ECMNet:Lightweight Semantic Segmentation with Efficient CNN-Mamba Network

Feixiang Du 1,2 and Shengkun Wu 1 1* TLU. 2 SUT.

Abstract

In the past decade, Convolutional Neural Networks (CNNs) and Transformers have achieved wide application in semantic segmentation tasks. Although CNNs with Transformer models greatly improve performance, the global context modeling remains inadequate. Recently, Mamba achieved great potential in vision tasks, showing its advantages in modeling long-range dependency. In this paper, we propose a lightweight Efficient CNN-Mamba Network for semantic segmentation, dubbed as **ECMNet**. ECMNet combines CNN with Mamba skillfully in a capsule-based framework to address their complementary weaknesses. Specifically, We design a Enhanced Dual-Attention Block (EDAB) for lightweight bottleneck. In order to improve the representations ability of feature, We devise a Multi-Scale Attention Unit (MSAU) to integrate multi-scale feature aggregation, spatial aggregation and channel aggregation. Moreover, a Mamba enhanced Feature Fusion Module (FFM) merges diverse level feature, significantly enhancing segmented accuracy. Extensive experiments on two representative datasets demonstrate that the proposed model excels in accuracy and efficiency balance, achieving 70.6% mIoU on Cityscapes and 73.6% mIoU on CamVid test datasets, with 0.87M parameters and 8.27G FLOPs on a single RTX 3090 GPU platform.

Keywords: Semantic segmentation, Lightweight, Convolutional neural network, Mamba

1 Introduction

Semantic segmentation aims to assign a label to each pixel in a given image, which is widely applied in autonomous driving[1], remote sensing[2], and agriculture[3], and more.

Early semantic segmentation primarily relied on CNNs, employing techniques like large convolutional kernels [4], dilated convolutions[5], and feature pyramids[6] to extend receptive fields. However, these CNN-based approaches remained limited in capturing long-range dependencies. The advent of Transformers[7] enabled more effective global context modeling in subsequent segmentation methods. Learning global context dependencies is essential for extracting global semantic features, particularly in intensive tasks like semantic segmentation. The rise of Visual Transformer (ViT)[8] has injected a new paradigm for semantic segmentation. SETR[9] slices images into sequences for the first time and captures global context feature through a self-attentive mechanism, outperforming traditional CNN models on complex scene datasets such as Cityscapes. Meanwhile, SegFormer[10] further optimized the architectural design by proposing a hierarchical Transformer encoder with a lightweight MLP decoder to achieve multi-scale feature fusion. However, the square-level computational complexity of Transformer limits its application to high-resolution images with insufficient sensitivity to local details.

To tackle the limitation of the above single model and extract fine spatial details, some models treated semantic segmentation tasks by integrating CNN with Transformer. For instance, HResFormer[11], PFormer[12], and DMFC-UFormer[13] have achieved satisfactory results in the field of medical image segmentation. However, the self-attention mechanism in CNN-Transformer methods still poses challenges in terms of speed and memory usage when dealing with long-range visual dependencies, especially processing high-resolution images.

Unlike previous Transformer, Mamba[14] shows great potential for high-resolution image by efficient sequence modeling with linear complexity. Vision Mamba[15] have recently demonstrated remarkable success in various computer vision tasks. For example, in the field of 3D medical imaging, SegMamba[16] achieves real-time inference on the colorectal cancer dataset CRC-500, with a speedup of 30% compared to 3D UNet. In addition, CM-UNet[17] introduces a Mamba decoder into a CNN encoder to bridge local and global features through a channel-space attention mechanism, achieving higher mIoU on the ISPRS Vaihingen dataset.

To accommodate limited computational resources and mobile device application, lightweight semantic segmentation models receive higher attention. For example, LEDNet[18] employed channel split-and-shuffle operations within residual blocks, significantly lowering computational complexity. While CFPNet[19] designed Channelwise Feature Pyramid (CFP) module to significantly reduces model parameters and model scale by extracting various level feature map and contextual feature information jointly. LETNet[20] used an LDB module and FE module for enhanced efficiency and accuracy with reduced model complexity.

