MPNet: Multiscale predictions based on feature pyramid network for semantic segmentation

Van Toan Quyen
Electronic and Electrical Engineering
Kyungpook National University
Daegu, South of Korea
yersin@knu.ac.kr

Min Young Kim
Electronic and Electrical Engineering
Kyungpook National University
Daegu, South of Korea
minykim@knu.ac.kr

***Abstract*—Semantic segmentation is a complex topic where they assign each pixel of an image with a corresponding class and demand accuracy at objective boundaries. The method plays a vital role in scene-understanding scenarios. For self-driving applications, the input source includes various types of objects such as trucks, people, or traffic signs. One receptive field is only effective in capturing a short range of sizes. Feature pyramid network (FPN) utilizes different fields of view to extract information from the input. The FPN approach obtains the spatial information from the high-resolution feature map and the semantic information from the lower scales. The final feature representation contains coarse and fine details, but it has some drawbacks. They burden the system with extensive computation and reduce the semantic information. In this paper, we devise an effective multiscale predictions network (MPNet) to address these issues. A multiscale pyramid of predictions effectively processes the prominent characteristics of each feature. A pair of adjacent features is combined together to predict the output separately. A lower-scale feature of each prediction is assigned as the contextual contributor, and the other provides coarser information. The contextual branch is passed through the atrous spatial pyramid pooling to improve performance. The segmentation scores are fused to obtain advantages from all predictions. The model is validated by a series of experiments on open data sets. We have achieved good results 76.5% mIoU at 50 FPS on Cityscapes and 43.9% mIoU on Mapillary Vistas.**

***Index Terms*—Semantic segmentation, feature pyramid network, multiscale prediction, real-time application.**

I. Introduction

Due to the development of technology infrastructure, semantic segmentation plays an important role in various fields such as medical imaging, unmanned aerial vehicles, or self-driving cars. The method is applied to understand the input images at pixel levels. Image segmentation classifies each pixel and assigns it to one corresponding class.

Many researchers have studied the deep learning algorithm for semantic segmentation. Early, Fully convolution networks [1] show promising results for dense prediction compared to traditional approaches. The method uses a series of convolutions to extract information from input images called an encoder and has a decoder part to segment output classes. The deep convolution network help generates rich contextual information, which is essential for semantic segmentation applications [2], [3]. However, semantic segmentation also requires rich spatial information from low features to obtain objective boundaries. Residual learning of ResNet method [4] is proposed to address the hindrance. Network, employing residual learning, can decrease the gradient problem by adding the coarse information from the lower level to the output of each bottleneck. On the other hand, many approaches utilize multiple branches at different-level features to contribute to the final map. They obtain both semantic and spatial information by using the three deepest layers of convolution network [5] or ResNet backbone [6] and different levels of ZigZag Network [7] or high-resolution network [8].

For the autonomous driving task, input scenes include various objective sizes. The field of view directly affects semantic segmentation performance. While the large field of view can observe the complete information of large objects, the small view only captures a part of the objects. Oppositely, the small receptive field suitably covers thin objects, while the large receptive field will include many different objects in a single view. DeeplabV3+ [9] propose an architecture with the ASPP module. The proposed atrous convolution can obtain different receptive fields by changing the rates R. The deepest feature of the backbone is passed through the ASPP module to gain multi-scale contextual information, and a low-level branch provides coarse information to the final map. The FPN method [10] uses four features of the backbone to capture information from inputs. The method obtains both contextual and spatial information through this combination. Additionally, many approaches gather diverse receptive fields by extracting information from different input-image sizes. Shiqi Yang et al. [11] employ two scales as image inputs, and they apply the different dilation of convolution for specific scales. The MscfNet [12] resizes image inputs into four scales and supplements each scale inference into a pipeline at different times.

