DDRNet-SDR: A Spatial-Frequency Refinement Network for Real-Time Semantic Segmentation

Shuhui Zhu, Jing Lu, Gang Shi(✉)

Xinjiang University, Urumqi 830017, China
shigang@xju.edu.cn

Abstract. Achieving dense scene parsing under real-time constraints demands a careful balance between prediction accuracy and runtime performance, posing a persistent challenge within the domain of computer vision. Although dual-resolution networks such as DDRNet achieve competitive performance, they are still limited by insufficient multi-scale context modeling, loss of high-frequency details caused by repeated downsampling, and suboptimal static feature fusion strategies. In response to the above limitations, this work builds upon a dual-resolution network architecture with targeted enhancements, termed DDRNet-SDR. Specifically, a Spatial-Frequency Fusion Module (SFFM) is introduced to exploit frequency-domain priors for attention generation, enabling effective spatial feature refinement and preservation of high-frequency information. In addition, a DWR-Conv module is designed to independently model high- and low-resolution branches, facilitating efficient multi-scale context encoding through multi-branch and multi-dilation structures. Furthermore, a dynamic synergistic supervision strategy that combines Online Hard Example Mining (OHEM) and Dice loss is adopted to balance pixel-level accuracy and region-level consistency. Experiments on Cityscapes show DDRNet-SDR reaches 78.43% mIoU at 66.1 FPS on a single 2080Ti GPU. Compared to the baseline, it gains 1.03% in accuracy while keeping real-time speed, which suits latency-sensitive tasks like autonomous driving and UAV vision.

Keywords: Real-time semantic segmentation, dual-resolution networks, spatial-frequency fusion module, multi-scale context modeling

1 Introduction

Semantic segmentation under real-time constraints remains a persistent challenge within the domain of computer vision, robotic perception, and UAV navigation. Achieving competitive accuracy while maintaining low inference cost remains difficult, as models optimized for performance tend to be computationally heavy, whereas lightweight alternatives often lose fine structural details and miss small objects.

To reconcile semantic context modeling with fine detail preservation, multi-scale feature integration has become a prevalent strategy in the community. Representative methods such as FPN [1] fuse hierarchical features via a top-down pathway. However,

low-resolution features tend to lose high-frequency details, while high-resolution features lack sufficient semantic information. Moreover, direct feature fusion often introduces misalignment and noise, limiting fine-grained representation.

We propose a dual-resolution network to handle these issues. DDRNet-SDR, following a "enhance-then-fuse" paradigm. Three main components form the backbone of this framework: a Spatial-Frequency Fusion Module (SFFM), which incorporates frequency-domain priors to strengthen structural feature representations; a DWR-Conv [2] module that strengthens multi-scale context modeling in dual branches; and a dynamic collaborative supervision strategy combining OHEM [3] and Dice loss [4] to balance pixel-level accuracy and structural consistency.

Benchmarking on the Cityscapes [5] dataset confirm that the proposed approach improves performance while keeping real-time efficiency.

2 Related Work

2.1 High-Accuracy Semantic Segmentation

Fully Convolutional Networks (FCN) [6] stand as a pioneering work in semantic segmentation, in which fully connected layers are substituted with convolutional ones, enabling dense pixel-level prediction in a fully integrated manner.

The encoder–decoder architectures such as U-Net [7] and its counterparts (e.g., U-Net++ [8], Attention U-Net [9]) significantly improve spatial detail recovery through skip connections. SegNet [10] further enhances computational efficiency by reusing pooling indices during upsampling.To improve contextual modeling, the DeepLab series [11,12] introduces atrous convolution and ASPP is employed to gather contextual representations across different spatial scales. More recently, such as ViT [13], Swin Transformer [14], and SegFormer [15] demonstrate strong capability in modeling long-range dependencies.

Despite their high accuracy, these methods typically rely on complex architectures and substantial computational overhead, which restricts their practical utility in real-time deployment scenarios.