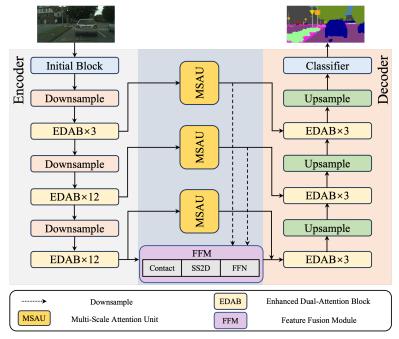
Motivated by the success of Mamba and lightweight approaches in semantic segmentation tasks, We propose ECMNet, an efficient CNN-Mamba hybrid network for lightweight semantic segmentation, optimized for minimizing model size and computational requirements. The main contributions of our paper are four folds:

• We firstly propose a novel lightweight Efficient CNN-Mamba Network (EMCNet) for for semantic segmentation. ECMNet utilizes U-shape encoder-decoder structure as backbone and regards the Feature Fusion Module (FFM) as a capsule network

- to capture global context information. Specially, FFM introduces SS2D block, a variant of Mamba, to learn long-range dependencies.
- We design a lightweight Enhanced Dual-Attention Block (EDAB) to extract multi-dimensional semantic information. EDAB consists of Dual-Direction Attention(DDA), Channel Attention (CA) and various convolution modules, realizing less model parameters and computational quantities.
- We develop a Multi-Scale Attention Unit (MSAU) to improve the representations ability of feature, which further refines the local details and global contextual information.
- ECMNet achieved 70.6% mIoU on the Cityscapes dataset on the single RTX 3090 GPU with only 0.87M of parameters, realizing the better trade-off between performance and parameters. Meanwhile, our proposed method achieved 73.6% of the highest performance on the CamVid dataset, which demonstrates the effectiveness and generalization of our proposed ECMNet.

2 Proposed Method

2.1 Overall Network Architecture



 $\textbf{Fig. 1} \ \ \textbf{The overall network architecture of Efficient CNN-Mamba Network (EMCNet)}$

As shown in Figure 1, the overall network architecture of our proposed EMCNet consists of four components: a CNN encoder improved with enhanced dual-attention

blocks, a CNN decoder with subtle difference from encoder, an efficient Mamba-based feature fusion module and three long skip connections enhanced with multi-scale attention unit. Specifically, the CNN-based encoder-decoder architecture extracts localized features for detailed spatial representation. The Mamba-based FFM can capture complex spatial information and long-range feature dependencies by state space model (SSM) to optimize global feature representations and computational complexity. The three long-distance skip connections generate more high-quality segmentation by focusing on low-level spatial information and high-level semantic information respectively. The above mentioned elaborated modules make it more efficient for ECMNet to fully integrate local and global feature information.

2.2 Enhanced Dual-Attention Block

The EDAB module is designed to focus different level feature information and keep network parameters as few as possible. Firstly, the input feature passes through a bottheneck structure that utilizes a 1×1 convolution to reduce the number of channels to half, significantly reducing the computational complexity and the number of parameters. Obviously, this will sacrifice a part of the accuracy, it will be more beneficial to introduce 3×1 convolution and 1×3 convolution more than make up for the loss at this point. Meanwhile, the two decomposed convolutions not only obtain a wider respective field for capturing a larger range of contextual feature information but also consider the model parameters and calculation complexity. The core of EDAB lies in its two-branch path, which captures local and global feature information respectively. Decompose convolution in one branch processes local and short-distance feature information, complemented by atrous convolution in the parallel branch for global feature integration. Then, the channel contains most of the feature information and the spatial feature information is key to enhance performance and suppress noise interference. Therefore, the two branches utilize Channel Attention (CA) and Dual-Direction Attention (DDA), which aims to build different attention matrix to learn multi-dimensional feature information and improve feature expression. Finally, the outputs from both designed pathes and intermediate features are integrated and processed by a 1×1 point-wise convolution to restore the original channel dimensionality. A channel shuffle strategy is applied at the end of EDBA to establish inter-channel correlations and overcome information fragmentation The detail operation is shown as follows:

$$F_{up_branch} = Conv_{1\times 3}(Conv_{3\times 1}(Conv_{1\times 1}(x))), \tag{1}$$

$$F_{mid_branch_1} = Conv_{CA}(Conv_{1\times3,D}(Conv_{3\times1,D}(F_{up_branch}))), \tag{2}$$

$$F_{mid_branch_2} = Conv_{DDA}(Conv_{1\times3,D,R}(Conv_{3\times1,D,R}(F_{up_branch}))), \tag{3}$$

$$Y_1 = Conv_{CS}(f_{1\times 1}(F_{up_branch} + F_{mid_branch_1} + F_{mid_branch_2}) + x), \tag{4}$$

where x denotes the input of the EDAB, Y_1 denotes the output feature map of the EDAB, and $Conv_{k\times k}(\cdot)$ are normal convolution operation. Among the suffix, D denotes depth-wise convolution, R is the atrous rates of atrous convolution, CA represents Channel Attention, DDA represents Dual-Direction Attention and CS denotes the shuffle operation of channel.

