Appendix

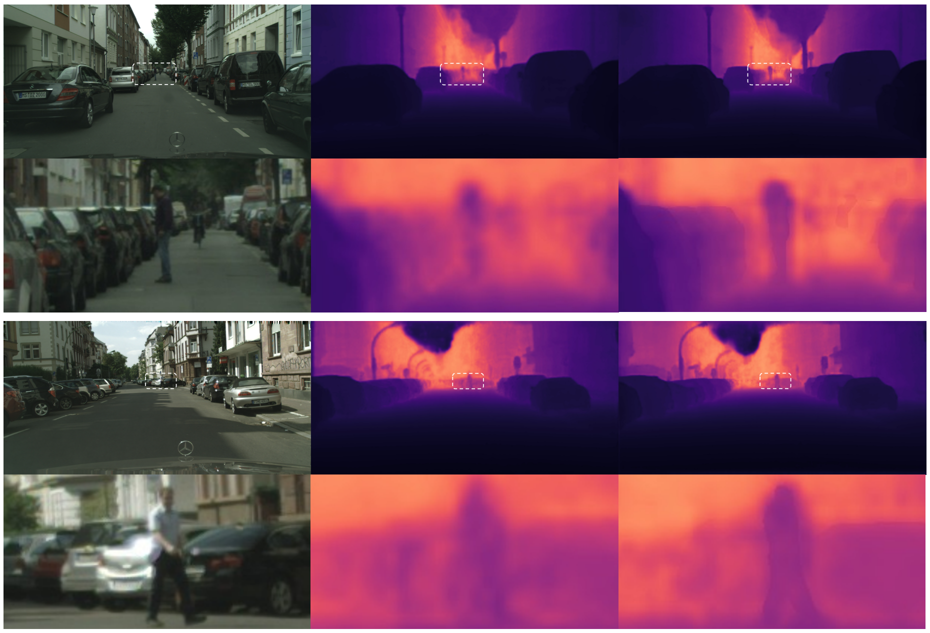

Fig. 1: Comparison of depth predictions of traditional dense depth predictions and depth predictions of PolyphonicFormer. From left to right: input images, traditional dense depth predictions, and depth predictions of PolyphonicFormer.

A Implementation Details

We report the implementation details in this section. To train the Polyphonic-Former, we first pre-train the backbone on ImageNet-1K [8] and the pre-train the panoptic path with Mapillary [5] and Cityscapes [3] datasets, following the ViP-Deeplab [6]. When performing the pre-training on Mapillary, we resize the original images to a random scale from 2048×1024 to 4096×2048 and randomly crop a 1024×1024 sample. We do the Mapillary pre-training for 300 epochs. For Cityscapes pre-training, we also perform random resize from 2048×1024 to 4096×2048, but crop to 2048×1024. After pre-training, we train the image baseline of PolyphonicFormer on Cityscapes-DVPS. Training on Cityscapes-DVPS requires 192 epochs and takes the same data augmentation strategy as the Cityscapes dataset. The depth ground truth needs to be divided by the resize scale factor because resizing an image means zooming the image for depth perception. The ablation studies are performed on the image baseline. With

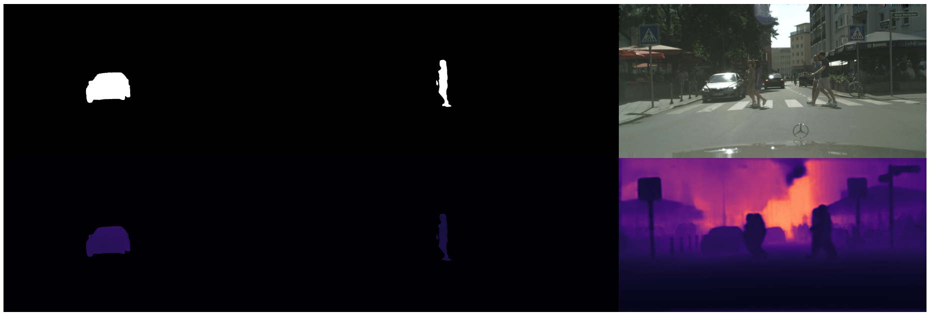

Fig. 2: We present several videos here. The left and middle are mask and depth results of tracked instance examples output by PolyphonicFormer. The depth results are merged into the final prediction (bottom right).

the image baseline, we fine-tune the PolyphonicFormer with a tracking head on Cityscapes-DVPS and SemKITTI-DVPS respectively for 48 epochs. For each sample, we randomly choose a reference frame from time $t-2$, $t-1$, $t+1$, and $t+2$ for source frame at time t. We do not perform random scale resizing on SemKITTI-DVPS, and only pad the images to 1280×384 instead, which is the minimum size that can be divided by 32 to cover all of the KITTI images. All of the datasets we used are *without* extra data with pseudo labels [2] for self-supervised or semi-supervised training. During inference, for simplicity, we use single scale inference with the original image size from the datasets, and we do not use the test time online depth refinement [1]. **In general**, except for that we do not use the *test-time augmentation* and *semi-supervised learning*, we adopt similar settings with ViP-Deeplab [6].