2.2 Real-Time Semantic Segmentation

To meet online requirements, numerous lightweight and efficient segmentation models have been proposed. The dual-branch architecture can separate spatial detail modeling from semantic context modeling.

BiSeNet [16] and its improved version BiSeNetV2 [17] adopt separate spatial and context paths to balance accuracy and efficiency. In addition, lightweight networks such as ENet [18], ICNet [19], and Fast-SCNN [20] achieve fast inference through simplified architectures.

Subsequent works focus on improving feature interaction and efficiency. SwiftNet [21] leverages lightweight pretrained backbones, CABiNet [22] enhances context aggregation, and SFNet [23] introduces semantic flow for better cross-resolution align-

ment. DDRNet [24] shares early-stage features and performs repeated bidirectional fusion between dual-resolution branches, achieving an excellent balance between performance and efficiency.

However, existing methods still face two key limitations: (1) significant representation gaps between multi-resolution features lead to inefficient fusion, and (2) most interaction mechanisms rely on simple operations, limiting feature collaboration. Therefore, designing more effective and efficient cross-resolution feature enhancement and fusion mechanisms remains an important research direction.

3 Method

3.1 Overall Architecture

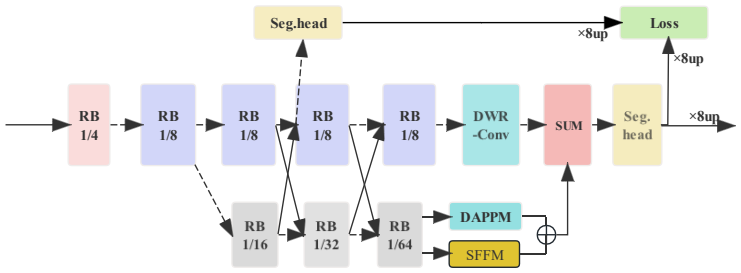

Fig. 1. Full structure of the proposed DDRNet-SDR.

Fig. 1 show the whole design of the proposed spatial-frequency collaborative refinement network, termed DDRNet-SDR, built upon the efficient dual-resolution framework of DDRNet and incorporating three main components.

(1) Multi-scale feature extraction and semantic enhancement: A dual-stream structure built on residual blocks (RB) serves as the backbone of the network. Spatial detail is preserved in the high-resolution stream, whereas the low-resolution stream captures global context through progressive downsampling. The DWR-Conv module is placed at the end of the semantic flow to enhance multi-scale context modeling and reduce semantic fragmentation.

(2) Spatial-frequency fusion module (SFFM): To mitigate information loss during feature fusion, we introduce an SFFM that models high-frequency components using frequency-domain priors and generates spatial attention to guide feature aggregation. A DWR-Conv is further embedded in the spatial refinement path, where frequency-guided multi-scale dilated convolutions enhance high-frequency responses and improve fine-grained structural representation.

(3) Global reconstruction and dynamic supervision: Fused features pass through a segmentation head that projects them into pixel space for final predictions. We apply a dynamic collaborative supervision strategy that merges Online Hard Example Mining (OHEM) with Dice loss. This joint optimization directs focus to challenging regions while enhancing boundary precision and structural integrity.

3.2 Spatial-Frequency Fusion Module

In two-branch real-time semantic segmentation architectures, the semantic branch obtains global context through downsampling, which acts like a low-pass filter and causes the loss of high-frequency structures such as thin lane markings. Conventional fusion in the spatial domain, such as concatenation or element-wise addition, tends to fall short in recovering such fine-grained details.

To handle this problem, we propose the Spatial-Frequency Fusion Module (SFFM), which models features in multiple dimensions to jointly represent global semantics and local geometry. SFFM leverages frequency-domain high-frequency priors and constructs two parallel branches for channel recalibration and spatial refinement. For an input feature map, the module sequentially performs high-frequency information generation, channel modeling, and spatial refinement, as illustrated in Fig. 2.

