[Supplementary Material] Center Focusing Network for Real-Time LiDAR Panoptic Segmentation

Xiaoyan Li[1,2] Gang Zhang[3] Boyue Wang[1,2] Yongli Hu[1,2] Baocai Yin[1,2]
[1]Beijing Municipal Key Lab of Multimedia and Intelligent Software Technology
[2]Beijing Institute of Artificial Intelligence, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[3]Mogo Auto Intelligence and Telemetics Information Technology Co. Ltd.
{xiaoyan.li,wby,huyongli,ybc}@bjut.edu.cn zhanggang11021136@gmail.com

In the supplementary material, the implementation details, extensive ablation studies, and some representative visualization results are shown in section 1, section 2, and section 3, respectively.

1. Implementation Details

In this section, we illustrate the detailed architectures of the feature enhancement module in the proposed CFFE and the CFNet framework with the backbone of the CPGNet [7].

Feature enhancement module. As shown in Fig. 1, the feature enhancement module (FEM) fuses the semantic feature maps $\boldsymbol{F}_m^{sem}$ and the re-projected shifted center feature maps $\boldsymbol{F}_{p\to m}^{sem}$ to generate center-focusing semantic feature maps $\boldsymbol{F}_m^{CFsem}$ and instance feature maps $\boldsymbol{F}_m^{CFins}$ (m is the specific 2D view, such as RV [1, 9, 12], BEV [4], and polar view [15].), which are used by the subsequent semantic and instance branches for more accurate predictions. Specifically, it first concatenates the two feature maps. Then, the concatenated feature maps undergo three convolution layers, where the dilation coefficients are set as 1, 2, and 4, respectively, to enlarge the receptive field. In the experiments, it is found that a larger receptive field can improve performance. Finally, the outputs of the three convolution layers are concatenated and then undergo two extra convolution layers to get semantic and instance feature maps, respectively. In our implementation, C_2 denotes the number of output channels of the corresponding 2D projection-based backbone. C_3 and C_4 are set as 64 and 48, respectively.

CFNet with the backbone of the CPGNet. Fig. 2 presents the proposed CFNet with the backbone of the CPGNet [7], which is a powerful and efficient multi-view fusion backbone and consists of the 2D projection-based bird's-eye view (BEV) and range view (RV) branches. Fig. 3 shows the corresponding center focusing feature encoding (CFFE) that is integrated with the CPGNet [7] backbone. These two modules are similar to those of the single-view backbone but add another view to alleviate the information loss during 2D projection.

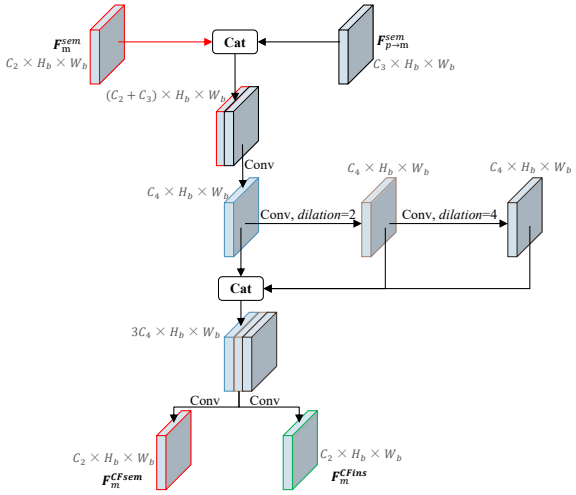


Figure 1. The feature enhancement module in the proposed CFFE. "Conv" represents a 2D convolution with 3×3 kernels, a batch normalization, and a ReLU layer. *dilation* denotes the dilation coefficient of the 2D convolution and is set as 1 unless specified.

As shown in Fig. 2, for efficiency, only one stage version of the CPGNet is adopted in our CFNet. In the CPGNet, the P2G operation aims to project the LiDAR point features onto the BEV and RV feature maps. Specifically, C_1 and C_2 are set as 64. H_b and W_b are set as 600. H_r and W_r are set as 64 and 2048, respectively. The 2D FCN extracts features on each view. On the contrary to the P2G, the G2P operation transmits the features from each view back to the LiDAR points. The PF is responsible for fusing the features from the 3D points, BEV, and RV to generate point-wise features for the following predictions. For the details of the CPGNet backbone, please refer to the CPGNet [7].

