucial for capturing subtle changes in pixel intensity, which are often indicative of important structural information. This enhancement facilitates the extraction of meaningful features during the convolutional process, leading to more precise and reliable segmentation results.

However, simply combining feature maps through addition can lead to the loss of detailed features. Such simplistic approaches fail to adequately preserve the rich information contained within each feature map, potentially degrading the network's performance. To address this limitation, in this paper, we propose a novel module called High-Frequency Feature Fusion (HFF). This module integrates two types of feature maps while preserving their detailed characteristics.

As illustrated in Figure 1, the HFF process involves several key steps: Channel Alignment: Initially, 1x1 convolution operations are applied to both feature maps to align their channel numbers. This ensures that each feature map has the same number of channels, facilitating subsequent processing steps. Batch Normalization: After performing channel alignment, we apply two distinct batch normalization layers to normalize each feature map separately. This technique enhances training stability and speed by mitigating internal covariate shift, thereby maintaining uniform input distributions throughout the network layers. Feature Fusion: After normalization, the two feature maps are fused using element-wise summation. This method allows for seamless integration of the enhanced high-frequency features with the original feature maps, maintaining the integrity of detailed information. Activation Function: Lastly, a ReLU activation function is utilized to add non-linearity to the combined features. The ReLU function helps retain only the positive elements, effectively highlighting significant features while suppressing less relevant ones. By carefully integrating these steps, the HFFF module ensures that detailed features are preserved and effectively utilized, this results in improved segmentation precision and enhanced network robustness.

IV. Experiments

A. Experimental Settings

Experimental Dataset: For dataset evaluation, we selected two widely recognized city road datasets, Cityscapes [25] and Camvid [26], to assess the performance of our network. To broaden the dataset and mitigate sample imbalance issues, this study incorporates data augmentation techniques on the training dataset, including random cropping (RC), vertical axis mirroring (MVA), and color shifting (CS). In the training stage using the Cityscapes dataset, input images are randomly cropped to a size of 512x1024. For the CamVid dataset, images are randomly cropped to 360x480. Such data augmentation techniques improve the model's robustness and ability to generalize by introducing it to a broader range of variations and conditions found in real-world imagery.

Experimental Metrics: In experimental settings, choosing suitable evaluation metrics is crucial for accurately assessing network performance. This study employs evaluation metrics that have been widely used in prior relevant research [19]. The Mean Intersection over Union (MIoU) is adopted as a