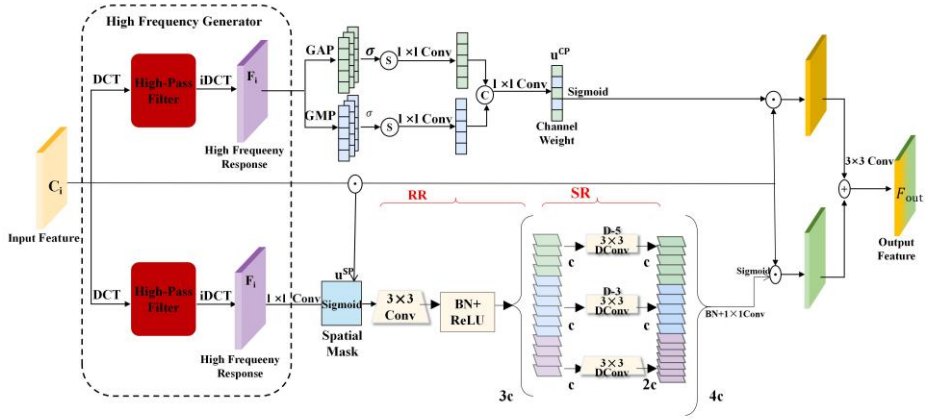

Fig. 2. Spatial-Frequency Fusion Module (SFFM).

High-Pass Filter Module (HPF). A High Frequency Generator is first employed to extract high-frequency response maps from the input feature representation. These maps are subsequently used to derive two complementary attention weights. To effectively enhance tiny objects—a task often compromised in conventional architectures—a specialized High-Pass Filter (HPF) is designed, as illustrated in Fig. 3. This design is predicated on the Discrete Cosine Transform (DCT) spectrum principle: low-frequency components (global structures) cluster in the top-left corner, while essential high-frequency components (fine details) distribute toward the bottom-right. We introduce a hyper-parameter α (set to 0.25) to regulate the cut-off frequency. As show in Fig. 3, α directly determines the spatial extent of the top-left region in the filter that is explicitly masked to zero. This zeroed region effectively suppresses the low-frequency responses. Specifically, the values of the filter Z are formally defined by the following formulation:

$$z(u,v) = \begin{cases} 0, u < \alpha H, v < \alpha W \\ 1, otherwise \end{cases}, \alpha \in [0,1] \quad (1)$$

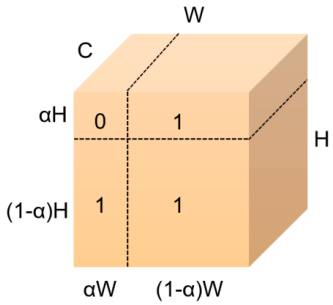

Fig. 3. Structure of the High-Pass Filter (HPF) module.

Channel Attention Generation. The channel attention stream highlights feature channels critical for small-scale objects by leveraging high-frequency responses to suppress background interference. Global Average Pooling (GAP) and Global Max Pooling (GMP) are applied in parallel to obtain robust channel descriptors, whose outputs are convolved, concatenated, and re-convolved to capture inter-channel dependencies. A Sigmoid activation then produces channel-wise attention weights that scale the input features point-wise, suppressing noise while enhancing informative channels.

The overall process of this branch can be formulated as:

$$C_i^{'} = C_i \odot u^{cp} \quad (2)$$

Spatial Attention Generation. The spatial path aims to highlight regions corresponding to small-scale objects and refine their spatial representations.

The process for this stream are defined by the following equations:

$$F_{refined} = f_{DSR}(C_i \odot u^{SP}) \quad (3)$$

$$F_{sr}^{'} = F_{refined} \odot C_i \quad (4)$$

Feature Fusion and Output Integration. The spatial attention branch highlights small-scale target regions and refines their spatial representations by generating a spatial mask for element-wise fusion with input features—enhancing key structures while suppressing background noise. Enhanced features then pass through cascaded Residual Refinement (RR) and multi-branch atrous Spatial Refinement (SR) blocks to capture multi-scale context and robustly model target structures. Branch outputs concatenate, undergo BN-conv-Sigmoid processing, and multiply point-wise with the original input to recalibrate the feature response.

