SUPPLEMENTARY MATERIAL FOR Efficient Parallel Multi-Scale Detail and Semantic Encoding Network for Lightweight Semantic Segmentation

Xiao Liu
Sichuan university
Chengdu, China
liux@stu.scu.edu.cn

Xiuya Shi
Sichuan university
Chengdu, China
shixiuya@stu.scu.edu.cn

Lufei Chen
Sichuan university
Chengdu, China
chenlufei@stu.scu.edu.cn

Linbo Qing
Sichuan university
Chengdu, China
qing_lb@scu.edu.cn

Chao Ren*
Sichuan university
Chengdu, China
chaoren@scu.edu.cn

A OVERVIEW

Here we provide more experiment results and analysis for PMSDSEN. The appendix document is arranged as follows:

(1) The specific implementation details are described in the Section B and Table 1.

(2) The per-class IoU (%) results on the Cityscapes test set are displayed in the Section C and Figure 1.

(3) More visual results on Cityscapes and CamVid dataset are displayed in the Figure 2 and Figure 3.

(3) More ablation experiments are displayed in the Tables 2, 3 and 4.

(4) More discussions incorporating the structural and results aspects are presented in the Section D.

B TRAINING DETAILS

For Cityscapes dataset [2], we adopt a two-stage training strategy. In the first stage, a smaller image resolution (512 × 256) is used as input to fit a larger batch-size and faster convergence speed. We train model for 500 epochs using SGD with an initial learning rate of 4.5×10^{-2}. In the second stage, we freeze the batch normalization layers and finetune the model at a slightly higher image resolution (1024 × 512). We train the second stage model for 500 epochs using SGD with initial learning rate of 1×10^{-2}. For Camvid dataset [1], we use only one-stage training strategy and train the model for 1000 epochs. More training details are provided in Table 1 for a more clearer understanding.

During training, we implement common data augmentation techniques. (1) Color jittering. The offset magnitude of brightness, contrast and saturation are set to 0.5, 0.5, 0.5, respectively. (2) Random resize. The scale ranges are set to [0.25, 0.5] for input resolution 512 × 256 and [0.35, 1.0] for input resolution 1024 × 512 in Cityscapes dataset. For the CamVid dataset, we set [0.35, 1] to input resolution 960 × 480. (3) Random crop. The cropped resolutions are set to 512 × 256 and 1024 × 512 for two-stage train of Cityscapes datatset, and 480 × 360 for Camvid dataset.

During inference, we do not employ any evaluation tricks, *e.g.*, sliding-window evaluation and multiscale testing, which can improve accuracy but are time consuming. For Cityscapes dataset, we first resize images to 1024 × 512 for inference and then resize the prediction to the original image size.

We implement our method with Pytorch 1.12.0 and train it on a single NVIDIA RTX 3090 GPU.

Table 1: Hyper-parameters of the training process.

Training Config	Settings	
Dataset	Cityscapes [2]	Camvid [1]
Optimizer	SGD (momentum=0.9)	SGD (momentum=0.9)
Initial learning rate	4.5e-2 + 1e-2	1e-3
Weight decay	1e-4	5e-4
Batch size	12 + 6	12
Training epochs	500 + 500	1000
Learning rate schedule	Poly (power=0.9)	Poly (power=0.9)
Warmup epochs	5 epochs	5 epochs
Warmup schedule	Linear	Linear
Loss function	CrossEntropy Loss	CrossEntropy Loss
Loss weights	λ_1, λ_2 = 0.1, 0.01	λ_1, λ_2 = 0.1, 0.01

C EXPERIMENTAL SETUP.

Firstly, we display per-class IoU (%) results on the Cityscapes test set. Obviously, our method performs the segmentation better for the main categories, *e.g.*, building, car, person, road, traffic sign, etc.

In Fig. 2 and Fig. 3, we display some segmentation results on Cityscapes test dataset [2] and CamVid test dataset [1]. We can observe that our method, named PMSDSEN, achieves better segmentation results for both large objects, such as cars, road and sidewalk, and small objects, such as traffic light, person, rider and pole, etc. Visual results also validate the effectiveness of the proposed PMSDSEN, which can parallel extract rich and detailed local information, as well as coarse and complex large-range relationships, respectively, enabling the recognization of object boundaries and object-level area. By stacking PMSDSEs in multiple stages, network can learn sufficiently fine-grained details and textures, as well as abstract category and semantic information, achieving efficiently receptive field, which helps network to employ a larger range of surrounding context information to make robust segmentation, thus obtaining a more powerful learning capability.