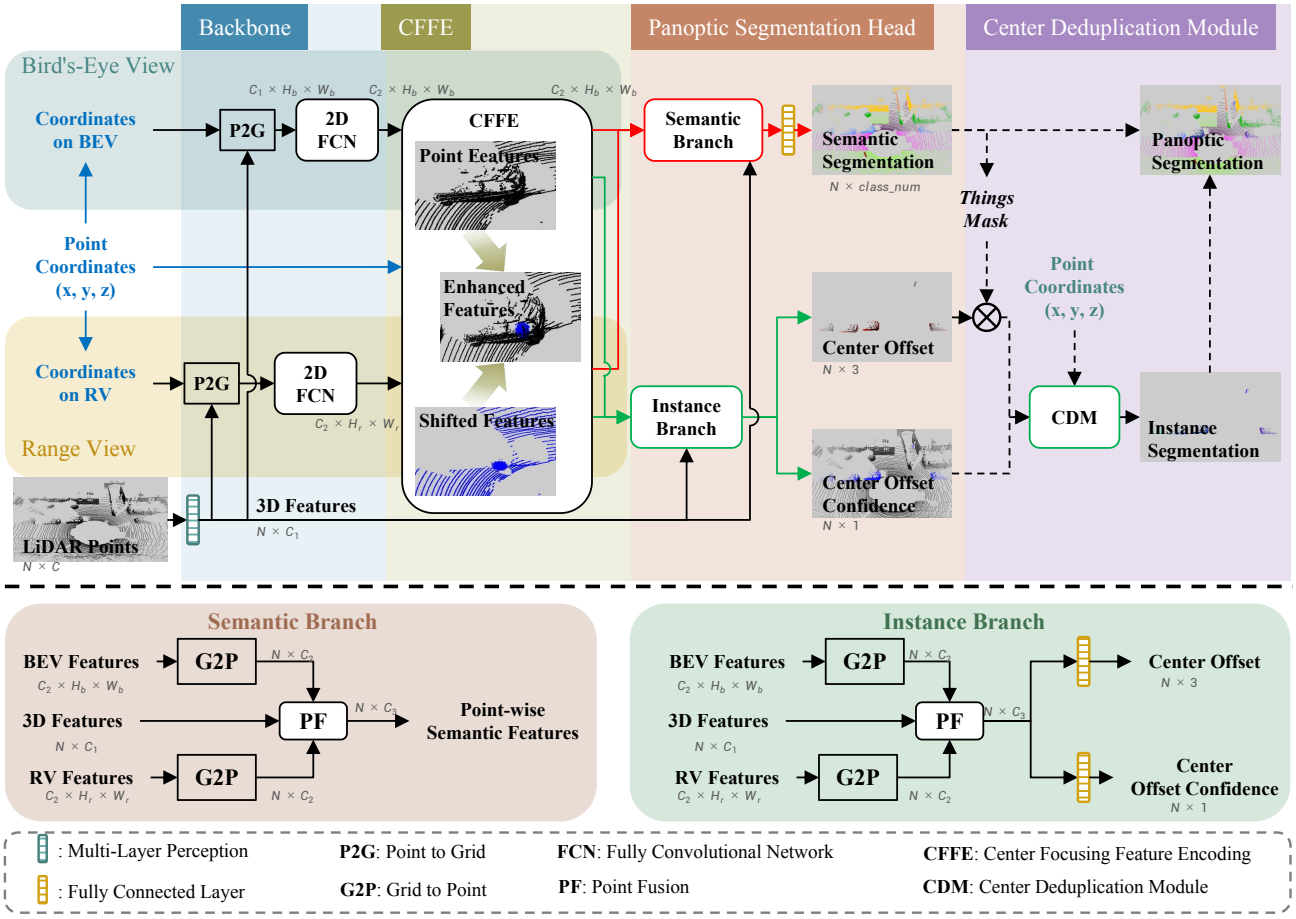


Figure 2. The overview of our CFNet with the backbone of the CPGNet [7].