For the ICCV-2021 SemKITTI-DVPS Challenge submission, we take advantage of the validation set for training. Before and after adding the extra validation samples, PolyphonicFormer can achieve 63.6 and 64.6 DSTQ respectively in the SemKITTI-DVPS test set.

B More Visualization analyses

We show more visualization analysis results in Figure 1. The depth predictions of PolyphonicFormer are merged from depth predictions for each thing or stuff mask. As shown in Figure 1, the final depth predictions of PolyphonicFormer successfully distinguish the boundary between the instances or instances and corresponding background and thus are more accurate than the dense prediction results. We note that the results are presented with a high resolution, so we recommend the readers zoom in to check the details about the depth results of other instances.

We also illustrate the unified query learning with a video, as shown in Figure 2. The PolyphonicFormer generates temporal-consistent instance-level mask and depth predictions and merges them into the final results.

$DVPQ^k_\lambda$ on Cityscapes-DVPS	k = 1			k = 2			k = 3			k = 4			Average			FLOPs
PolyphonicFormer $\lambda = 0.50$	64.3	56.0	70.3	57.1	43.1	67.2	54.0	37.0	66.3	52.3	34.3	65.3	56.9	42.6	67.3	-
PolyphonicFormer $\lambda = 0.25$	59.7	53.3	64.4	53.0	41.3	61.5	49.9	35.3	60.5	48.6	33.0	60.0	52.8	40.7	61.6	-
PolyphonicFormer $\lambda = 0.10$	39.3	31.8	44.7	34.3	23.3	42.3	32.7	20.3	41.7	31.5	18.6	40.8	34.5	23.5	42.4	-
Average: PolyphonicFormer	54.4	47.0	59.8	48.1	35.9	57.0	45.5	30.9	56.2	44.1	28.6	55.4	**48.1**	35.6	57.1	411G
$DVPQ^k_\lambda$ on SemKITTI-DVPS	k = 1			k = 5			k = 10			k = 20			Average			FLOPs
PolyphonicFormer $\lambda = 0.50$	50.5	44.0	55.3	45.7	34.8	53.7	44.4	32.4	53.1	43.7	31.4	52.7	46.1	35.7	53.7	-
PolyphonicFormer $\lambda = 0.25$	47.9	42.2	52.1	43.2	33.3	50.4	42.0	31.1	49.9	41.3	30.3	49.4	43.6	34.2	50.5	-
PolyphonicFormer $\lambda = 0.10$	35.9	33.6	37.6	31.2	25.2	35.5	29.6	22.9	34.5	28.5	21.5	33.6	31.3	25.8	35.3	-
Average: PolyphonicFormer	44.8	39.9	48.3	40.0	31.1	46.5	38.7	28.8	45.8	37.8	27.7	45.2	**40.3**	31.9	46.5	99G

Table 1: Experimental results on Cityscapes-DVPS and SemKITTI-DVPS datasets with Resnet-50 backbone. Each cell shows $DVPQ^k_\lambda$ | $DVPQ^k_\lambda$-Thing | $DVPQ^k_\lambda$-Stuff where λ is the threshold of relative depth error, and k is the number of frames. Smaller λ and larger k correspond to a higher accuracy requirement. We also estimate the computational cost (FLOPs) of ViP-Deeplab with Resnet-50 backbone and get 1,096G and 280G on Cityscapes-DVPS and SemKITTI-DVPS respectively.

method	abs rel	sq rel	RMSE	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
DPT-Hybrid [7]	0.0697	0.4515	4.115	0.1106	0.9434	0.9914	0.9976
PolyphonicFormer	**0.0647**	**0.3454**	**3.800**	**0.1013**	**0.9524**	**0.9950**	**0.9985**

Table 2: Comparison results of PolyphonicFormer and the representative depth estimation method. The metrics with orange background means "**lower** is better". The metrics with blue background means "**higher** is better". ViP-Deeplab [6] has a **0.0721** abs rel.

C More Experiments

We report more results of PolyphonicFormer in this section. The results of PolyphonicFormer with Swin-B backbone are already provided, and we report the DVPQ results with Resnet-50 backbone in this section. As shown in Table 1, with a Resnet-50 backbone, the PolyphonicFormer achieves **48.1**, and **40.3** in DVPQ on the Cityscapes-DVPS and SemKITTI-DVPS datasets, respectively.