In this study, we propose feature pyramid network with multiscale predictions for semantic segmentation. Based on the characteristics of the baseline FPN backbone [10], we devise a novel decoder module to process useful features of each layer effectively. The multiscale predictions include three parallel predictions with different scales separately. It is proposed to capture a wide range of objective sizes from input scenes. The proposed architecture achieves an mIoU of 76.5% and 50 FPS on Cityscapes dataset and 43.9% on Mapillary Vistas.

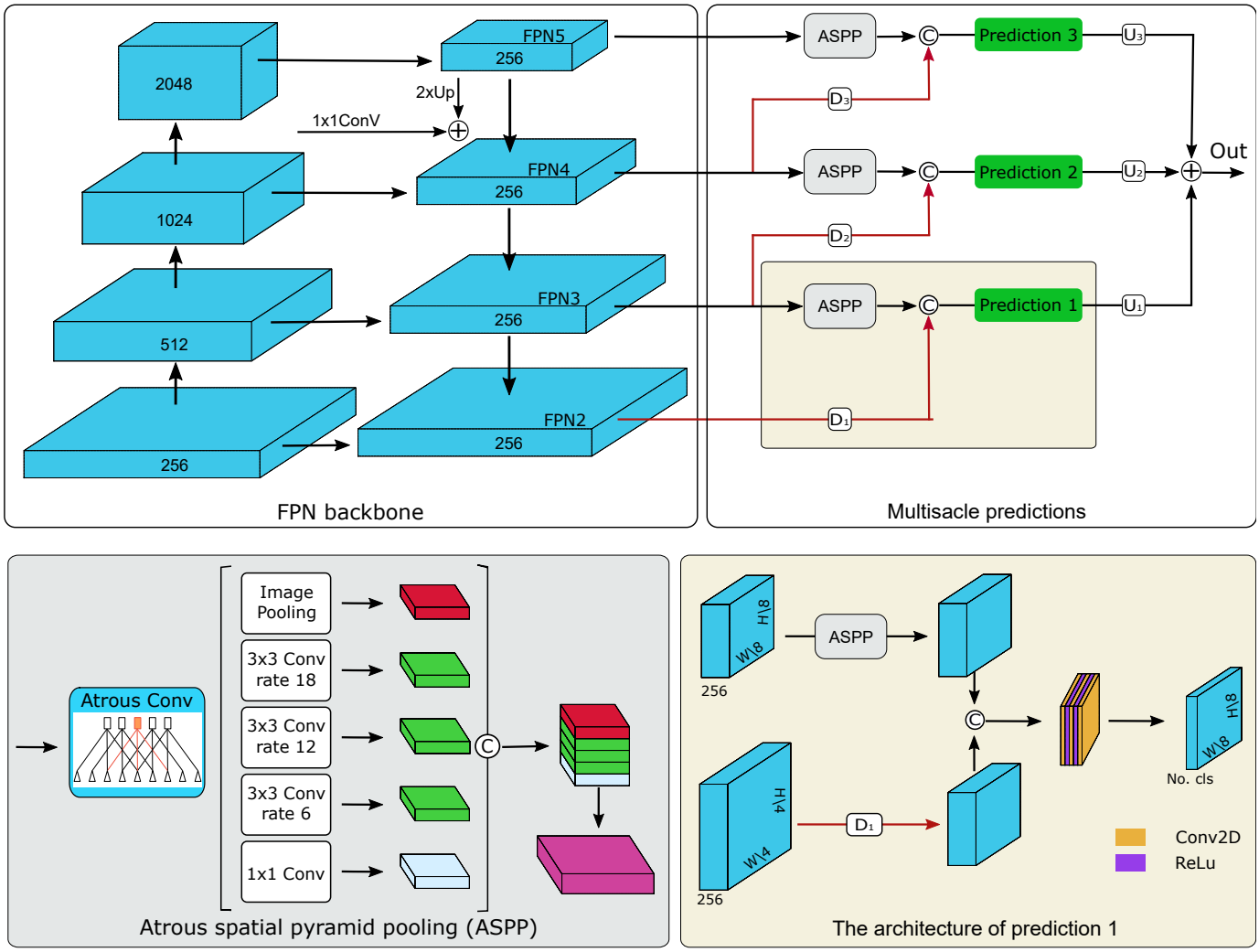

Fig. 1: A general framework of feature pyramid network with multiscale predictions for semantic segmentation is shown in the first row. The bottom-left block is a detailed structure of the ASPP module. The bottom-right block shows the structure of prediction 1.