The last stage in this process is expressed as:

$$F_{out} = Conv3 \times 3(C_i^{'} + F_{sr}^{'}) \quad (5)$$

3.3 DWR-Conv for High-Resolution Feature Refinement

In dual-resolution networks, the high-resolution branch preserves fine details but has a limited receptive field, hindering large-scale context capture. To boost multi-scale representation, we introduce a lightweight multi-dilation convolution module, DWR-Conv, replacing standard residual blocks in this branch.

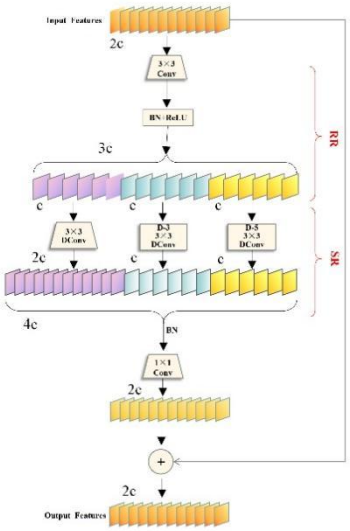

Fig. 4. Structure of the proposed DWR-Conv module.

For an input feature map $X \in R^{H \times W \times C}$, the processing flow of DWR-Conv is described as follows:

(1) Bottleneck Reduction: 3×3 a convolutional layer is employed to compress the channel dimension from C to $C/2$,improving computational efficiency.

(2) Multi-scale Context Extraction: The reduced features are fed into parallel axis separable convolution branches with varying dilation rates (e.g., $d \in \{1,3,5\}$) to gather representations across varying spatial scales.

(3) Feature Fusion and Restoration: Outputs from all branches concatenate along the channel axis and fuse via a 1×1 convolution, which also restores the channel dimension to C .

(4) Residual Connection: The merged features are combined with the input via element-wise summation feature X to preserve original spatial information while enhancing representation.

By leveraging parallel multi-dilation structures, DWR-Conv effectively enlarges the receptive field while maintaining high-resolution representations with low computational overhead.

3.4 Design of the Total Objective Loss Function

Main Loss Function Design. The main loss function incorporates OHEM cross-entropy loss and Dice loss in a weighted manner, balancing pixel-wise classification and regional structural coherence. It is defined as:

$$L_{main} = \lambda_1 \cdot L_{ohem} + \lambda_2 \cdot L_{dice} \tag{6}$$

Among them, L_{ohem} dynamically focuses on hard samples with large prediction errors, thereby enhancing the segmentation performance in boundary and ambiguous regions while alleviating class imbalance. In contrast, L_{dice} optimizes the intersection between predictions and ground truth labels, improving structural continuity and integrity of elongated objects.

These two loss functions have different optimization objectives. λ_1 is fixed to 1, while λ_2 is treated as the key hyperparameter and searched within the range of [0.1,0.9]. The coefficients are not subject to normalization constraintst.

The work achieves peak performance at λ_2=0.5, with an mIoU of 78.43% on the validation set. This indicates that an appropriate Dice loss weight effectively enhances structural consistency while maintaining pixel-level classification accuracy.

Auxiliary Supervision Mechanism. To enhance high-resolution feature representation and mitigate potential gradient vanishing in deep network training, an auxiliary segmentation head with an auxiliary loss is introduced at an intermediate stage. The auxiliary loss, formulated as standard cross-entropy, provides direct supervision to shallow spatial features. Its weight is fixed at 0.4, following the empirical setting in PSPNet [25], ensuring effective gradient guidance without interfering with the primary loss optimization.

$$L_{total} = L_{main} + \gamma \cdot L_{aux} \tag{7}$$

Through this joint optimization strategy [27,27], the model strikes a trade-off between pixel-wise accuracy and regional structural coherence, while auxiliary supervision further enhances high-resolution feature learning.