We compare PolyphonicFormer with recently proposed DPT [7], which is one of the state-of-the-art supervised monocular depth estimation methods on KITTI (eigen split) [4]. As the KITTI dataset lacks the panoptic segmentation annotation and the SemanticKITTI dataset has a very different split strategy compared with eigen split, we cannot directly get the results on the KITTI eigen split. We adopt the pre-trained model of DPT-Hybrid on MIX6 [7] (meta-datasets containing 10 datasets) and KITTI eigen split, and fine-tune on Cityscapes-DVPS with the same schedule of PolyphonicFormer. As in Table 2, our proposed PolyphonicFormer outperforms the DPT-Hybrid [7] and ViP-Deeplab [6].

D More Visualization Results

We show some of the visualization results from the Cityscapes-DVPS and SemKITTI-DVPS datasets along with PolyphonicFormer (Swin-B backbone) predictions in Figure 3 and Figure 4.

References

1. Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In: AAAI (2019)
2. Chen, L.C., Lopes, R.G., Cheng, B., Collins, M.D., Cubuk, E.D., Zoph, B., Adam, H., Shlens, J.: Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In: ECCV (2020)
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
4. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV (2015)
5. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: ICCV (2017)
6. Qiao, S., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. CVPR (2021)
7. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021)
8. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV (2015)

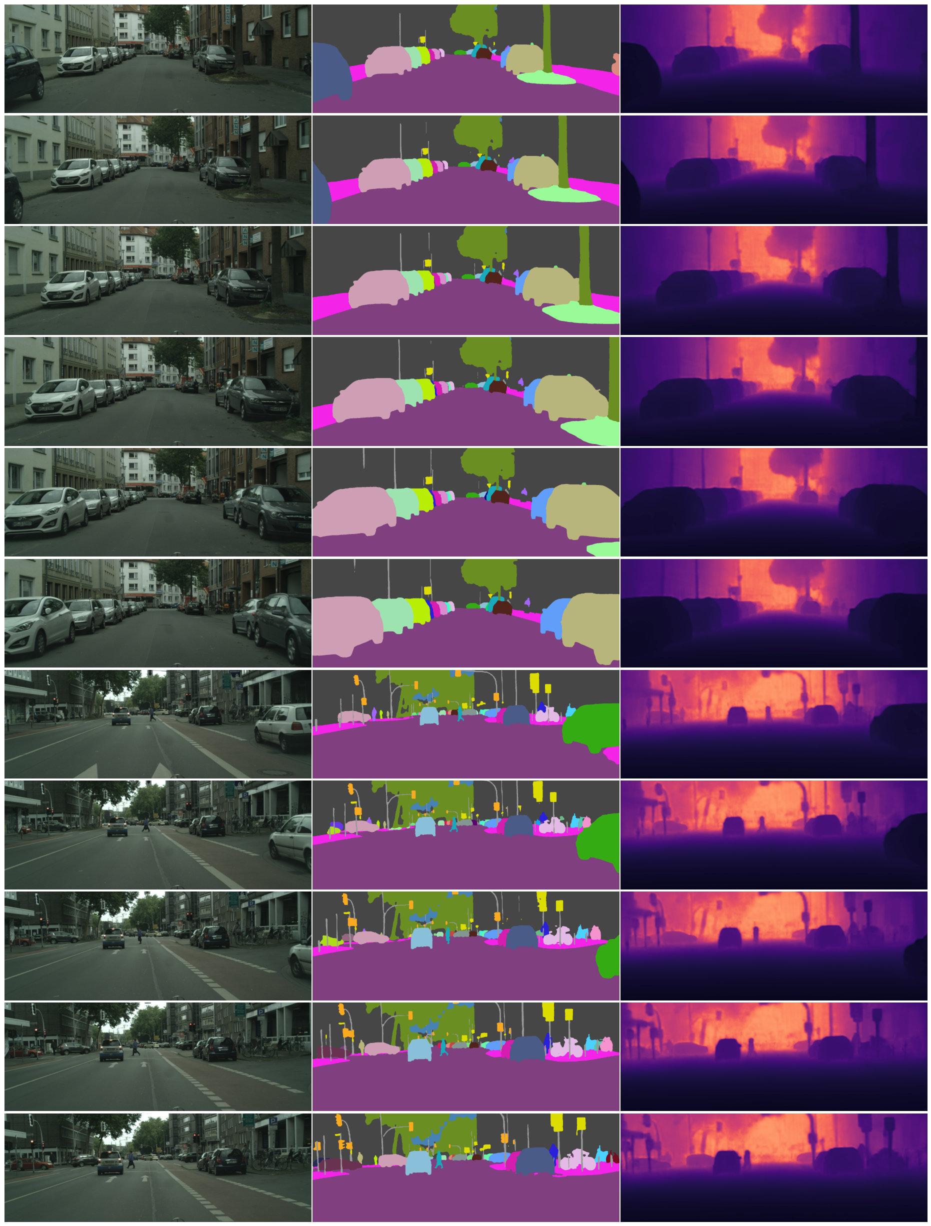

Fig. 3: Prediction visualizations on Cityscapes-DVPS. From left to right: input images, temporally consistent panoptic segmentation (TCPS), and depth predictions. Color change of the same instance of TCPS indicates an id switch.