II. Method

A. Overall architecture

In this paper, the proposed architecture consists of two main components illustrated in Fig. 1. The FPN backbone component is utilized to extract information from input images, and the multiscale predictions effectively decode the highlighted characteristics of each feature.

The FPN backbone includes parallel pathways of pyramidal feature hierarchy. The bottom-up pathway is the output of ResNet-50 encoder in which semantic information is gradually increased from low to high levels, but each step reduces the feature resolution. Each layer of the left column is fed into 1x1 convolution to downsize the channel depth to 256. To build a top-down pathway with richer semantic information, each feature is contributed by one ResNet layer and an above feature. The traditional decoder continuously uses convolution layers and concatenated functions to predict the final segmentation map. However, this approach incurs heavy computation and has not satisfied the accuracy demand.

In the second part, we deploy the novel decoder to address the limitations of the baseline method. The multiscale predictions consist of three parallel predictions, and each prediction is utilized to generate a completed semantic map separately. A pair of adjacent features is processed to provide the inputs for an individual prediction. A branch of the lower-scale feature is assigned as the contextual contributor, and the other is a coarser provider. The contextual branch is passed through atrous spatial pyramid pooling (ASPP) to obtain denser high-level feature maps. The ASPP module has different receptive fields to capture multiscale context information, especially without increasing parameters effectively. The coarser branch is downsampled to have the exact size of the lower one.

B. Feature characteristics of the backbone

In this section, we analyze the feature characteristics of the FPN backbone to indicate the prominence of each feature. Semantic segmentation tasks require both spatial and contextual information to predict the final heads. Convolution networks are a sequence of feature layers in which the resolution and the semantic information are inversely proportional. The first stages contain basic information with high resolution, and the deeper layers contain rich semantic information with low resolution.

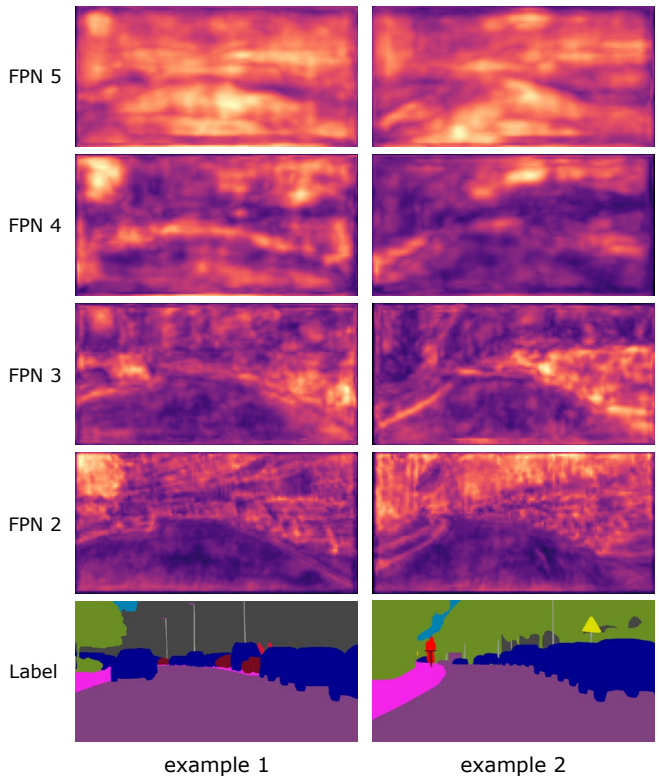

Fig. 2: Feature characteristics of the FPN backbone at multi-scale layers. In the heatmaps, the dark color indicates the small pixel value and the brighter areas are high values.