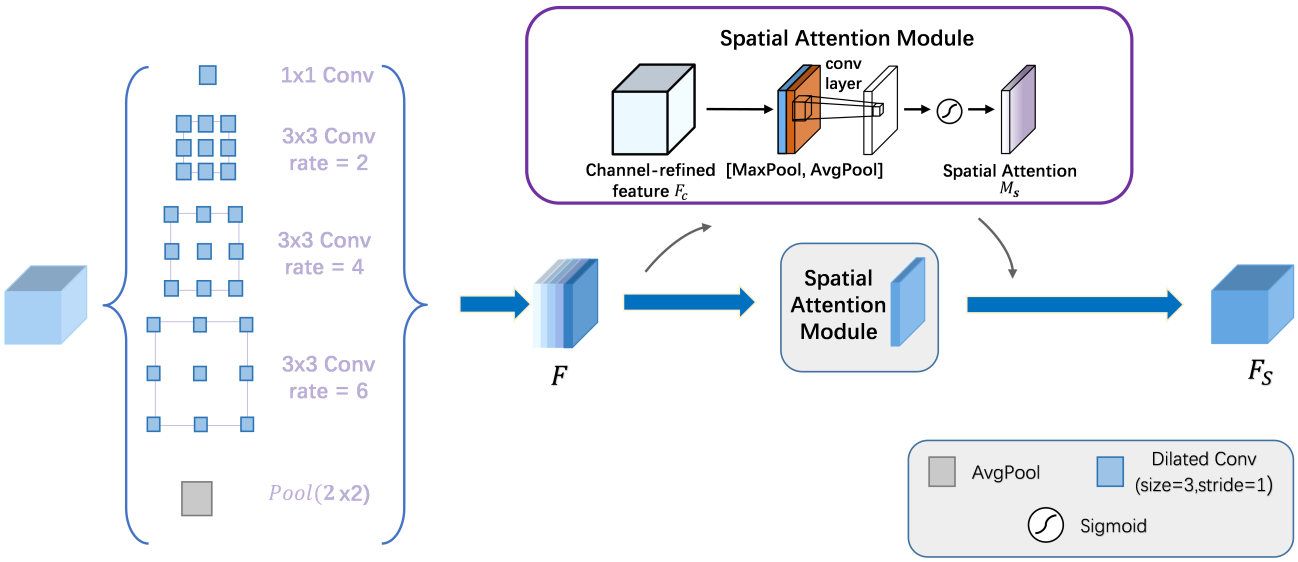

Fig. 3. Architecture of the Multi-scale Feature Extraction Module. The MFM incorporates dilated convolutions at different dilation rates along with several feature extraction pathways to efficiently gather comprehensive multi-scale contextual details.

TABLE I
OVERVIEW OF THE EVALUATION DATASETS FOR OUR NETWORK (THE NUMBERS INDICATE THE QUANTITY OF IMAGES UTILIZED FOR TRAINING AND TESTING. SYMBOLS ✓ AND ✗ DENOTE THE INCLUSION OR EXCLUSION OF SPECIFIC DATA AUGMENTATION TECHNIQUES, RESPECTIVELY)

Datast	Number			Type		
	Train	Val	Test	RC	MVA	CS
Cityspaces [25]	2975	500	1525	✓	✗	✓
Camvid [26]	367	101	233	✓	✓	✓

key metric to measure the pixel-level accuracy of segmentation models. MIoU is computed by averaging the IoU scores across all classes. The IoU is determined by comparing the predicted segmentation masks against the actual ground truth labels. The formula for IoU is provided in Equation 2.

$$IoU = \frac{TP}{TP + FP + FN} \tag{2}$$

Here, TP denotes the count of pixels for which both the actual label and the predicted label match a specific class. FP refers to the number of pixels where the predicted label indicates the class, but the actual label does not. FN represents the number of pixels where the actual label indicates the class, but the predicted label does not.

Compared Methods: To assess the advancement of the method proposed in this paper, we conducted experiments on the Cityscapes and CamVid test sets with LMSNet and other established networks. Our evaluation not only compares against large semantic segmentation networks but also juxtaposes the current lightweight real-time semantic segmentation networks. The purpose of this comparative study is to demonstrate the effectiveness and efficiency of the proposed approach in relation to current state-of-the-art methodologies.

Implementation Details: The implementation of our method was carried out on a server outfitted with an NVIDIA GeForce RTX 3090 GPU using PyTorch. The network was trained end-to-end employing the Adam optimizer with a momentum of 0.8, a weight decay of $1e^{-5}$, and an initial learning rate of $3e^{-3}$. To maximize GPU memory utilization, a batch size of 16 was employed for training on the Cityscapes dataset, and a batch size of 32 was utilized for training on the CamVid dataset. These configurations were optimized to facilitate efficient training and to leverage the computational capabilities of the GPU for enhanced performance. Our approach was implemented on a server equipped with an NVIDIA GeForce RTX 3090 GPU, utilizing the PyTorch framework. We trained the network in an end-to-end manner using the Adam optimizer, configured with a momentum value of 0.8, a weight decay of $1e^{-5}$, and an initial learning rate set to $3e^{-3}$. To optimize GPU memory usage, we used a batch size of 16 for training on the Cityscapes dataset and a batch size of 32 for the CamVid dataset. These settings were fine-tuned to ensure efficient training processes and to take full advantage of the GPU's computational power for improved performance.