2.3 Multi-Scale Attention Unit

On the one hand, lower layers preserve fine spatial details with limited semantics, on the other hand, higher layers offer strong semantic representation at lower spatial resolution. Therefore, it is an efficient strategy to combine the low-level rich spatial information and high-level rich semantic information for semantic segmentation tasks. Inspired by U-Net, we use the same resolution connections to integrate the high-level feature maps and low-level feature maps. In order to better process the three long connections, we design a Multi-Scale Attention Unit (MSAU) to enhance the ability of feature representation. The MSAU is carried out from two branches, one is the Multi-Scale Spatial Aggregation, the other is the Channel Aggregation.

In the Multi-Scale Spatial Aggregation, the input feature map is utilized 1×1 convolution to convert from C channel to C/2 channel. In order to reduces the amount of parameter and computation while retaining the ability of multi-scale feature extraction, the next feature map goes through different sizes of depth-separable convolution, such as 3×3 , 5×5 and 7×7 . Meanwhile, the outputs of different scale convolutions obtain multi-scale feature information enhancing the multi-scale perception capability of the model. Then, the multi-scale fused feature map compresses the height dimension to 1 by adaptive average pooling, and generates a spatial attention map by 7×7 depth separable convolution, 1×1 convolution and Sigmoid activation function. At the same time, by multiplying with the multi-scale fused feature map, the processed feature highlights the key spatial regions and suppresses the irrelevant information. At last, the channel of model is converted from C/2 back to C by using 1x1 convolution, and the attention map reflects the importance of the different locations of feature map. For channel aggregation, the input feature map uses average pooling and maximum pooling to obtain average channel features and maximum channel features respectively, which captures channel statistics from different angles. The MSAU multiplies the spatial and channel aggregation results and adds them with the original input feature maps to obtain the output feature maps.

This design allows the MSAU module to fused the low-level spatial information to the high-level semantic information more effectively, and further enhance the ability of feature expression The detail operation can be defined as:

$$X_{1} = Conv_{(3\times3)}(Conv_{(1\times1)}(x)) + Conv_{(5\times5)}(Conv_{(1\times1)}(x)) + Conv_{(7\times7)}(Conv_{(1\times1)}(x))$$
(5)

$$X_2 = Conv_{(1\times 1)}(X_1 \otimes Sigmoid(Conv_{(7\times 7)}(Pool(x))))$$
 (6)

$$X_3 = Conv_{(1\times1)}(ReLU(Conv_{(1\times1)}(AvgPool(Conv_{(3\times3)}(x)))))$$
 (7)

$$X_4 = Conv_{(1\times1)}(ReLU(Conv_{(1\times1)}(MaxPool(Conv_{(3\times3)}(x)))))$$
(8)

$$Y_2 = x + (X_2 \otimes (X_3 + X_4)) \tag{9}$$

where x denotes the input of the MSAU and Y_2 represents the output feature map of the MSAU. Among the formulas, $Conv_{k\times k}(\cdot)$ are normal convolution operation. \otimes denotes element-by-element multiplication, $Pool(\cdot)$ denotes the adaptive average pooling, $AvgPool(\cdot)$ is average pooling, $MaxPool(\cdot)$ is maximum pooling, $ReLU(\cdot)$ is rectified linear unit and $Sigmoid(\cdot)$ is the Sigmoid activation function.