*Corresponding author.

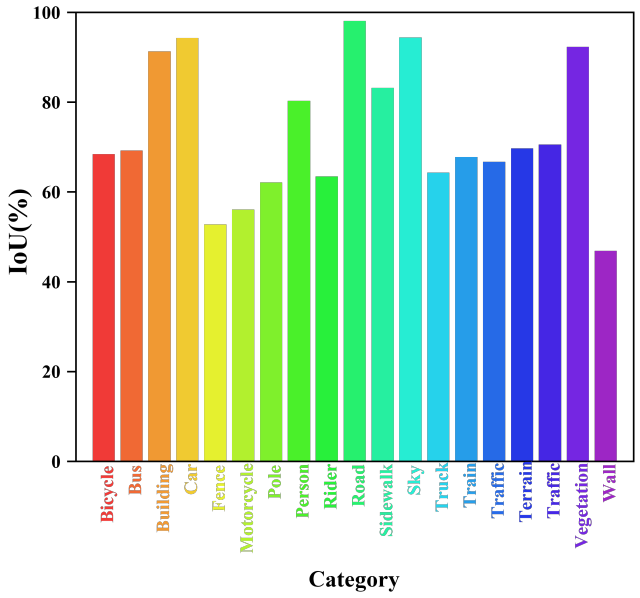

Figure 1: Per-class IoU (%) results on the Cityscapes test set.

Table 2: The effect of different number of branches in the LDI/MSLRI for the model results.

Branch counts in LDI/MSLRI	mIoU(%)
2	69.65
4 (Ours)	73.60
8	59.95

Table 3: The effect of different dilation rates in the MSLRI for the model results.

Dilation rate in MSLRI	mIoU(%)
1-2-3-6(Ours)	73.60
1-4-8-12	73.31
1-6-12-18	73.10

Table 4: The effect of different pooling kernel sizes in the MSSE for the model results.

Pooling Kernel Size in MSSE	mIoU(%)
3-7-15-Global	72.71
5-9-17-Global (Ours)	73.60
7-11-19-Global	72.23

D DISCUSSION

Architecture discussion: (1) Differences to ASPP/PPM: The ASPP and PPM modules employ atrous convolution with different rates or pooling with varying stride to extract features at different scales. These features are then combined to capture multi-scale context information. However, these semantic context modules are only added at the end of the pre-trained backbone network, which may not always suffice in extracting multi-scale spatial and semantic context. There is still potential for enhancing the extraction of multi-scale features for semantic segmentation. Therefore, this study aims to develop advanced multi-scale spatial and semantic context modules and improve their integration with hand-crafted multi-scale backbone networks to achieve more efficient and effective feature extraction in each stage. The PMSDSE integrates two parallel branches, namely MSLRI and LDI, to effectively extract both rich and detailed local information, as well as coarse and complex large-range information. As depicted in Figure 6, the fusion features exhibit finely detailed localization and powerful long-range relationships, enabling accurate recognition of object boundaries and object-level areas. In the PMSDSE, each $FAL_i(\cdot)$ function can receive features from all subparts $x_j, j \leq i$. Moreover, when a feature subpart x_j passes through a $FAL_i(\cdot)$ function, the resulting output can have an enlarged receptive field, facilitating the aggregation of contextual information from longer ranges. The MSSE architecture, similar to MSLRI, adopts parallel processes and hierarchical residual-like connections. It leverages the low-resolution feature maps as input to capture multi-scale semantic contextual information and rich spatial details, while minimizing computational burden by expanding the receptive field. This approach effectively incorporates global context information through multi-scale context aggregation, enabling the model to capture fine-grained details and high-level semantic information.

Summary, comparing our module with ASPP and PPM, the PMSDSE can be integrated into the manually designed backbone network to achieve more comprehensive multi-scale feature extraction at each stage, while ASPP and PPM are usually placed at the end of the network to extract semantic information further. In addition, through a hierarchical residual strategy, the PMSDSE and MSSE enables a larger receptive field than ASPP's and PPM's and can fully fuse multi-stage and multi-scale contextual information, so as to effectively capture detailed local information as well as complex large-range relationships. Notably, as shown in Table4(a), we achieve more optimal results with significantly fewer parameters and computational efforts compared to ASPP and PPM (Parameters (Multi-Adds): ASPP and PPM are 34% (26%) and 61% (43%) higher than ours, respectively). The main focus of our work is on extremely lightweight methods to achieve more efficient trade-offs between accuracy and complexity.