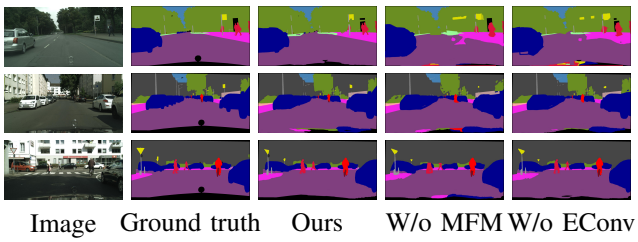

Fig. 4. For visual comparison: Ground Truth (GT) Masks, Full Model Predictions, Predictions without Multi-scale FE Module, and Predictions without ECov Module.

B. Compared Detection Methods

To assess the performance of our proposed network, we perform a comparative study against various leading methodologies. Our comparison encompasses both extensive large-scale semantic segmentation techniques and efficient real-time semantic segmentation models.

TABLE II
THE MIOU AND FRAME RESULT (% AND FPS) OF OUR METHOD AND THE OTHER METHODS ON CITYSPACES DATASET.

Network Type	Methods	MIoU(%)	Frame(fps)
Large	SegNet [29]	57.0	17
	DeepLabV2 [5]	70.4	<1
	SFNet [30]	78.4	<1
Lightweight	ENet [8]	58.3	77
	CGNet [27]	64.8	50
	BiSeNet [28]	68.4	106
	DABNet [16]	70.1	104
	LCANet [19]	72.7	86
	Ours	75.8	101

TABLE III
THE MIOU AND FRAME RESULT (% AND FPS) OF OUR METHOD AND THE OTHER METHODS ON CAMVID DATASET.

Network Type	Methods	MIoU(%)	Frame(fps)
Large	SegNet [29]	46.4	5
	DeepLab [4]	61.6	5
	PSPNet [31]	69.1	5
Lightweight	ENet [8]	51.3	61
	CGNet [27]	65.6	-
	BiSeNet [28]	65.6	-
	FDDWNet [17]	66.9	79
	LCANet [19]	67.1	105
	Ours	70.3	106

Table II presents the quantitative results obtained by each method on the Cityscapes dataset. It is evident that even without utilizing additional training data, LMSNet achieved an MIoU of 75.8% while operating at a frame rate of 101fps. Our network surpasses all lightweight networks in terms of the MIoU metric, even outperforming some large-scale semantic segmentation networks like DeepLabV2 and SegNet. Although SFNet achieved an MIoU of 78.4%, its processing speed is insufficient for real-time semantic segmentation tasks. In contrast to lightweight networks, while our network runs slightly slower than BiSeNet and DABNet, our method's MIoU metric is the highest, surpassing them by 3.3% and 5%, respectively.

As illustrated in Figure 5, a scatter plot comparison of various lightweight networks on the Cityscapes test set reveals LMSNet's superior performance in both segmentation accuracy and running speed. Positioned at the apex of this plot, our proposed network outperforms its counterparts in segmentation precision while concurrently maintaining high-speed inference capabilities. LMSNet's efficient architecture, characterized by a reduced parameter count, exemplifies an optimal balance between computational efficiency and predictive accuracy. This positioning not only signifies LMSNet's leadership among lightweight semantic segmentation models but also highlights its capability to deliver state-of-the-art segmentation results without sacrificing operational efficiency. The equilibrium achieved by LMSNet between accuracy and speed sets a new standard for real-time applications and contributes significantly to the advancement of lightweight deep learning architectures. This achievement underscores the potential for deploying advanced semantic segmentation models in environments with limited computational resources.

Table III summarizes the performance metrics obtained by various methods on the CamVid dataset. Remarkably, LMSNet achieves an mIoU of 75.8% and processes images at a rate of 101 frames per second, all without using extra training data. These results underscore LMSNet's superior mIoU performance relative to other lightweight real-time semantic segmentation models that were assessed. In conclusion, our network strikes a good balance in segmentation accuracy, running speed, and other aspects.