2.4 Feature Fusion Module

Motivated by by the effectiveness of Mamba in linear-complexity sequence modeling, we design a Feature Fusion Module (FFM) by introduce 2D-Selective-Scan (SS2D) block for better capturing global representations with less network parameters and computational quantities. The FFM enriches the feature diversity by integrating different scale feature information from the multi-level the MSAU and the encoder through the concatenation operation. Then, the SS2D block further extracts and fuses the features through a series of linear transformations and 2D convolution operations, which employs a selective scanning mechanism to enhance the feature representation ability. Finally, Feed-Forward Network (FFN) performs a nonlinear transformation to adjust the weight distribution of features, highlighting the key features and suppressing the redundant information, so as to improve performance in handling complex tasks. The designed FFM can effectively fuse multi-scale features and capture both local detail information and overall semantic features, great improving the performance of the model in semantic segmentation tasks. The complete operation is shown as follows:

$$X_{FFN} = FFN(SS2D(Concat(x_{encoder}, x_{MSAU1}, x_{MSAU2})))$$
(10)

$$Y_3 = X_{FFN} + x_{encoder} (11)$$

where $x_{encoder}, x_{MSAU1}, x_{MSAU2}$ denotes the out of the Encoder and MSAU resectively, Y_3 denotes the output feature map of the FFM. Among the formulas, $Concat(\cdot)$ is normal concatenation operation. $SS2D(\cdot)$ is the 2D-Selective-Scan block and $FFN(\cdot)$ is the eed-Forward Network.

3 Experiments

3.1 Datasets

- Cityscapes. This dataset is composed of high-quality 5,000 images, annotated at the pixel level. The images are primarily scenes of driving within urban settings, captured across 50 different cities with a resolution of 2,048×1,024. The dataset was divided into training sets(2,975 images), validation sets (500 images), and test sets (1,525 images)
- CamVid. The CamVid dataset, developed by the University of Cambridge, contains urban road scene images captured from a driving perspective (960×720 resolution). Its 700+ annotated samples support supervised learning, featuring 11 representative object classes that effectively capture urban road elements. This diversity in objects and well-annotated classes makes it particularly suitable for our segmentation accuracy research.

3.2 Ablation Studies

We design a series of ablation experiments to validate the effectiveness of each module in our proposed model. As shown in Figure 2, The baseline model used for comparison is structured as simple U-shape type, including the standard Encoder and Decoder.

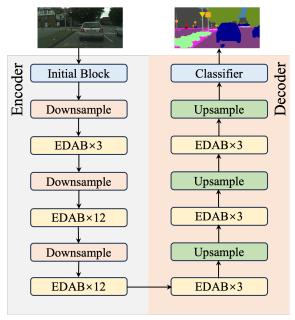


Fig. 2 The simple structure of the baseline model

The Encoder and Decoder consist of multiple lightweight enhanced dual-attention blocks (EDABs), which are modeled to achieve an average mIoU of 69.92% on the Camvid validation set.

In the long connection ablation experiments (A Group), the effect of gradually adding Line 1, Line 2, and Line 3 is investigated. The observed 0.61% enhancement after adding Line 1 substantiates that shallow information effectively aids semantic feature information reconstruction. Meanwhile, With three long skip connections, the model achieves a 1.29% mIoU enhancement. These results further demonstrate the significance of long-range skip connections for semantic segmentation task. In the MSAU ablation experiments (B group), the MSAU module is added gradually in the long connection. A comparison between B1 and A1 reveals that adding the MSAU module to long connections only adds 9.43K parameters, but improves the performance by 0.92% of mIoU. In the last ablation experiments (C Group), the introduction of the Feature Fusion Module (FFM) improves the performance of the model by 1.11% of mIoU. Finally, as the finalized architecture (C3), our proposed ECMNet improves performance by 3.7% mIoU compared to the baseline model. All the above experiments shown in Table 1 fully validate the efficacy of our proposed modules and strategies

3.3 Comparisons with SOTA Methods

In this section, we compare state-of-the-art semantic segmentation methods in recent years on the Cityscapes and CamVid datasets to verify that our approach achieves a better balance between performance and parameters. Our evaluation is based on three key metrics: model parameters, floating-point operations (FLOPs) and mIoU.

Table 1 Extensive ablation study for the proposed ECMNet on Camvid dataset.