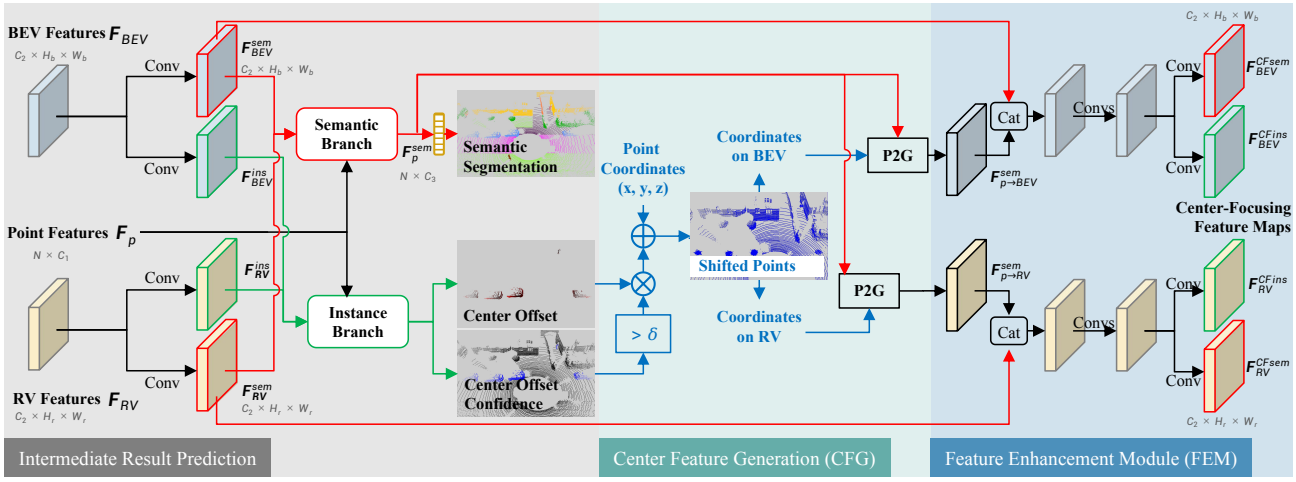


Figure 3. The proposed center focusing feature encoding (CFFE) that is integrated with the CPGNet [7] backbone. The "Conv" represents a 2D convolution with 3×3 kernels, a batch normalization, and a ReLU layer. The details of the semantic branch and instance branch are shown in Fig. 2. The blue arrows are coordinate-related operations.

2. Ablation studies

In this section, the ablative experiments are enriched for a comprehensive understanding of our CFNet.

	Methods	Backbone	PQ	PQ†	PQTh	PQSt	mIoU
a	CFNet [Ours]	CPGNet [7]	62.7	67.5	70.0	57.3	67.4
b	CFNet [Ours]; $dilation = 1$		62.2	66.7	68.5	57.2	67.1
c	CFNet [Ours]; GT Offsets		**65.5**	**69.7**	**76.4**	**57.5**	**69.5**

Table 1. Ablation studies on the SemanticKITTI validation set. $dilation = 1$ denotes that all dilation coefficients in the feature enhancement module (FEM) are set as 1. "GT Offsets" means that the center feature generation (CFG) generates the re-projected feature maps $\boldsymbol{F}^{sem}_{p\to m}$ according to the ground-truth center offsets instead of the predicted ones.

distance threshold d	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4
car	20.4	64.3	90.4	94.2	94.9	95.1	**95.2**	**95.2**	**95.2**	**95.2**	**95.2**	95.0	94.6
truck	4.7	29.2	40.6	59.5	71.3	71.7	**71.9**	**71.9**	**71.9**	71.5	70.8	70.0	68.9
person	16.6	76.9	84.7	**85.5**	**85.5**	84.9	83.3	82.7	81.9	81.0	79.9	77.8	76.8
bicycle	10.2	42.3	55.3	58.0	**58.6**	58.2	58.2	58.1	57.8	57.2	56.4	56.1	55.9

Table 2. Different values of distance threshold d in the proposed center deduplication module (CDM) on the SemanticKITTI validation set.