C. Robustness Evaluation

Beyond benchmarking against existing state-of-the-art techniques, we performed supplementary experiments to test the resilience of our proposed model. Specifically, we applied several types of image distortions to the validation set images from the Cityscapes dataset. These distortions comprised resizing, JPEG compression using a quality factor η, and Gaussian blurring with a kernel size κ.

Details of the parameters used and the results for manipulation detection, evaluated using MIoU and Frame metrics, are summarized in Table IV. Our model shows strong performance across different types of distortions. Particularly for compressed images, the MIoU decreases by just 0.4% compared to undistorted images at a quality factor of 100, and by 1.0% at a quality factor of 80. This indicates that our network exhibits significant robustness against various image distortion methods.

TABLE IV
THE MIOU AND FRAME RESULT (% AND FPS) ON CITYSPACES DATASET UNDER VARIOUS DISTORTION.

Distortion	MIoU	Frame
No distortion	75.8 ↓0.0	101 ↓0.0
Resize (0.75×)	74.9 ↓0.9	101 ↓0.0
Resize (0.30×)	72.6 ↓3.2	103 ↑2.0
GaussianBlur (κ = 3)	72.8 ↓3.0	99 ↓2.0
GaussianBlur (κ = 7)	70.4 ↓5.4	97 ↓4.0
JPEGCompress (η = 100)	72.9 ↓3.1	100 ↓1.0
JPEGCompress (η = 80)	70.7 ↓5.1	98 ↓3.0

D. Ablation Study

In this study, we integrated the EConv module and MFM module into conventional semantic segmentation networks. To systematically evaluate the contributions of these two proposed modules, we conducted ablation studies by selectively removing each module from our network architecture. The performance of the modified models was assessed using the Cityscapes and CamVid datasets.

When MobileNetV3 was substituted for EConv as the backbone network, no significant changes were observed in the Mean Intersection over Union (MIoU) metric on both