			Metho	od						
Architecture	Lo	ng	Connection	Μ	\mathbf{S}	\U	FFM	Parameter $(K)\downarrow$	$FLOPs (G) \downarrow$	mIoU (%)↑
	1	2	3	1	2	3				
Baseline	_	-	-	-	-	-	-	775.57	7.56	69.92
$\mathbf{A1}$	\checkmark	-	-	-	-	-	-	777.93	7.57	$70.53^{\circ .61\uparrow}$
$\mathbf{A2}$	\checkmark	✓	-	-	-	-	_	796.41	7.64	$70.92^{1.00\uparrow}$
$\mathbf{A3}$	\checkmark	✓	\checkmark	-	-	-	-	805.67	7.79	$71.21^{1.29\uparrow}$
B 1	-	-	-	\checkmark	-	-	_	787.34	7.57	$71.45^{1.53\uparrow}$
$\mathbf{B2}$	-	-	-	✓	\checkmark	-	-	805.82	7.67	$72.65^{2.73\uparrow}$
$\mathbf{B3}$	-	-	-	\checkmark	\checkmark	\checkmark	_	815.08	7.90	$73.22^{3.30\uparrow}$
C1	-	-	-	-	-	-	\checkmark	827.80	7.80	$70.75^{0.83\uparrow}$
C2	\checkmark	✓	\checkmark	-	-	-	\checkmark	863.93	8.06	$71.03^{1.11\uparrow}$
C3 (ours)	\checkmark	✓	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	871.11	8.27	$73.62^{3.70\uparrow}$

A, B, C denote the long connection, the feature enhancement and the feature fusion respectively. No. of A, B and C denotes the the stack of same or different modules.

Evaluation Results on Cityscapes Dataset. As shown in Table 2, the model with a larger number of parameters and computation obviously achieves excellent segmentation results. However, computational complexity of model is high and its operation speed is slow, which is unsuitable for real-time intelligent embedded devices. In contrast, lightweight models like NDNet[24], CGNet[25], CFPNet[19], LEDNet[18] and LETNet[20] are computationally efficient, which lack overall performance, especially in accuracy. Obviously, the LBN-AA[42] achieved the highest mIoU with 6.2M model parameters which are far more than our proposed approach. Meanwhile, the ESPNet[23] utilized the least parameters to realize 60.3% mIoU, which is significantly lower than the performance of our method. Our proposed ECMNet only with a 0.87M parameters achieved a higher 70.6% mIoU. Our proposed ECMNet can get better segmentation results with less model parameters, which benefits from well-design structure, and the utilization of the Mamba. These results fully demonstrate that our proposed model can achieve a excellent balance between model parameters and performance.

Evaluation Results on CamVid Dataset. As shown in Table 3, to further verify the effectiveness and generalization capacity of our proposed ECMNet, we conducted comparative experiments with our method and other lightweight models on the CamVid dataset. Obviously, the MGSeg[38] just achieved the 72.7% mIoU with 13.3M model parameters which is lower performance and larger model parameters compared our proposed method. Therefore, our method has achieved the best accuracy by only using 0.87M parameters. Compared to Cityscapes, the higher overall performance on the CamVid dataset is due to our designed modules and strategies, which better capture feature of small size datasets. Per-class results are detailed in Table 4 further demonstrate the advantages of our proposed ECMNet.

Table 2 Performance comparison of our proposed ECMNet and other representative methods on the Cityscapes dataset.

Method	Year	$\begin{array}{c} {\rm Resolution} \\ {\rm (width} \times {\rm height)} \end{array}$	Backbone	Parameter (M)↓	FLOPs (G)↓	$\begin{array}{c} {\rm Speed} \\ {\rm (FPS)} \uparrow \end{array}$	mIoU (%)↑
SegNet[21]	2017	640×360	VGG-16	29.50	286.0	17	57.0
ENet[22]	2016	512×1024	No	0.36	3.8	135	58.3
ESPNet[23]	2018	512×1024	ESPNet	0.36	_	113	60.3
NDNet[24]	2021	512×1024	No	0.50	3.5	120	61.1
CGNet[25]	2021	360×640	No	0.49	6.0	-	64.8
ADSCNet[26]	2020	512×1024	No	_	-	77	67.5
ERFNet[27]	2017	512×1024	No	2.10	_	42	68.0
BiseNet-v1[28]	2018	768×1536	Xception	5.80	14.8	106	68.4
ICNet[29]	2018	1024×1024	PSPNet-50	26.50	28.3	30	69.5
DABNet[30]	2019	1024×2048	No	0.76	42.4	28	70.1
CFPNet[19]	2021	1045×2048	No	0.55	_	30	70.1
FPENet[31]	2019	512×1024	No	0.40	12.8	55	70.1
LEDNet[18]	2019	512×1024	No	0.94	_	40	70.6
DFANet[32]	2019	1024×1024	Xception	7.80	3.4	100	71.3
STDC1-50[33]	2021	512×1024	_	8.40	_	87	71.9
SegFormer[10]	2021	512×1024	MiT-B0	3.80	17.7	48	71.9
MSCFNet[34]	2022	512×1024	No	1.15	17.1	50	71.9
FPANet[35]	2022	512×1024	-	14.10	_	_	72.0
MLFNet[36]	2023	512×1024	ResNet-34	13.03	15.5	72	72.1
BiseNet-v2[37]	2021	512×1024	Xception	3.40	21.2	156	72.6
MGSeg[38]	2021	1024×1024	ShuffleNet-v2	4.50	16.2	101	72.7
PCNet[39]	2022	1024×2048	Scratch	3.40	21.2	156	72.6
LETNet[20]	2023	512×1024	No	0.95	13.6	150	72.8
SCTNet-S[40]	2024	512×1024	No	4.6	451.2	160.3	72.8
HSB-Net[41]	2021	512×1024	ResNet-34	12.10	_	124	73.1
LBN-AA[42]	2021	448×896	No	6.20	49.5	51	73.6
ECMNet (Ours)	_	1024×1024	No	0.87	8.27	43	70.6