Different configurations of the CFFE. As shown in Table 1, row (a) is the proposed CFNet with the backbone of the CPGNet [7]. Row (b) is that all dilation coefficients in the feature enhancement module (FEM) are set as 1. Row (c) denotes that the center feature generation (CFG) generates the re-projected feature maps $\boldsymbol{F}^{sem}_{p\to m}$ according to the ground-truth center offsets instead of the predicted ones. It can be discovered that: 1) the larger receptive field results in the better performance (a,b); 2) the ground-truth center offsets facilitate the biggest performance improvements (a,c), which illustrates the importance and upper bound of the CFFE.

Distance threshold d on different classes. To figure out the effects of the distance threshold d, Table 2 shows the distance threshold d versus the PQ metric on some representative classes. The *car* and *truck* denote large objects, while the *person* and *bicycle* are small objects. The *car* and *truck* get the best PQ when the distance threshold d is in the range of 1.2 to 1.6. The *person* and *bicycle* get the highest PQ when the distance threshold d is set as 0.8. Thus, the optimal distance threshold d varies from different classes. However, when d is set as 0.8 in the main body, the performances of different classes are comparable to the optimal ones.

Comparison results on the nuScenes test set. As shown in Table 3, the proposed CFNet can be comparable with the state-of-the-art Panoptic-PHNet [6] on the nuScenes test set. However, the proposed CFNet runs much faster, as referred to the Table 3 of the main body.

Methods	PQ	PQ†	SQ	RQ	mIoU
EfficientLPS [11]	62.4	66.0	83.7	74.1	66.7
Panoptic-PolarNet [16]	63.6	67.1	84.3	75.1	67.0
Panoptic-PHNet [6]	**80.1**	**82.8**	**91.1**	**87.6**	80.2
CFNet [Ours] w/CPGNet [7]	79.4	81.6	90.7	87.0	**83.6**

Table 3. Comparison results on the nuScenes test set.

3. Visualization

In this section, our CFNet with the backbone of the CPGNet [7] is inferred on the SemanticKITTI test set.

Comparison visualization results. We run the official code of the Panoptic-PolarNet [16] and DS-Net [2] with the provided model parameters on the SemanticKITTI test set. For better visualization comparison, it only presents the instance segmentation results in Fig. 4. It can be observed that the over-segmented and under-segmented problems frequently occur in the Panoptic-PolarNet and DS-Net, while our CFNet can avoid this problem. By the way, the over-segmented problem means that an instance is split into several parts and the under-segmented problem means that adjacent instances are predicted as a single instance.

Visualization results. We present more visualization results of our CFNet on the SemanticKITTI test set in Fig. 5. Our CFNet can distinguish adjacent objects. Besides, the boundaries of instances can be accurately segmented.

References

[1] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In *International Symposium on Visual Computing*, pages 207–222. Springer, 2020. 1

[2] Fangzhou Hong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Lidar-based panoptic segmentation via dynamic shifting network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13090–13099, 2021. 3, 5

[3] Juana Valeria Hurtado, Rohit Mohan, Wolfram Burgard, and Abhinav Valada. Mopt: Multi-object panoptic tracking. *arXiv preprint arXiv:2004.08189*, 2020.

[4] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019. 1

[5] Enxu Li, Ryan Razani, Yixuan Xu, and Liu Bingbing. Smac-seg: Lidar panoptic segmentation via sparse multi-directional attention clustering. *arXiv preprint arXiv:2108.13588*, 2021.

[6] Jinke Li, Xiao He, Yang Wen, Yuan Gao, Xiaoqiang Cheng, and Dan Zhang. Panoptic-phnet: Towards real-time and high-precision lidar panoptic segmentation via clustering pseudo heatmap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11809–11818, 2022. 3

[7] Xiaoyan Li, Gang Zhang, Hongyu Pan, and Zhenhua Wang. Cpgnet: Cascade point-grid fusion network for real-time lidar semantic segmentation. *arXiv preprint arXiv:2204.09914*, 2022. 1, 2, 3

[8] Andres Milioto, Jens Behley, Chris McCool, and Cyrill Stachniss. Lidar panoptic segmentation for autonomous driving. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8505–8512. IEEE, 2020.