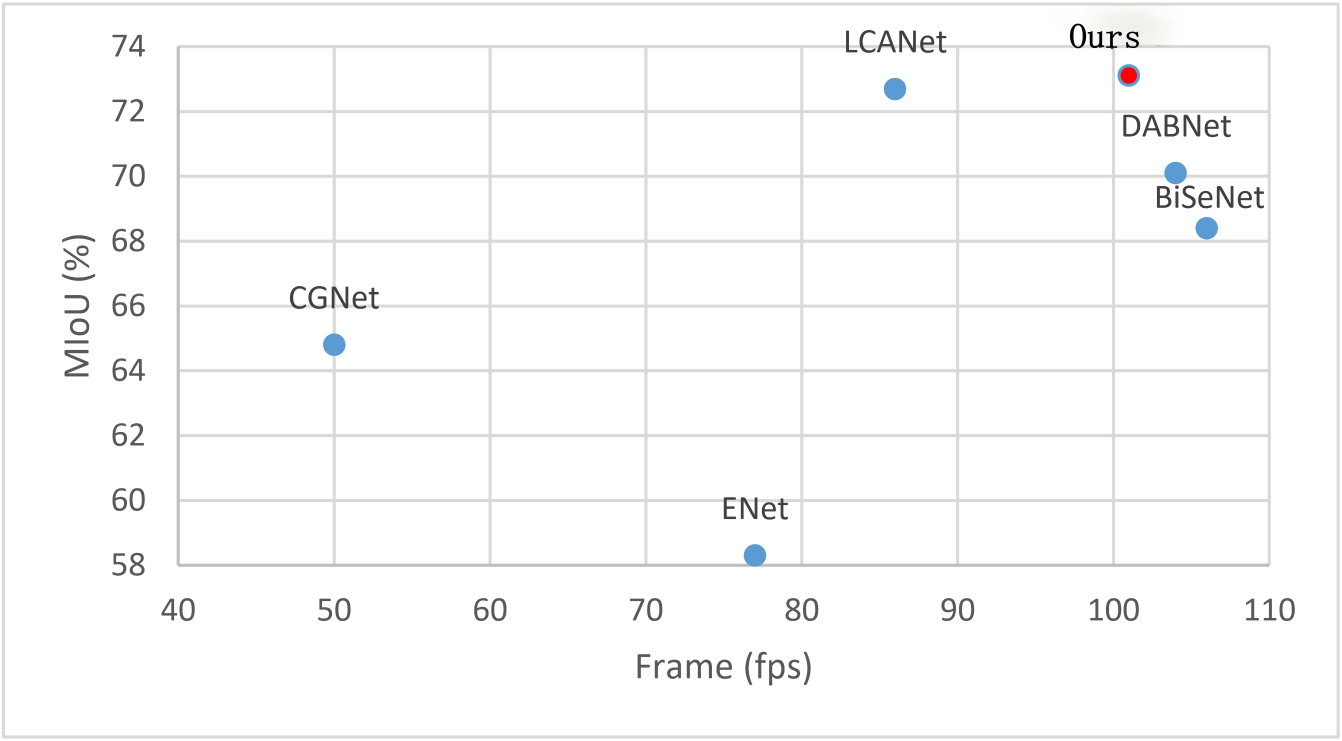

Fig. 5. A Comparative Analysis of Accuracy and Operational Speed in Lightweight Networks.

datasets. However, a notable decrease in inference speed was recorded, with reductions of 5.2% and 5.8% on the respective datasets. This performance degradation is likely attributable to EConv's streamlined architecture, which features one fewer fully connected layer compared to MobileNetV3. Our experimental findings indicate that employing EConv as the backbone network enhances computational efficiency without compromising segmentation accuracy.

Excluding the MFM module between the encoder and decoder layers led to an increase in running speed but resulted in a substantial drop in accuracy, with MIoU decreasing by 6.7% and 4% on the Cityscapes and CamVid datasets, respectively. These results underscore the critical role of MFM in maintaining high segmentation accuracy under reduced computational loads by effectively preserving and transmitting multi-scale contextual information.

Omitting the High-Frequency Feature (HF) stream yielded a slight improvement in running speed, However, it also caused a decline in segmentation accuracy, with MIoU reductions of 2.7% and 1.1% on the respective datasets.

As illustrated in Figure IV-A, the visual results corresponding to Table V demonstrate that LMSNet achieves accurate segmentation with well-defined object boundaries. Compared to other methods, LMSNet exhibits superior performance in capturing detailed segmentation features across various object categories. For example, in the segmentation of pedestrian walkways, which occupy significant spatial proportions, LMSNet excels in accurately delineating local regions. Additionally, less frequent objects such as trucks, bicycles, and pedestrians in smaller spatial proportions are precisely identified and segmented with clear boundaries.

TABLE V
ABLATION RESULTS ON CITYSCAPES AND CAMVID DATASETS, MIOU AND FRAME (% AND FPS) ARE REPORTED

Variants	Cityspaces		CamVid	
	MIoU	Frame	MIoU	Frame
w/o EConv module	72.9	96	69.3	102
w/o MFM module	69.1	108	66.3	117
w/o HF stream	73.1	113	69.2	120
Ours	75.8	101	70.3	108

V. CONCLUSIONS

In this paper, we propose LMSNet for real-time semantic segmentation. In the encoder part of LMSNet, we leverage the advantages of the ECA module to introduce a more lightweight backbone network while preserving network performance. Subsequently, between the encoder and decoder, in order to better retain image context information, we introduce a novel Multi-Feature Extraction Module (MFM), capable of extracting object information at multiple scales. Furthermore, to further enhance the network's performance, we extract high-frequency features from the images, forming a new stream, and fuse these two types of feature maps through a High-Frequency Feature Fusion (HFF) module. This strategy helps capture more detailed information, thereby improving segmentation accuracy. Through experiments, We assess the performance of each component in our network and benchmark it against other leading networks. The experimental outcomes on the Cityscapes and CamVid datasets showcase the competitive edge of LMSNet, outperforming most advanced lightweight semantic segmentation networks and even surpassing some large-scale semantic segmentation networks. This fully illustrates that LMSNet achieves a favorable balance in segmentation accuracy, network scale, and operational speed.