The gray box denotes the best value of the current metric.

4 Conclusion

In this study, we proposed a lightweight semantic segmentation network that combines Mamba and Convolutional Neural Networks (CNNs). We fused the local feature extraction capability of convolutional neural networks with long-range dependencies of Mamba to model. Specifically, we introduced a Feature Fusion Module (FFM) as a capsule-based framework in the middle of the model, which can better capture global feature information. Additionally, an Enhanced Dual Attention Module (EDAB) designed in the convolutional neural network learned more local feature information while ensuring simplicity and lightweight. Meanwhile, in order to compensate for the local feature information lost by CNNs, multi-scale long connections are utilized in the model. Moreover, We design a Multi-Scale Attention Unit (MSAU) for cross-layer connections, effectively boosting discriminative features and attenuating

 ${\bf Table~3}~{\rm Performance~comparison~of~our~proposed~ECMNet~and~other~representative~methods~on~the~CamVid~dataset.}$

Method	Year	Resolution	Backbone	$\begin{array}{c} \mathrm{Parameter} \\ \mathrm{(M)} \downarrow \end{array}$	$\begin{array}{c} \text{Speed} \\ (\text{FPS}) \downarrow \end{array}$	mIoU (%)↑
ENet[22]	2016	360×480	No	0.36	68	51.3
SegNet[21]	2017	360×480	VGG-16	29.45	87	55.6
NDNet[24]	2021	360×480	No	0.50	78	57.2
DFANet[32]	2019	360×480	Xception	7.80	120	64.7
DABNet[30]	2019	360×480	No	0.76	136	66.4
FDDWNet[43]	2020	360×480	No	0.80	79	66.9
ICNet[29]	2018	720×960	PSPNet-50	26.50	28	67.1
FBSNet[44]	2023	360×480	No	0.62	120	68.9
MSCFNet[34]	2022	360×480	No	1.15	110	69.3
LETNet[20]	2023	360×480	No	0.95	21	70.5
HAFormer[45]	2024	360×480	No	0.60	118	71.1
MGSeg[38]	2021	736×736	ResNet-18	13.3	127	72.7
ECMNet (Ours)	-	360×480	No	0.87	55	73.6

The gray box denotes the best value of the current metric.

noise. Extensive experimental results demonstrate that our proposed model achieves an excellent balance between model scale and performance.

Table 4 Performance comparison of our proposed ECMNet and the state-of-arts lightweight methods about per-class results on the CamVid dataset.