4 Experiments

4.1 Datasets

We test DDRNet-SDR on Cityscapes and CamVid datasets. Cityscapes offers high-resolution street views from 50 cities for 19 classes, ideal for fine-detail recovery in complex scenes. CamVid has 701 annotated images across 32 classes to check generalization under varied lighting and views.

4.2 Benchmarking Against Leading Methods on Cityscapes

The present study benchmarks DDRNet-SDR against DDRNet, BiSeNet V2, and SFNet, evaluating accuracy (mIoU) and computational performance (FPS/GFLOPs) for

UAV and urban road applications. As shown in Table 1, all GFLOPs are computed at 2048×1024 resolution, with †-marked methods accelerated via TensorRT.

Table 1. Performance and efficiency comparison on the Cityscapes benchmark.

Model	mIoU(%)		Speed	GPU	Resolution	GFLOPs	Params
	val	test	(FPS)				
BiSeNet1[16]	69.0	68.4	105.8	GTX 1080Ti	1536×768	14.8	5.8M
BiSeNet2[16]	74.8	74.7	65.5	GTX 1080Ti	1536×768	55.3	49M
BiSeNetV2† [17]	73.4	72.6	156	GTX 1080Ti	1024×512	21.1	-
BiSeNetV2-L† [17]	75.8	75.3	47.3	GTX 1080Ti	1024×512	118.5	-
Fast-SCNN [20]	68.6	68.0	123.5	TitanXp	2048×1024	-	1.1M
DFANet A [30]	-	71.3	100	TitanXp	1024×1024	3.4	7.8M
DFANet B [30]	-	67.1	120	TitanXp	1024×1024	2.1	4.8M
SwiftNetRN-18[21]	75.5	75.4	39.9	GTX 1080Ti	2048×1024	104.0	11.8M
SFNet(DF1)[23]	-	74.5	74	GTX 1080Ti	2048×1024	-	9.03M
SFNet(DF2)[23]	-	77.8	53	GTX 1080Ti	2048×1024	-	10.53M
SFNet(ResNet-18)[23]	-	78.9	18	GTX 1080Ti	2048×1024	247	12.87M
MSFNet* [31131]	-	71.3	117	GTX 2080Ti	1024×512	24.2	-
MSFNet [31]	-	77.1	41	GTX 2080Ti	2048×1024	96.8	-
LETNet [22]	72.8	72.8	150	GTX 2080Ti	512×1024	4.3	4.4M
SGCPNet[33]	69.5	69.3	178.5	GTX 2080Ti	1536×768	2.53	0.61M
DDRNet-23-Slim [24]	77.4	77.4	101.6	GTX 2080Ti	2048×1024	36.3	5.7M
DDRNet-23[24]	79.1	79.4	37.1	GTX 2080Ti	2048×1024	143.1	20.1M
DDRNet-39[24]	-	80.4	22.0	GTX 2080Ti	2048×1024	281.2	32.3M
DDRNet-SDR(Ours)	78.9	78.43	66.1	GTX 2080Ti	2048×1024	56.6	8.7M

Qualitative results in Fig. 5 indicate that DDRNet-SDR achieves better boundary delineation and fewer misclassifications on thin structures than DDRNet-23, demonstrating improved structural consistency. Although DDRNet-SDR yields slightly lower mIoU than DDRNet-23, it substantially reduces GFLOPs from 143.1 to 56.6, parameters from 20.1M to 8.7M, and increases FPS from 37.1 to 66.1, achieving a favorable trade-off between segmentation accuracy and computational efficiency for online deployment under constrained computational budgets.