[9] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4213–4220. IEEE, 2019. 1

[10] Ryan Razani, Ran Cheng, Enxu Li, Ehsan Taghavi, Yuan Ren, and Liu Bingbing. Gp-s3net: Graph-based panoptic sparse semantic segmentation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16076–16085, 2021.

[11] Kshitij Sirohi, Rohit Mohan, Daniel Büscher, Wolfram Burgard, and Abhinav Valada. Efficientlps: Efficient lidar panoptic segmentation. *IEEE Transactions on Robotics*, 2021. 3

[12] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2020. 1

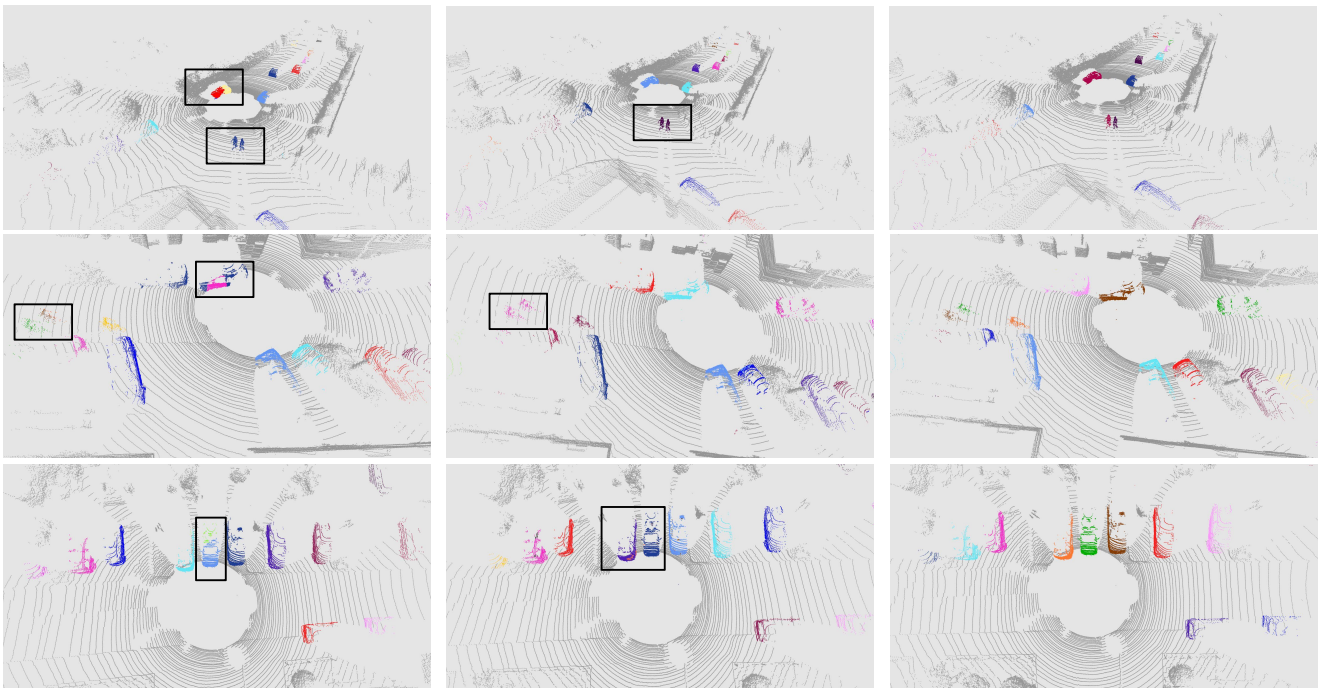
[13] Shuangjie Xu, Rui Wan, Maosheng Ye, Xiaoyi Zou, and Tongyi Cao. Sparse cross-scale attention network for efficient lidar panoptic segmentation. *arXiv preprint arXiv:2201.05972*, 2022.