Method	Roa	Sid Bui	Bui	Wal	Fen	Pol	TLi	TSi	Veg	Ter	Sky	Ped	Rid	Car	Tru	Bus	Tra	Mot	Bic	mIoU(%)
Enet[22]	96.3	74.2	75.0	32.2	33.2	43.4	34.1	44.0	88.6	61.4	9.06	65.5	38.4	9.06	36.9	50.5	48.1	38.8	55.4	58.3
ESPNet[23]	97.0	77.5	76.2	35.0	36.1	45.0	35.6	46.3	8.06	63.2	92.6	67.0	40.9	92.3	38.1	52.5	50.1	41.8	57.2	60.3
CGNet[25]	95.5	78.7	88.1	40.0	43.0	54.1	59.8	63.9	9.68	9.79	92.9	74.9	54.9	90.2	44.1	59.5	25.2	47.3	60.2	64.8
ESPNet-v2[46]	97.3	78.6	88.8	43.5	42.1	49.3	52.6	0.09	90.5	8.99	93.3	72.9	53.1	91.8	53.0	62.9	53.2	44.2	59.9	66.2
ERFNet[27]	7.76	81.0	8.68	42.5	48.0	56.3	59.8	65.3	91.4	68.2	94.2	8.92	57.1	92.8	50.8	60.1	51.8	47.3	61.7	0.89
DABNet[30]	8.96	78.5	6.06	45.4	50.2	59.1	65.2	8.02	92.5	68.2	94.6	80.5	58.5	92.7	52.7	67.2	50.9	50.4	65.7	70.0
CFPNet[19]	8.76	81.4	90.5	46.4	50.6	56.4	61.5	2.79	92.1	68.9	94.3	80.4	2.09	93.9	51.4	0.89	50.8	51.2	67.7	70.1
${ m LEDNet}[18]$	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	6.06	64.4	64.0	52.7	44.4	71.6	9.02
ECMNet (Ours) 97.1 80.8 90.9	97.1	80.8	6.06	44.2	53.4	8.09	61.9	72.4	91.7	60.4	93.9	75.2	52.7	92.0	65.3	76.5	66.5	37.8	69.1	5.07

The gray box denotes the best mIoU of the current class. Roa, Sid, Bui, Wal, Fen, Pol, TLi, TSi, Veg, Ter, Sky, Ped, Rid, Car, Tru, Mot and Bic reprsent Road, Sidewalk, Building, Wall, Fence, Pole, Traffic Light, Traffic Sign, Vegtation, Terrain, Sky, Pedestrain, Rider, Car, Truck, Motorcycle and Bicycle respectively.

References

- [1] Sanchez, J., Deschaud, J.-E., Goulette, F.: Cola: Coarse-label multi-source lidar semantic segmentation for autonomous driving. IEEE Transactions on Robotics (2025)
- [2] Jing, W., Zhang, W., Di, D., Li, C., Emam, M., Mian, A.: Hypergraph biformer for semantic segmentation of high-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing (2025)
- [3] Luo, Z., Yang, W., Yuan, Y., Gou, R., Li, X.: Semantic segmentation of agricultural images: A survey. Information Processing in Agriculture 11(2), 172–186 (2024)
- [4] Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters-improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4353–4361 (2017)
- [5] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40(4), 834–848 (2017)
- [6] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
- [7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
- [8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
- [9] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)
- [10] 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)

- [11] Ren, S., Li, X.: Hresformer: Hybrid residual transformer for volumetric medical image segmentation. IEEE Transactions on Neural Networks and Learning Systems (2025)
- [12] Gao, Y., Zhang, J., Wei, S., Li, Z.: Pformer: An efficient cnn-transformer hybrid network with content-driven p-attention for 3d medical image segmentation. Biomedical Signal Processing and Control 101, 107154 (2025)
- [13] Garbaz, A., Oukdach, Y., Charfi, S., El Ansari, M., Koutti, L., Salihoun, M.: Dmfc-uformer: Depthwise multi-scale factorized convolution transformer-based unet for medical image segmentation. Biomedical Signal Processing and Control 101, 107200 (2025)
- [14] Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
- [15] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. Advances in neural information processing systems 37, 103031–103063 (2024)
- [16] Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 578–588 (2024). Springer
- [17] Liu, M., Dan, J., Lu, Z., Yu, Y., Li, Y., Li, X.: Cm-unet: Hybrid cnn-mamba unet for remote sensing image semantic segmentation. arXiv preprint arXiv:2405.10530 (2024)
- [18] 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 (2019). IEEE
- [19] Ding, L., Jiang, H., Xu, R., Huang, R.: Cfpnet: Improving lightweight tof depth completion via cross-zone feature propagation. arXiv preprint arXiv:2411.04480 (2024)
- [20] Xu, G., Li, J., Gao, G., Lu, H., Yang, J., Yue, D.: Lightweight real-time semantic segmentation network with efficient transformer and cnn. IEEE Transactions on Intelligent Transportation Systems 24(12), 15897–15906 (2023)
- [21] 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)

