Supplementary Material for Towards Deeply Unified Depth-aware Panoptic Segmentation with Bi-directional Guidance Learning

1. Implementation Details

1.1. Network

We choose ResNet-50 [3] and Swin-B [6] as the shared backbone, initialized with ImageNet-1K [2] pre-trained checkpoints[1]. We also use the multi-scaled deformable attention Transformer (MSDeformAttn [12]) as the decoder for both the semantic and depth branches, for consistency with the previous work [1]. The Transformer decoder comprises $l = 9$ layers to process per-segment queries, with a total of $N = 100$ queries used. The latent representation has the same dimension as the per-segment query: $\mathcal{R}^l \in \mathbb{R}^{N \times E}$, where $E = 256$.

1.2. Training

To align with previous practices [8, 11], our training process comprises two steps. Initially, we train the segmentation branch exclusively for 50 epoches; then, we fine-tune the entire model for an additional 10 epoches with bi-directional guidance learning added. We use Detectron2 [10] and 8 Titan RTX GPUs for training, with a batch size of 16 and 8 in the two training steps, respectively. During training, We adopt AdamW [7] optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.99$) with a poly learning rate schedule, and a learning rate multiplier of 0.1 is applied to the backbone. Additionally, large-scale jittering (LSJ) [5] and horizontal flipping are also utilized.

In the first step, we randomly resize the original images to a scale ranging from 0.5 to 2.0, followed by a fixed-size crop of 512×1024 on Cityscapes-DVPS and 384×1280 on SemKITTI-DVPS. For all backbones, we set an initial learning rate of 0.0001 and a weight decay of 0.05. In the second step, we fine-tune the model using full-resolution images and reduce the initial learning rate to 0.00005.

λ_{depth}	$\lambda = 0.5$	$\lambda = 0.25$	$\lambda = 0.1$	DPQ	PQ	abs rel
1.0	69.28	67.10	51.87	62.75	**69.49**	0.0644
2.5	**69.30**	**66.82**	**52.78**	**62.97**	**69.49**	0.0632
5.0	69.07	66.64	52.20	62.64	69.25	**0.0630**

Table 1: Ablations of depth loss weight λ_{depth} on Cityscapes-DVPS. We keep the segmentation loss weights (λ_{cls} and λ_{mask}) unchanged and only modify λ_{depth}.

2. Hyper-parameter Analysis

We perform additional hyper-parameter analysis, specifically adjusting the weights for loss terms and hyper-parameter settings in bi-directional guidance learning. To avoid excessive hyper-parameter tuning, we maintain consistency with most of the hyper-parameters used in previous studies [1, 4], and we use ResNet-50 [3] backbones for our experiments.

Initially, we show the impact of different depth loss weights on Cityscapes-DVPS [9] in Tab. 1. We keep the other loss weights unchanged (where $\lambda_{cls} = 2$, $\lambda_{mask} = 5$ as in [1] and $\lambda_{sg} = \lambda_{dg} = 0.1$), and only modify λ_{depth}. Despite altering the depth loss weights, the panoptic segmentation (PQ) and quality of the depth map (abs rel) remain relatively stable. This outcome is akin to the results found in [11], where a unified architecture is less susceptible to weight loss choices, providing a significant advantage over prior approaches [9].

K	PQ	abs rel	DPQ
3	69.45	0.0633	62.95
5	**69.49**	**0.0632**	**62.97**
7	69.42	0.0635	62.91

Table 2: Ablations of bi-directional guidance learning with different patch size K.

[1]We use official pre-trained models of ResNet [3] from https://github.com/facebookresearch/detectron2 and Swin [6] from https://github.com/microsoft/Swin-Transformer

Furthermore, we conduct experiments using various configurations in bi-directional guidance learning, encompassing patch size K, applying layers L, and gap parameter α. The layers include $L = 0, 1, 2, 3$, which correspond to features of $\times 1/32$, $\times 1/16$, $\times 1/8$, $\times 1/4$ resolutions of the original images.

In Tab. 2, Tab. 3, and Tab. 4, we contrast the outcomes of various patch sizes, applying layers, and gap parameters, respectively. The results indicate that the performance remains relatively stable across different settings, and we select the optimal configuration based on experimental outcomes (*i.e.*, $K = 5$, $L =$ 1-3, and $\alpha = 0.3$).

L	PQ	abs rel	DPQ
0-3	**69.49**	0.0634	62.96
1-3	**69.49**	**0.0632**	**62.97**
0-2	69.47	0.0636	62.94

Table 3: Ablations of bi-directional guidance learning with different applying layers L.