To explain clearly the differences in the backbone features, we visualize the convolutional layers shown in Fig. 2. The visualization illustrates four-feature scales of the top-down pathway and the corresponding labels of two different examples. At the first layer, the FPN 2 has high resolution and effectively extracts information about simple shapes represented in a bright color. The layer contains the coarsest feature having information on large objects, especially the boundary. Oppositely, The deepest FPN 5 layer has high-level semantics and involves information about complicated shapes. This layer contains the finest information, and it is hard to determine the shapes or edges of the objects. The middle FPN 3 and FPN 4 comprise both coarse and fine information with different densities. Each layer comprises some useful features and has different contributions to the semantic segmentation final. The baseline approach only adds features together, so bad values average the prominent characters from other layers.

C. Multiscale predictions module

As a detailed analysis of feature characteristics, we embed multiscale predictions module to process the useful points of the layers effectively. Each pair of adjacent features contains similar characteristics, so we combine them as an individual prediction. The novel module composes of three parallel predictions, and each prediction has the exact size of the lower branch. Our approach addresses the drawback of averaging-pixel values and reduces the computing burden on hardware.

As illustrated in the second row of Fig 1, we show an example of prediction 1 architecture. The inputs for this prediction are the FPN 2 and FPN 3 layers. All layers have the same 256 channels, but width and height are different. The FPN 3 layer is passed through the ASPP to obtain multi-scale contextual information. The ASPP method extract feature maps with different receptive fields and do not increase the number of parameters depicted in the bottom-left of Fig 1. The FPN 2 branch is downsampled and then concatenated with the lower branch to gain the information. The concatenated layer is fed into a series of 3x3 convolution filters and ReLu activation to predict the class labels for each pixel. The dimension of the score map has the height and width of the lower-scale feature and the class number of a corresponding dataset. The other predictions have the same process.

In order to gain the completed information from three predictions, the final map is calculated by equation 1

$$\mathcal{SS} = \sum_{i=0}^{C} \sum_{j=0}^{H*W} (U(L_{ij}) + U(M_{ij}) + U(S_{ij}) \qquad (1)$$

where SS is the semantic segmentation result. L, M, and S are prediction 1, prediction 2, and prediction 3 of multiscale predictions module. H and W are the height and width of input resolution, and C is the class number of the dataset. Lastly, i and j are instant locations of pixels.

III. Experiments

In the experiment section, we validate our approach on Cityscapes and Mapillary Vistas datasets to show the improvement in both terms of quantitative and qualitative results. We train the model by 150 epochs on Pytorch deep learning framework. We evaluate the proposed method by some standard metrics such as mean intersection of union (mIoU) for accuracy and frame per second (FPS) for speed.

A. Cityscapes dataset

Cityscapes is a public segmentation dataset for autonomous applications. The data involves the urban street scenes collected from 50-different countries around the world. It contains 19 objective classes and total 5,000 images with resolution 1024x2048. The whole images are divided into three separate sets. The training and validation sets include RGB images and corresponding ground truths, and the test set only

TABLE I: Per-class results between proposed method and existing approaches on Cityscapes validation set. Objective categories are road, side walk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, and bicycle.