- [22] 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)
- [23] 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)
- [24] Yang, Z., Yu, H., Fu, Q., Sun, W., Jia, W., Sun, M., Mao, Z.-H.: Ndnet: Narrow while deep network for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems **22**(9), 5508–5519 (2020)
- [25] 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)
- [26] Wang, J., Xiong, H., Wang, H., Nian, X.: Adscnet: asymmetric depthwise separable convolution for semantic segmentation in real-time. Applied Intelligence 50(4), 1045–1056 (2020)
- [27] 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)
- [28] Zhao, S., Wu, X., Tian, K.: Real-time semantic segmentation network based on improved bisenet v1. In: Proceedings of the 2022 11th International Conference on Networks, Communication and Computing, pp. 38–44 (2022)
- [29] 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)
- [30] Li, G., Yun, I., Kim, J., Kim, J.: Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. arXiv preprint arXiv:1907.11357 (2019)
- [31] Liu, M., Yin, H.: Feature pyramid encoding network for real-time semantic segmentation. arXiv preprint arXiv:1909.08599 (2019)
- [32] Li, H., Xiong, P., Fan, H., Sun, J.: Dfanet: Deep feature aggregation for realtime semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9522–9531 (2019)
- [33] Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z., Luo, J., Wei, X.: Rethinking bisenet for real-time semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9716–9725 (2021)

- [34] Gao, G., Xu, G., Yu, Y., Xie, J., Yang, J., Yue, D.: Mscfnet: A lightweight network with multi-scale context fusion for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems 23(12), 25489–25499 (2021)
- [35] Wu, Y., Jiang, J., Huang, Z., Tian, Y.: Fpanet: Feature pyramid aggregation network for real-time semantic segmentation. Applied Intelligence **52**(3), 3319–3336 (2022)
- [36] Fan, J., Wang, F., Chu, H., Hu, X., Cheng, Y., Gao, B.: Mlfnet: Multi-level fusion network for real-time semantic segmentation of autonomous driving. IEEE Transactions on Intelligent Vehicles 8(1), 756–767 (2022)
- [37] Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. International journal of computer vision 129, 3051–3068 (2021)
- [38] He, J.-Y., Liang, S.-H., Wu, X., Zhao, B., Zhang, L.: Mgseg: Multiple granularity-based real-time semantic segmentation network. IEEE Transactions on Image Processing 30, 7200–7214 (2021)
- [39] Lv, Q., Sun, X., Chen, C., Dong, J., Zhou, H.: Parallel complement network for real-time semantic segmentation of road scenes. IEEE Transactions on Intelligent Transportation Systems 23(5), 4432–4444 (2021)
- [40] Xu, Z., Wu, D., Yu, C., Chu, X., Sang, N., Gao, C.: Sctnet: Single-branch cnn with transformer semantic information for real-time segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 6378–6386 (2024)
- [41] Li, G., Li, L., Zhang, J.: Hierarchical semantic broadcasting network for real-time semantic segmentation. IEEE Signal Processing Letters 29, 309–313 (2021)
- [42] Dong, G., Yan, Y., Shen, C., Wang, H.: Real-time high-performance semantic image segmentation of urban street scenes. IEEE Transactions on Intelligent Transportation Systems **22**(6), 3258–3274 (2020)
- [43] Liu, J., Zhou, Q., Qiang, Y., Kang, B., Wu, X., Zheng, B.: Fddwnet: a lightweight convolutional neural network for real-time semantic segmentation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2373–2377 (2020). IEEE
- [44] Gao, G., Xu, G., Li, J., Yu, Y., Lu, H., Yang, J.: Fbsnet: A fast bilateral symmetrical network for real-time semantic segmentation. IEEE Transactions on Multimedia 25, 3273–3283 (2022)
- [45] Xu, G., Jia, W., Wu, T., Chen, L., Gao, G.: Haformer: Unleashing the power of hierarchy-aware features for lightweight semantic segmentation. IEEE Transactions on Image Processing (2024)

[46] Lin, S., Hao, X., Liu, Y., Yan, D., Liu, J., Zhong, M.: Lightweight deep learning methods for panoramic dental x-ray image segmentation. Neural Computing and Applications **35**(11), 8295–8306 (2023)