α	PQ	abs rel	DPQ
0.1	69.47	0.0632	62.95
0.2	69.48	0.0631	**62.97**
0.3	**69.49**	0.0632	**62.97**
0.4	69.48	**0.0629**	62.96

Table 4: Ablations of bi-directional guidance learning with different gap parameter α.

3. Visualization Results

We provide additional visualizations of panoptic segmentation and depth estimation results for comparison on Cityscapes-DVPS and SemKITTI-DVPS datasets.

In Fig. 1, we present the ablation qualitative results of depth prediction with the proposed components applied. The results indicate that, with the extra backup query incorporated, the predicted depth map produces more previse depth values on filtered-out regions (depicted in black color in the segmentation result in the last row). Moreover, the bi-directional guidance learning incorporated in our model enhances its ability to capture more details, specifically on object boundaries, thus enhancing the quality of the predicted depth map.

In Fig. 2, we conduct a comparison of the visualizations of prediction results with recent state-of-the-art methods, namely PanopticDepth [8] and PolyphonicFormer [11]. We utilized the publicly available pre-trained networks with ResNet-50 backbones provided by the authors. Furthermore, in Fig. 3, we present additional visualization results on SemKITTI-DVPS.

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[4] Hyunyoung Jung, Eunhyeok Park, and Sungjoo Yoo. Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12642–12652, 2021. 1

[5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 1

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[8] Gao Naiyu, He Fei, Jia Jian, Shan Yanhu, Zhang Haoyang, Zhao Xin, and Huang Kaiqi. Panopticdepth: A unified framework for depth-aware panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 4

[9] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3997–4008, 2021. 1, 4, 5

[10] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1

[11] Haobo Yuan, Xiangtai Li, Yibo Yang, Guangliang Cheng, Jing Zhang, Yunhai Tong, Lefei Zhang, and Dacheng Tao. Polyphonicformer: Unified query learning for depth-aware video panoptic segmentation. *arXiv preprint arXiv:2112.02582*, 2021. 1, 2, 4