(2) Differences to channel attention: Channel attention uses feature compression and nonlinear expansion to obtain per-channel significance weights for modulating a single input Tensor. The DWFF strategy is employed to selectively emphasize the most informative parts of the feature map, effectively combining shallow and deep features and enhancing segmentation accuracy, as illustrated in Figure 7. DWFF assigns higher weight to shallow features in regions with fine-grained detail and places greater emphasis on deep features in regions with higher-level semantic information, thus improving feature combination and segmentation precision. The DWFF strategy dynamically adjusts weight allocation through fusion and selection operations. The fusion operator merges information from two streams to generate global feature descriptors, utilizing channel down-/up-scaling convolution layers to produce two feature descriptors, denoted as v_1 and v_2. The selection operator applies the SoftMax function to v_1 and v_2, generating attention activations that adaptively recalibrate the dual feature flows before their aggregation.

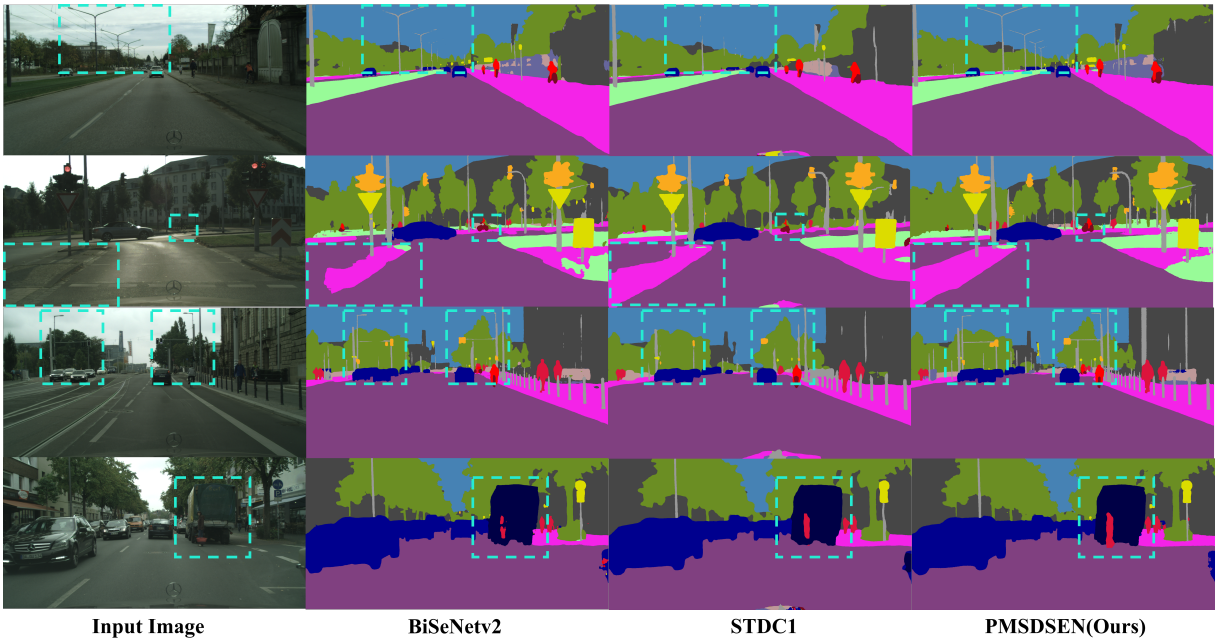

Figure 2: The visualization of the Input Image, BiSeNetv2 [5], STDC1 [3] and the proposed PMSDSEN on Cityscapes test set. Ground Truth images are not released. Our method achieves better segmentation results for both large objects, such as cars, and small objects, such as traffic light and pole, etc.