4.3 Ablation Study

DDRNet23-slim (mIoU of 77.40%) is adopted as the baseline, and a set of ablation experiments are performed on the Cityscapes validation set. The results are presented in Table 2.

Table 2. Ablation studies of different modules.

Model Configuration	mIoU(%)
Baseline (DDRNet-23-slim)	**77.40**

Baseline + SFFM	**78.02**
Baseline + SFFM + DWR-Conv	**78.23**
DDRNet-SDR(Full Model)	**78.43**

(1) Effect of SFFM: Integrating the Spatial-Frequency Fusion Module (SFFM) yields a significant gain of +0.72% (reaching 78.02% mIoU). This verifies that restoring high-frequency priors effectively alleviates the information degradation introduced by spatial downsampling.

(2) Effect of DWR-Conv: Incorporating DWR-Conv further boosts the mIoU to 78.23% (+0.21%), validating its efficacy in multi-scale context modeling.

(3) Effect of Dynamic Loss: Finally, applying the Dynamic Collaborative Loss [20, 21] optimizes the final performance to 78.43% mIoU, an overall gain of 1.03% over the baseline.

These results demonstrate that our components work synergistically to enhance feature representation while maintaining real-time efficiency.

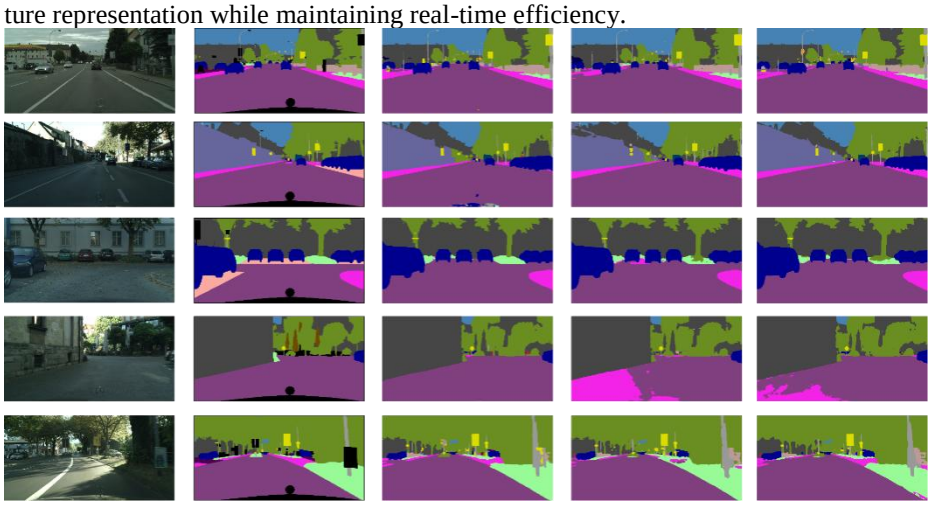

Fig. 5. Columns from left to right represent: (a) input image, (b) ground truth, (c) DDRNet-SDR (ours), (d) DDRNet-23, and (e) DDRNet-23-slim. The proposed model produces sharper object boundaries and fewer segmentation artifacts, particularly for small-scale targets including poles and traffic signs.

5 Conclusion

The present study introduces a compact network designed for real-time semantic parsing, DDRNet-SDR, which builds upon the DDRNet framework by introducing a frequency-aware feature fusion module (SFFM) and a dilated multi-scale convolution (DWR-Conv) to enhance the modeling of high-frequency details and fine-grained structures.

It should be noted that this work does not aim to surpass the original DDRNet-23 in all metrics, but rather to achieve a trade-off between segmentation performance and inference efficiency. Experiments on the Cityscapes benchmark dataset demonstrate that the proposed method achieves stable results, especially in challenging regions with complex textures and low contrast.

Furthermore, the SFFM module incorporates frequency domain prior information to supplement spatial feature learning, thereby effectively enhancing structure awareness capabilities. Future research will explore more effective frequency domain modeling methods and combine them with a compact architecture to achieve more stable performance in various practical application scenarios.