[12] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1

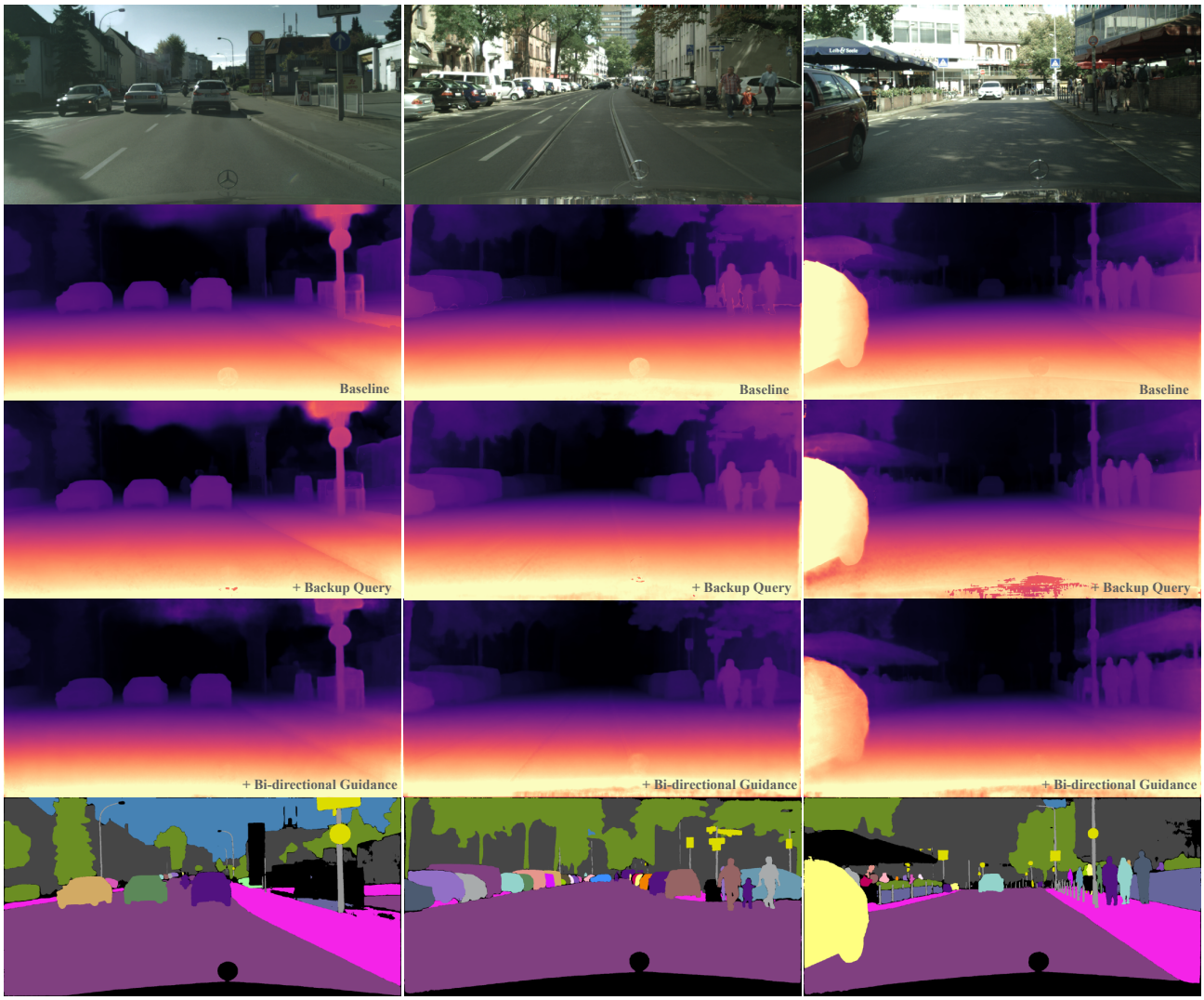

Figure 1: Qualitative results of depth predictions with proposed components applied.

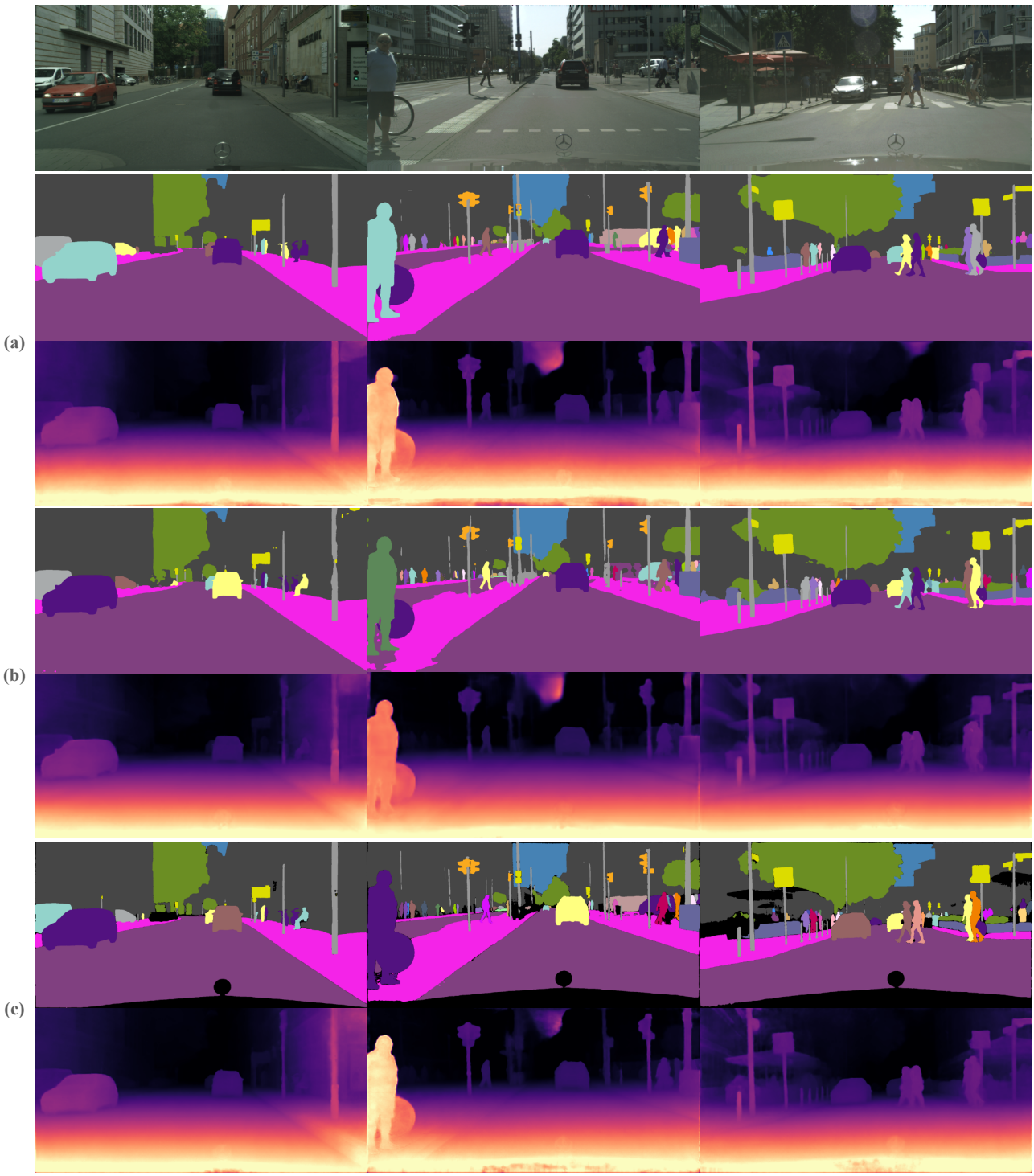

Figure 2: We compared the prediction visualizations of our method with those of previous methods on Cityscapes-DVPS [9]. Specifically, we evaluated the visualizations of (a) PanopticDepth [8], (b) PolyphonicFormer [11], and (c) our own method. Results are obtained from authors' public pretrained networks.