Method	Road	swalk	build	wall	fence	pole	tlight	tsign	veg.	terr	sky	pers	rider	car	truck	bus	train	mcle	bicle	mIoU
AGLNet [13]	97.8	81.0	91.0	51.3	50.6	58.3	63.0	68.5	92.3	71.3	94.2	80.1	59.6	93.8	48.4	68.1	42.1	52.4	67.8	70.1
ESPNet [14]	97.3	78.6	88.8	43.5	42.1	49.3	52.6	60.0	90.5	66.8	93.3	72.9	53.1	91.8	53.0	65.9	53.2	44.2	59.9	66.2
FSCNN [15]	97.4	77.8	87.4	39.7	41.8	35.0	39.4	50.5	88.5	63.3	92.7	65.7	46.4	91.0	57.0	70.3	56.5	40.9	52.6	62.8
DABNet [16]	97.8	80.7	90.2	47.9	48.1	56.4	61.8	67.0	92.0	69.5	94.3	80.3	59.2	93.7	46.0	57.1	35.0	50.4	66.8	68.1
CFPNet [17]	97.8	81.4	90.5	46.4	50.6	56.4	61.5	67.7	92.1	68.9	94.3	80.4	60.7	93.9	51.4	68.0	50.8	51.2	67.7	70.1
FANet [18]	97.9	83.3	91.6	55.5	55.1	60.3	66.2	74.9	91.7	61.8	94.7	78.5	58.1	94.1	76.8	85.1	74.5	50.7	73.9	75.0
FPN [10]	97.5	81.6	90.9	46.3	54.2	59.1	63.9	74.3	91.3	58.8	93.6	78.4	56.0	93.3	54.0	71.9	54.7	59.6	74.8	71.3
MPNet (ours)	97.7	82.1	92.0	57.9	61.3	61.1	64.2	74.2	91.6	61.4	94.0	79.5	59.5	93.9	80.4	88.0	74.5	63.4	75.6	76.5

TABLE II: Performance and inference speed comparison between our approach and other methods on Cityscapes dataset.

Method	Resolution	mIoU	FPS
AGLNet [13]	512×1024	70.1	52
TCNet [19]	1024x2048	74.6	16
SwiftNet [20]	1024×2048	75.4	39
ICNet [21]	1024x2048	67.7	38
FPN [10]	1024×2048	71.3	18
MPNet (ours)	1024×2048	76.5	50

has RGB images. The train, validation, and test sets have 2975, 500, and 1525 images, respectively.

We conduct experiments on the Cityscapes to verify the superiority of our method. Regarding per-class accuracy, our approach achieves the best mIoU performance on average, shown in Tab I. The stuff classes involving road, vegetation, or sky have high accuracy for all approaches. For thing classes, we analyze the effectiveness of our model based on categories of objective sizes. Firstly, we observed that our method improves the large objective accuracy with 3.6% for trucks and 2.9% for buses compared to the second place, respectively. Receptive fields of all methods are suitable to capture the medium-sized car, so the results emphasize outstanding accuracy. Lastly, the thin-structure classes are fence, motorcycle, or bicycle, so our method also proves the improvement. Our method does not achieve the highest accuracy for all classes, but we have a good trade-off for a wide range of objective sizes on the streets. The FANet [18] is an example of the class-accuracy unbalance. They are good for roads or the sky, but the results of walls, fences, or motorcycles are bad.

To further verify the effectiveness of our method, we conduct experiments in terms of performance and inference speed shown in Tab II. Our method accomplishes outstanding results with 76.5% mIoU and 50 FPS on Cityscapes dataset. Compared with the baseline FPN method that does not use multiscale predictions module, the proposal approach overcomes the baseline FPN method [10] in both aspects of performance and speed where we achieve 5.2% mIoU improvement and have 32 FPS faster. Compared to the highest accuracy (SwiftNet method [20]), our method achieves an improvement of 1.1% mIoU and significantly surpasses them in the perspective of speed. AGLNet [13] study has solved the time consumption hindrance, but the accuracy is deficient with only 70.1% mIoU. The results show that our methodology sacrifices the performance and reduces the computation for semantic segmentation.

The qualitative results are validated on Cityscapes dataset and shown in Fig. 3. We categorize objective classes into three groups to analyze our study's effectiveness. Our approach has good results in different examples. The traditional FPN only has good results for medium sizes such as cars. For narrow structures of poles, traffic lights, or people, our study recognizes them much better than the FPN method. The baseline cannot capture the complete information of large objects, so it lacks the global context. Proposed multiscale predictions have effectively boosted prominent characteristics of each feature, so the method is suitable to segment a wide range of objective sizes.