REFERENCES

[1] Huang Meiyi. A brief introduction on autonomous driving technology. Science & Technology Information, 2017, 15(27): 1-2(in Chinese), **2017**,

[2] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation . In Proceedings of the on Computer Vision and Pattern Recognition, **2015**, *3431-3440*

[3] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In Proceedings of Proceedings of Medical Image Computing and Computer-Assisted Intervention, 2015, 234-241.

[4] Chen L C, Papandreou G, Kokkinos I. Semantic image segmentation with deep convolutional nets and fully connected CRFs Computer Science, 2014(4): 357-361.

[5] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Transactions on Pattern Analysis and Machine Intelligence 2018, 40(4): 834- 848

[6] Chen, Liang-Chieh, et al. "Rethinking atrous convolution for semantic image segmentation." arxiv preprint arxiv:1706.05587 (2017).

[7] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 833-851

[8] Lin G S, Milan A, Shen C H, et al. RefineNet: multi-path refinement networks for high-resolution semantic segmentation[C]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 5168-5177

[9] Paszke, Adam, et al. "Enet: A deep neural network architecture for real-time semantic segmentation." arxiv preprint arxiv:1606.02147 (2016).

[10] Zhang X Y, Zhou X Y, Lin M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 6848-6856

[11] Ma N N, Zhang X Y, Zheng H T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design. In Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 122-138

[12] Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arxiv preprint arxiv:1704.04861 (2017).

[13] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).

[14] Romera, E., Alvarez, J. M., Bergasa, L. M., & Arroyo, R. (2017). Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems, 19(1), 263-272.

[15] Wang Y, Zhou Q, Liu J, et al. Lednet: a lightweight encoder-decoder network for real-time semantic segmentation. In Proceedings of the IEEE International Conference on Image Processing. Los Alamitos: IEEE Computer Society Press, 2019:1860-1864.

[16] Li, G., Yun, I., Kim, J., & Kim, J. (2019). Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. arxiv preprint arxiv:1907.11357.

[17] Liu, Jia, et al. FDDWNet: a lightweight convolutional neural network for real-time semantic segmentation. In Proceedings of the ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2020.

[18] Jiang W, Xie Z, Li Y, et al. LRNNET: a light-weighted network with efficient reduced non-local operation for real-time semantic segmentation. In Proceedings of the IEEE International Conference on Multimedia & Expo Workshops. Los Alamitos: IEEE Computer Society Press, 2020: 1-6.

[19] Rongsheng D, Yi L,Yuqi M. Lightweight Network with Convolutional Attention Feature Fusion for Real-Time Semantic Segmentation. Journal of Computer-Aided Design & Computer Graphics, 2023 Jun. Vol.35 No.6.

[20] Ioffe, Sergey, and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning, pp. 448-456. pmlr, 2015.

[21] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[22] Howard, Andrew, et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[23] 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 IEEE/CVF conference on computer vision and pattern recognition (pp. 11534-11542).

[24] Woo, Sanghyun, et al. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV). 2018.

[25] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 3213-3223

[26] Brostow G J, Shotton J, Fauqueur J, et al. Segmentation and recognition using structure from motion point clouds. In Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2008: 44-57

[27] Wu T Y, Tang S, Zhang R, et al. CGNet: a light-weight context guided network for semantic segmentation[J]. IEEE Transactions on Image Processing, 2021, 30: 1169-1179

[28] Yu C Q, Wang J B, Peng C, et al. BiSeNet: bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 334-349

[29] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495

[30] Li X T, You A S, Zhu Z, et al. Semantic flow for fast and accurate scene parsing. In Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2020: 775-793

[31] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arxiv preprint arxiv:1511.07122, 2015.

[32] He, K.; Zhang, X.; Ren, S. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

[33] Korus, P.; Huang, J. Evaluation of random field models in multi-modal unsupervised tampering localization. In Proc. of IEEE Int. Workshop on Inf. Forensics and Security, 2016.