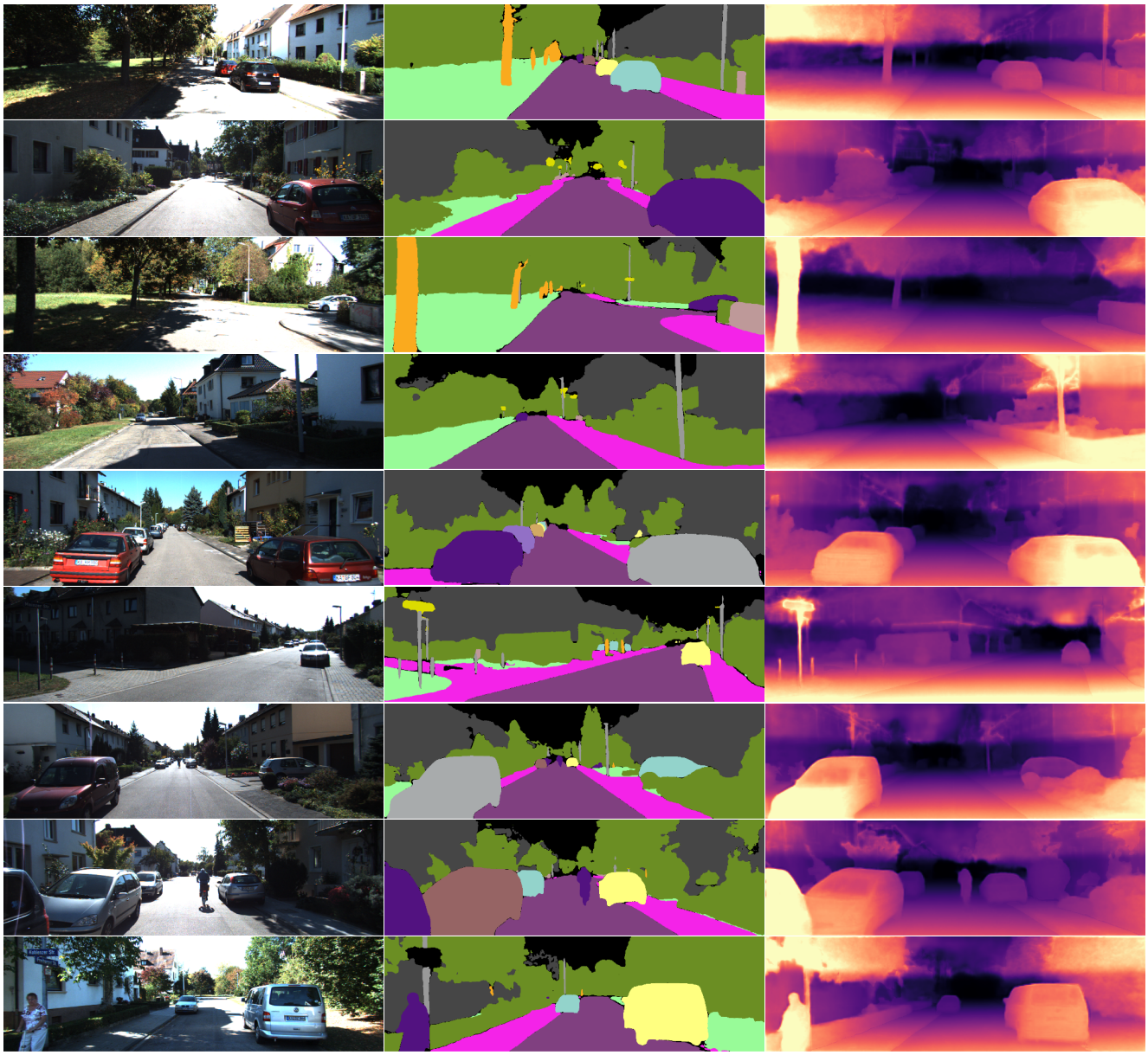

Figure 3: Visualizations results on SemKITTI-DVPS [9].