TABLE III: Performance comparison between our approach and SOTA methods on Mapillary Vistas dataset.

Method	Resolution	mIoU
AGLNet [13]	1024x2048	30.7
DABNet [16]	1024x2048	29.6
RGPNet [22]	1024×2048	41.7
FPN [10]	1024×2048	40.2
MPNet (ours)	1024×2048	43.9

B. Mapillary Vistas

Mapillary Vistas is a complex segmentation dataset. It contains a total of 65 objective classes in which some classes are tough challenges for all approaches, such as CCTV camera, ground animal, or parking. The data is gathered around the world under various weather conditions and has a long range of resolutions. It includes 20,000 images separated into two different sets. There is the training set with 18,000 images and the validation set with 2,000 images.

We also conduct experiments on the Mapillary Vistas dataset to show the effectiveness of our method. The performance comparison between our method and other state-of-the-art methods is illustrated in Tab III. As a result, we achieve an mIoU of 43.9%. Compared to the FPN approach, the performance of the model is improved by 3.7% mIoU. Compared to the highest accuracy obtained by RGPNet method [22], our method has increased by 2.2%. Our accuracy is much better than AGLNet [13] and DABNet [16] approaches with 13.2% and 14.3% improvement, respectively.

Finally, we visualize the qualitative results of the baseline FPN method and proposed method on Mapillary Vistas dataset shown in Fig. 4. Our method effectively captures a wide range of objective sizes, illustrated in the red bounding boxes. Compared with the baseline FPN visualization, we observed that our model segments small objects with more precise and smoother outputs.

IV. Conclusion

According to the characteristics of the FPN backbone, this study proposes multiscale predictions module to process useful features effectively. The novel module with three parallel predictions is suitable to segment a long range of objective sizes from complex street scenes. Additionally, the ASPP module is deployed to obtain multi-scale contextual information for each prediction and does not increase the parameter numbers. Both quantitative and qualitative results demonstrate that our proposal outperforms the other existing methods on different public datasets. Our model not only sacrifices the performance but also accelerates real-time application. In the future, we determine to improve the model accuracy by using a heavier architecture and devising the fusion module for multiscale predictions.

Acknowledgment

This work was supported by Basic Science Research Program through the National Research Institute Foundation of Korea (NRF) funded by the Ministry of Education(2021R1A6A1A03043144) and BK21 Four project funded by the Ministry of Education, Korea (4199990113966). And this work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1A2C2008133).

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] Q. Zhou, W. Yang, G. Gao, W. Ou, H. Lu, J. Chen, and L. J. Latecki, "Multi-scale deep context convolutional neural networks for semantic segmentation," *World Wide Web*, vol. 22, no. 2, pp. 555–570, 2019.

[6] B. Zhang, W. Li, Y. Hui, J. Liu, and Y. Guan, "Mfenet: Multi-level feature enhancement network for real-time semantic segmentation," *Neurocomputing*, vol. 393, pp. 54–65, 2020.

[7] D. Lin, D. Shen, S. Shen, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang, "Zigzagnet: Fusing top-down and bottom-up context for object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7490–7499.

[8] J. Zhang, S. Lin, L. Ding, and L. Bruzzone, "Multi-scale context aggregation for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 12, no. 4, p. 701, 2020.

[9] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[10] S. Seferbekov, V. Iglovikov, A. Buslaev, and A. Shvets, "Feature pyramid network for multi-class land segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 272–275.

[11] S. Yang and G. Peng, "Attention to refine through multi scales for semantic segmentation," in *Pacific Rim Conference on Multimedia*. Springer, 2018, pp. 232–241.