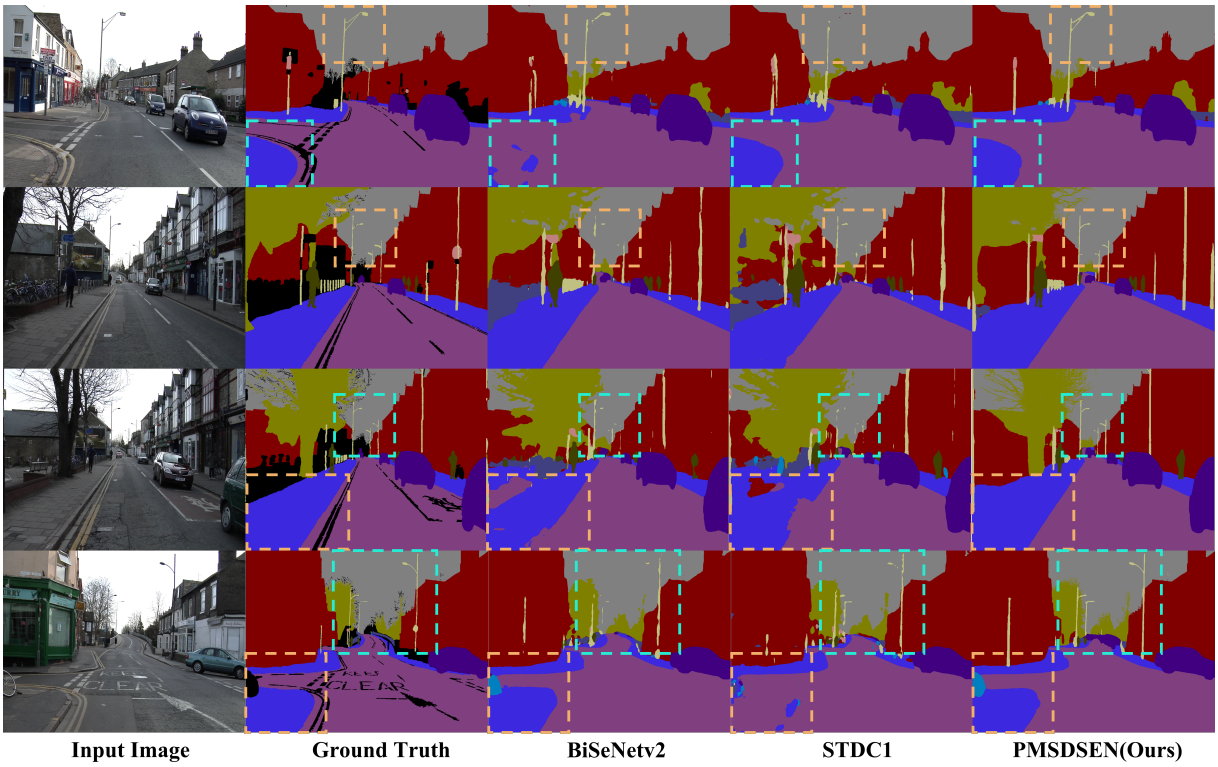

Figure 3: The visual results of the BiSeNetv2 [5], STDC1 [3] and the proposed PMSDSEN on test images from CamVid test set.

Summary, Unlike conventional channel attention, which mainly considers modulating a single feature Tensor, DWFF is able to jointly modulate the "shallow-deep" feature Tensor with the generated pairwise weights to better fuse shallow spatial and deep semantic information in our customized PMSDSEN.

Results discussion: (1) By analyzing the dataset, it is found that a significant number of images in the dataset exhibit global texture interference, with wave-like interference curves radiating in a multi-scale manner from top to bottom (*e.g.,* "0001TP_008550.png"). This will largely affect the segmentation performance. Our method leverages the MSLRI and LDI branches capture both detailed local information and complex large-range relationships, respectively. By employing a hierarchical residual strategy, our network effectively learns fine-grained details, textures, abstract categories, and semantic information. This approach efficiently expands the receptive field, enabling the network to leverage a broader context and fully fuse multi-scale contextual information for robust segmentation. Consequently, our method achieves maximum elimination of global (large-range) and multi-scale wave-like interference curves, and produces reliable segmentation results. (2) A portion of the test set exhibits more pronounced low-illumination effects compared to the training set, which may lead to performance degradation. To mitigate this issue, we employ transfer learning by utilizing pre-training weights from Cityscapes. This approach reduces the training difficulty on the CamVid dataset. It is worth noting that PIDNet [4], a very recent model with 7.6M parameters, also adopts the pre-training scheme for CamVid. We also show results with and without pre-training in Table 3.

REFERENCES

[1] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. 2009. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters* 30, 2 (2009), 88–97.

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3213–3223.

[3] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. 2021. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9716–9725.

[4] Jiacong Xu, Zixiang Xiong, and Shankar P Bhattacharyya. 2023. PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19529–19539.

[5] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. 2021. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision* 129 (2021), 3051–3068.