MonoDVPS: A Self-Supervised Monocular Depth Estimation Approach to Depth-aware Video Panoptic Segmentation

Supplementary Material

A. Experiments

A.1. Depth-aware Video Panoptic Segmentation

In Table 5 we present a comparison between our MonoDVPS network and concurrent work ViP-DeepLab [1] on the DVPS task. We train our network on the original Cityscapes-DVPS training set and obtain 43.4 DVPQ, while on the extended dataset with panoptic pseudo-labels, we achieve 48.8 DVPQ. We surpass ViP-DeepLab [1] with the ResNet-50 backbone on the DVPQ score, while having a fast inference speed.

Model	Backbone	DVPQ	DVPQ-Things	DVPQ-Stuff	Time (s)
MonoDVPS	ResNet-50	48.8	31.0	61.7	0.11
MonoDVPS*	ResNet-50	43.4	26.2	55.9	0.11
ViP-DeepLab [1]	ResNet-50	42.0	27.6	51.5	-

Table 5: **DVPS on Cityscapes-DVPS.** MonoDVPS* is our network trained on the reduced training set (without extension). ViP-DeepLab with ResNet-50 was evaluated with the author's code [1].

A.2. Panoptic-guided Moving Object Masking for Improved Depth Ablation

For each instance in frame t, we measure the IoU between its mask in the reconstructed panoptic label $P_{s\rightarrow t}$ and its mask in P_t. The geometric projection model used for generating $P_{s\rightarrow t}$ assumes the scene is static and considers only the ego-motion. Therefore, we observe a high overlap between instance masks for static objects and low overlap for moving objects, since object motion was not modeled. We set a threshold T such that if the IoU is lower than the threshold, the instance is considered a moving object and the pixels corresponding to its mask will be ignored in the photometric loss computation. In Table 6, we experiment with $T = \{0.3, 0.5, 0.7\}$ and a linear scheduling. We need to consider that errors from warping with optical flow, geometric reconstruction or occlusions could influence the IoU computation. In consequence, a high threshold $T = 0.7$ removes too many instances, while a low threshold $T = 0.3$ is too permissive. The linear scheduling obtains the best balance between panoptic and depth performance, with $T = 0.5$ being a close second.

IoU threshold	PQ ↑	absRel ↓
0.3	63.2	0.099
0.5	63.5	0.098
0.7	63.9	0.102
linear	**63.6**	**0.098**

Table 6: **Moving Object Masking.** Ablation study on the IoU threshold used to determine if an object is moving. *Linear* means that the IoU is decreased linearly from 0.7 with each training iteration.

A.3. Panoptic-guided Depth Losses Ablation

In Table 7 we perform an extensive ablation study for depth estimation. We evaluate the depth output of our multi-task depth-aware panoptic segmentation network on the Cityscapes-DVPS dataset. Specifically, we introduce three panoptic-guided depth losses and evaluate their individual contributions. As seen in Table 7, the panoptic-guided triplet loss $\mathcal{L}_{PGT}$ brings the largest improvement compared to the other two panoptic-guided losses $\mathcal{L}_{PGS}$ and $\mathcal{L}_{PED}$. This could be because $\mathcal{L}_{PGT}$ is less sensitive to errors in the panoptic predictions due to its patch-based formulation. However, we obtain the best results when all three losses $\mathcal{L}_{PGS}$, $\mathcal{L}_{PED}$, $\mathcal{L}_{PGT}$ are used during training.

B. Implementation Details

We adopt the ResNet-50 [4] backbone for the depth-aware video panoptic segmentation network. The network is pretrained on the Cityscapes dataset [2] for image panoptic segmentation. The pose estimation network follows [3] with a ResNet18 backbone and a decoder that predicts the 6DOF camera pose, the translation vector and rotation matrix, as Euler angles. During inference we discard the pose

Model	absRel ↓	sqRel ↓	RMS ↓
MTL Self-Supervised Depth	0.106	0.841	5.270
+ Loss Balancing	0.102	0.767	5.034
+ $\mathcal{L}_{PGS}$	0.101	0.781	5.010
+ $\mathcal{L}_{PED}$	0.101	0.778	5.024
+ $\mathcal{L}_{PGT}$	0.100	0.757	4.998
+ $\mathcal{L}_{PGS}$ + $\mathcal{L}_{PED}$ + $\mathcal{L}_{PGT}$	0.099	0.747	4.988
+ Moving Objects Masking	0.098	0.701	4.864
+ Extended dataset	**0.082**	**0.515**	**4.198**

Table 7: **Panoptic-guided Depth Evaluation.** Results on Cityscapes-DVPS. Ablation study for panoptic-guided depth losses and moving objects masking.

estimation network. During training, we employ a mini-batch of 4 images for 30k iterations, using the Adam optimizer with a base learning rate of $1e-3$ for decoders and heads and $1e-4$ for the backbone and polynomial learning rate decay. We adopt image augmentation, such as random horizontal flip and random color augmentation: brightness, contrast, saturation and hue jitter. We employ image resolution 1025×2049 for Cityscapes-DVPS and 385×1281 for SemKITTI-DVPS. For depth evaluation, we center crop the Cityscapes-DVPS image to 512×1664, in order to discard the sky and ego-vehicle regions, following [3].