[12] G. Gao, G. Xu, Y. Yu, J. Xie, J. Yang, and D. Yue, "Mscfnet: a lightweight network with multi-scale context fusion for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[13] Q. Zhou, Y. Wang, Y. Fan, X. Wu, S. Zhang, B. Kang, and L. J. Latecki, "Aglnet: Towards real-time semantic segmentation of self-driving images via attention-guided lightweight network," *Applied Soft Computing*, vol. 96, p. 106682, 2020.

[14] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 552–568.

[15] R. P. Poudel, S. Liwicki, and R. Cipolla, "Fast-scnn: Fast semantic segmentation network," *arXiv preprint arXiv:1902.04502*, 2019.

[16] G. Li, I. Yun, J. Kim, and J. Kim, "Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation," *arXiv preprint arXiv:1907.11357*, 2019.

[17] A. Lou and M. Loew, "Cfpnet: Channel-wise feature pyramid for real-time semantic segmentation," *arXiv preprint arXiv:2103.12212*, 2021.

[18] P. Hu, F. Perazzi, F. C. Heilbron, O. Wang, Z. Lin, K. Saenko, and S. Sclaroff, "Real-time semantic segmentation with fast attention," *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 263–270, 2020.

[19] Z. Wu, C. Shen, and A. v. d. Hengel, "Real-time semantic image segmentation via spatial sparsity," *arXiv preprint arXiv:1712.00213*, 2017.

[20] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic, "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*, 2019, pp. 12 607–12 616.

[21] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 405–420.

[22] E. Arani, S. Marzban, A. Pata, and B. Zonooz, "Rgpnet: A real-time general purpose semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3009–3018.

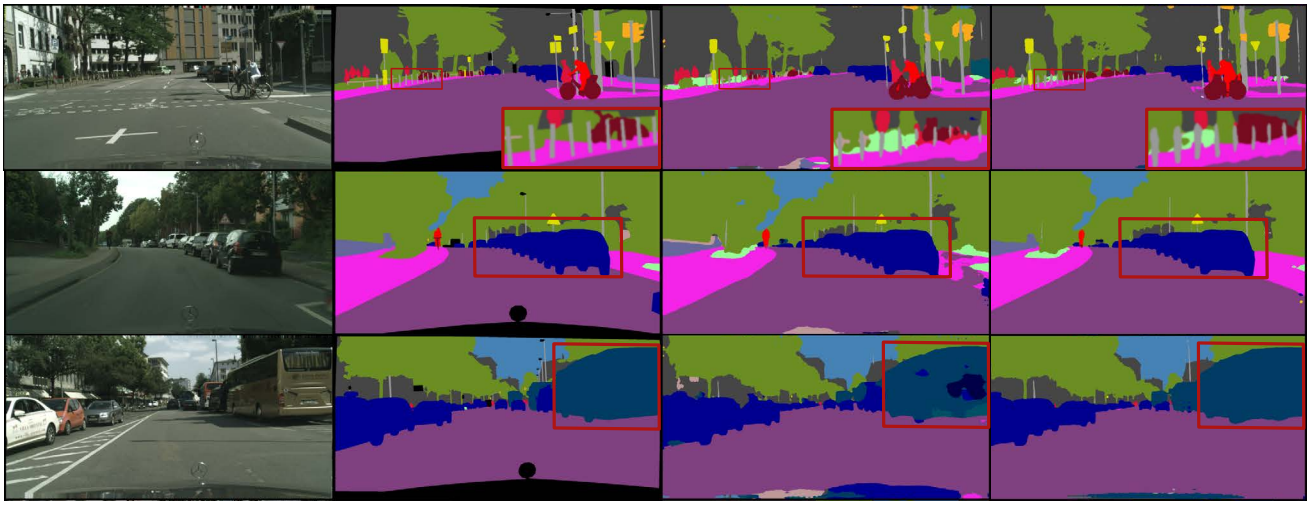

Fig. 3: Visualization of the baseline FPN method and our proposal approach on Cityscapes dataset