[14] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[15] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9601–9610, 2020. 1

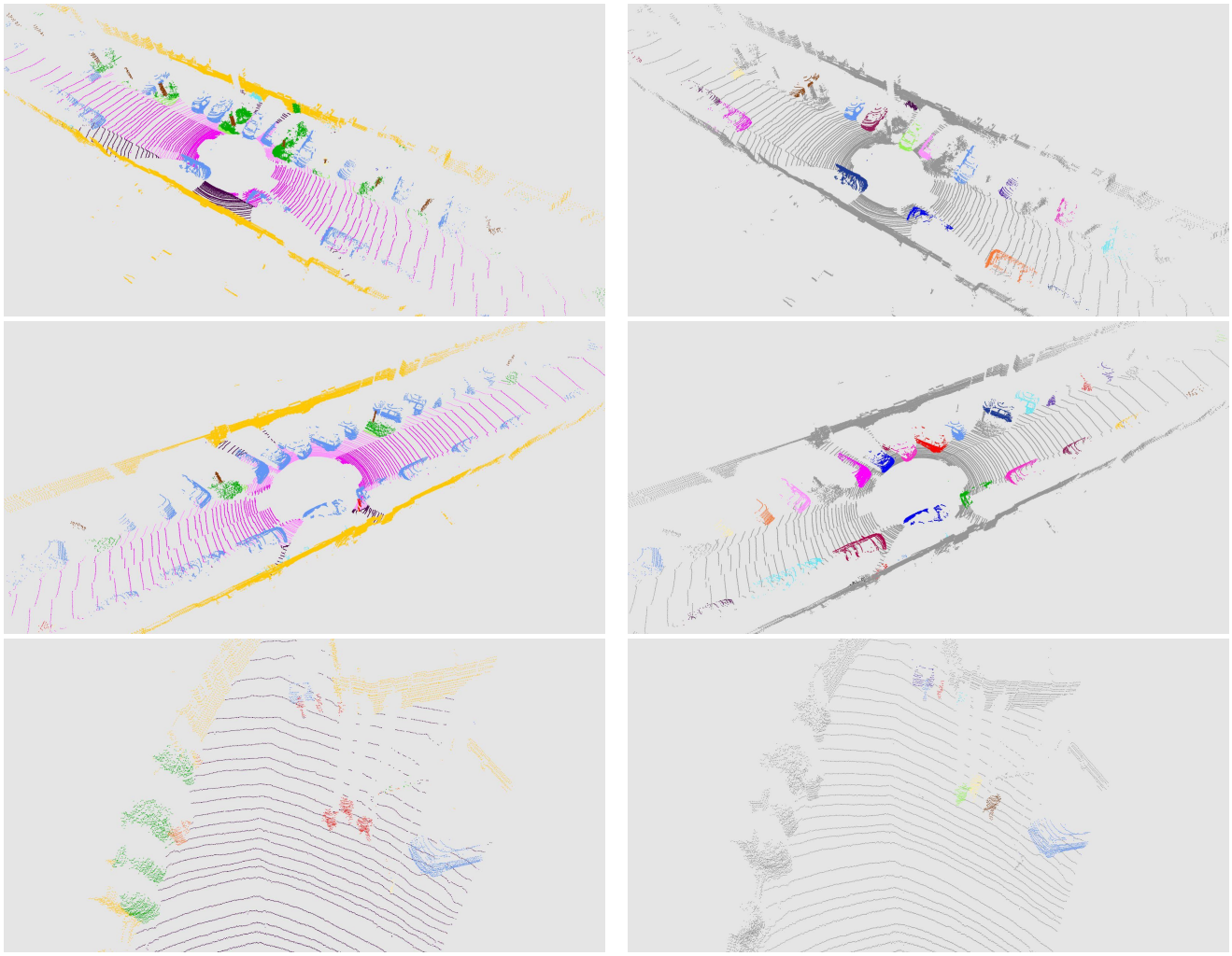
[16] Zixiang Zhou, Yang Zhang, and Hassan Foroosh. Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13194–13203, 2021. 3, 5

[17] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9939–9948, 2021.



Panoptic-PolarNet DS-Net CFNet (Ours)

Figure 4. Comparison visualization results of the instance segmentation from the Panoptic-PolarNet [16], DS-Net [2], and our CFNet on the SemanticKITTI test set. The black box marks the region of interest.



semantic segmentation instance segmentation

Figure 5. Visualization results of our CFNet on the SemanticKITTI test set. For semantic and instance segmentation, different colors represent different classes and instances, respectively.