Acknowledgements. This work was supported by the Key Research and Development Project in Xinjiang Uygur Autonomous Region (No.2022B01006).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Lin, T.Y., Dollár, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2117–2125 (2017)
2. Wei, H., Liu, X., Xu, S., et al.: DWRSeg: Rethinking efficient acquisition of multi-scale contextual information for real-time semantic segmentation. arXiv preprint arXiv:2212.01173 (2022)
3. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 761–769 (2016)
4. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
5. 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 (CVPR), pp. 3213–3223 (2016)
6. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440 (2015)
7. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241. Springer, Cham (2015)
8. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., et al.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis, pp. 3–11. Springer, Cham (2018)
9. Oktay, O., Schlemper, J., Le Folgoc, L., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
10. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39(12), 2481–2495 (2017)
11. Chen, L.C., Papandreou, G., Kokkinos, I., et al.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40(4), 834–848 (2017)

12. Chen, L.C., Zhu, Y., Papandreou, G., et al.: Encoder-decoder with atrous separable convolution for semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Liu, Z., Lin, Y., Cao, Y., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022 (2021)
15. Xie, E., Wang, W., Yu, Z., et al.: SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34, 12077–12090 (2021)
16. Yu, C., Wang, J., Peng, C., et al.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 325–341 (2018)
17. Yu, C., Gao, C., Wang, J., et al.: BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation. Int. J. Comput. Vis. 129(11), 3051–3068 (2021)
18. Paszke, A., Chaurasia, A., Kim, S., et al.: ENet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147 (2016)
19. Zhao, H., Qi, X., Shen, X., et al.: ICNet for real-time semantic segmentation on high-resolution images. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 405–420 (2018)
20. Poudel, R.P.K., Liwicki, S., Cipolla, R.: Fast-scnn: Fast semantic segmentation network. arXiv preprint arXiv:1902.04502 (2019)
21. Orsic, M., Kreso, I., Bevandic, P., et al.: 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 (CVPR), pp. 12607–12616 (2019)
22. Kumaar, S., Lyu, Y., Nex, F., et al.: Cabinet: Efficient context aggregation network for low-latency semantic segmentation. arXiv preprint arXiv:2011.00993 (2020)
23. Li, X., You, A., Zhu, Z., et al.: Semantic flow for fast and accurate scene parsing. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 775–793 (2020)
24. Pan, H., Hong, Y., Sun, W., et al.: Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. IEEE Trans. Intell. Transp. Syst. 24(3), 3448–3460 (2023)
25. Zhao, H., Shi, J., Qi, X., et al.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
26. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7482–7491 (2018)
27. Chen, Z., Badrinarayanan, V., Lee, C.Y., et al.: Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: International Conference on Machine Learning (ICML), pp. 794–803. PMLR (2018)
28. Fan, M., Lai, Z., Huang, J., et al.: Rethinking BiSeNet for real-time semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9716–9725 (2021)
29. Romera, E., Alvarez, J.M., Bergasa, L.M., et al.: ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation. IEEE Trans. Intell. Transp. Syst. 19(1), 263–272 (2018)
30. Li, H., Xiong, P., Fan, H., et al.: DFANet: Deep feature aggregation for real-time semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9522–9531 (2019)

31. Si, H., Zhang, Z., Lv, F., et al.: Real-time semantic segmentation via multiply spatial fusion network. arXiv preprint arXiv:1911.07217 (2019)
32. Xu, G., Li, J., Gao, G., et al.: Lightweight real-time semantic segmentation network with efficient transformer and CNN. IEEE Transactions on Intelligent Transportation Systems, 24(12), 15897–15906 (2023).
33. Hao, S., Zhou, Y., Guo, Y., et al.: Real-time semantic segmentation via spatial-detail guided context propagation. IEEE Transactions on Neural Networks and Learning Systems, 2022.