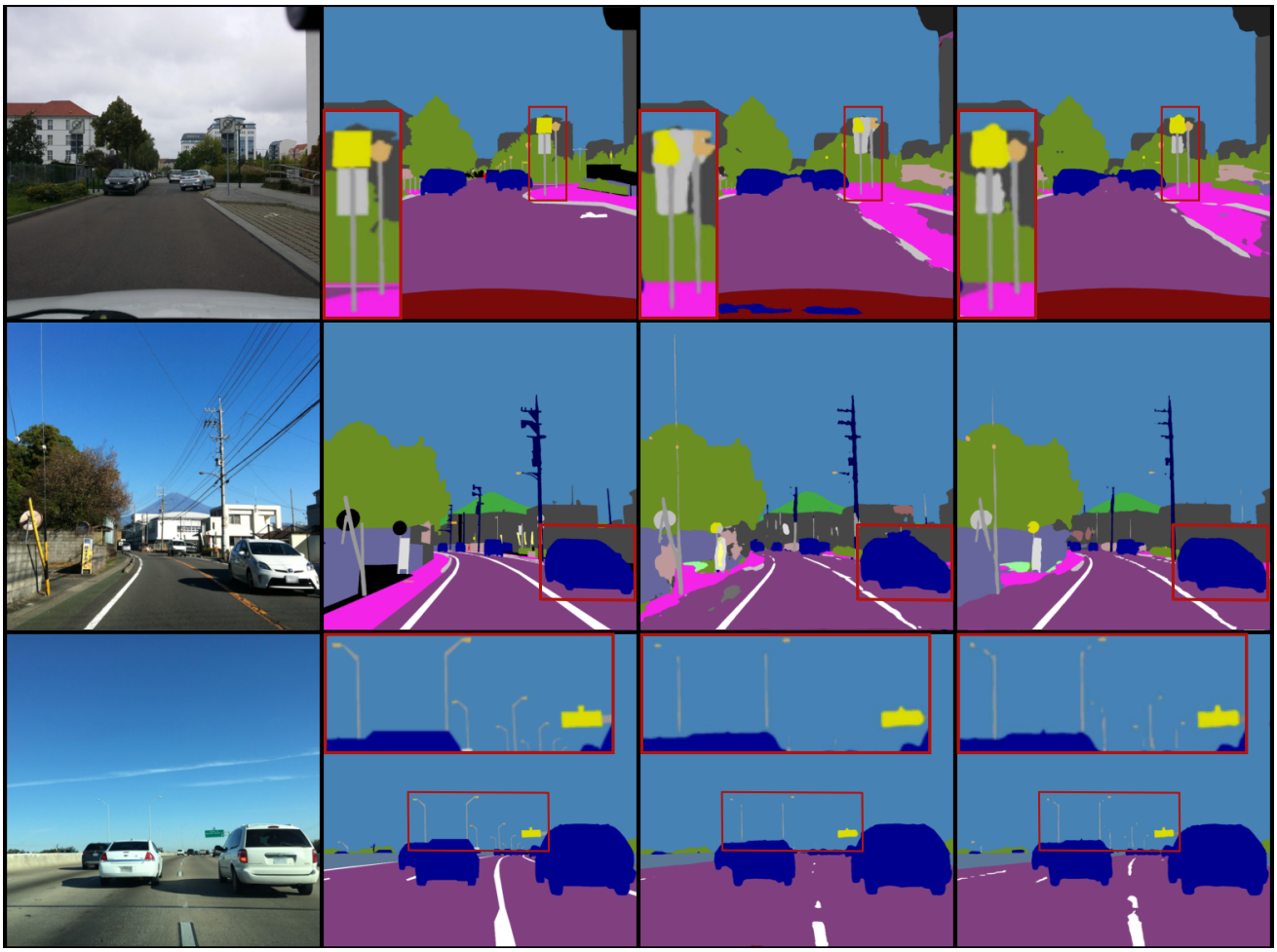

Fig. 4: Visualization of the baseline FPN method and our proposal approach on Mapillary Vistas dataset