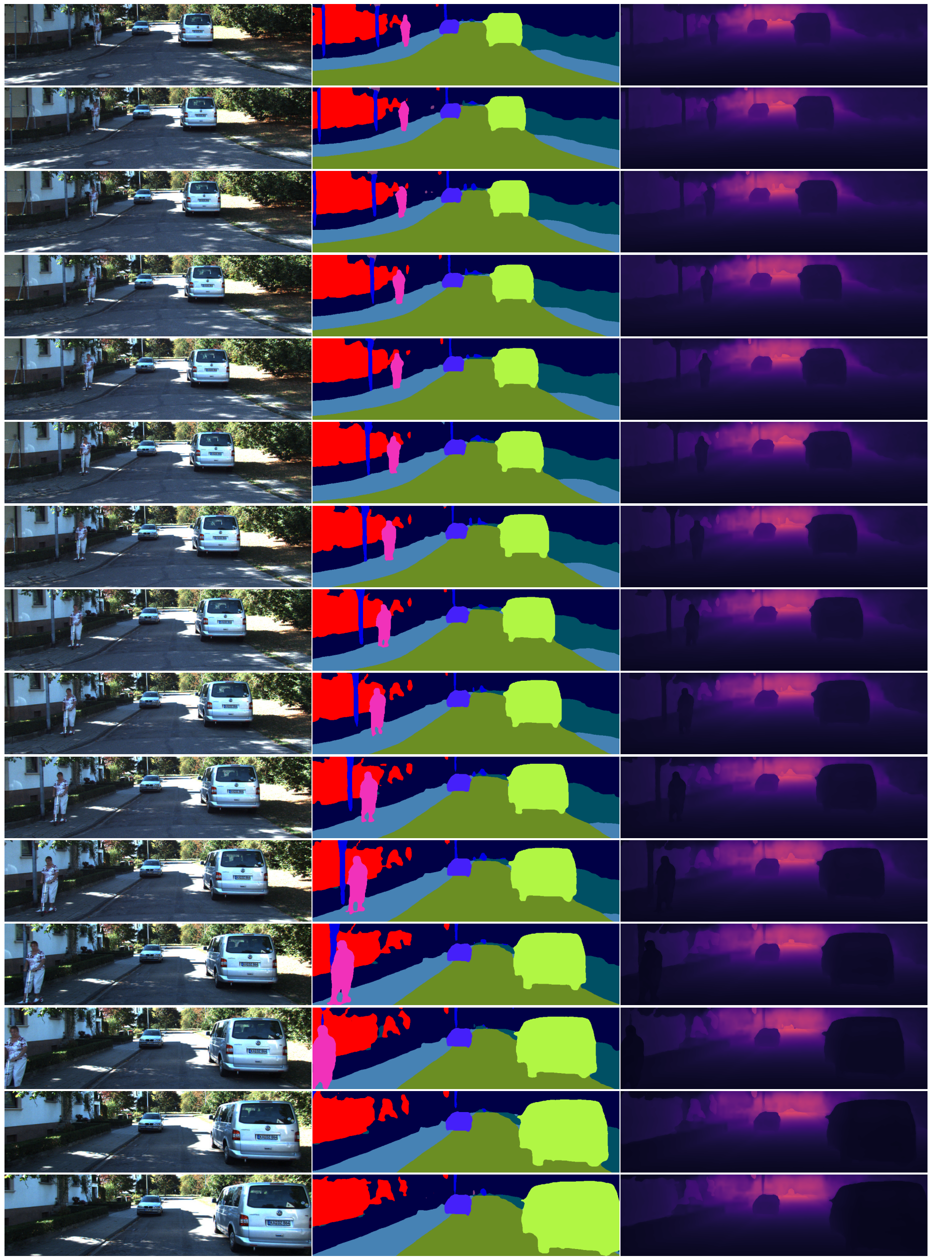

Fig. 4: Prediction visualizations on SemKITTI-DVPS. From left to right: input images, temporally consistent panoptic segmentation (TCPS), and depth predictions. Color change of the same instance of TCPS indicates an id switch.