Supplementary Material for "SurroundDepth: Entangling Surrounding Views for Self-Supervised Multi-Camera Depth Estimation"

Yi Wei[1,2]*, Linqing Zhao[3]*, Wenzhao Zheng [1,2], Zheng Zhu [4], Yongming Rao[1,2],

Guan Huang [4], Jiwen Lu[1,2]†, Jie Zhou[1,2]

[1]Beijing National Research Center for Information Science and Technology, China
[2]Department of Automation, Tsinghua University, China
[3]School of Electrical and Information Engineering, Tianjin University, China
[4]PhiGent Robotics

Figure 1: The occlusion masks of DDAD dataset [1].

A Datasets and Implementation Details

Datasets: We conduct experiments on both the Dense Depth for Automated Driving (DDAD) [1] and nuScenes datasets [2]. The DDAD dataset is a large-scale dataset captured with six synchronized cameras for autonomous driving. The dataset contains 12,650 training samples, including 75,900 images for six cameras. The validation set has 3,950 samples (15,800 images) and ground truth depth maps, which are only utilized for evaluation. During the evaluation procedure, we consider the distance up to 200m and do not crop depth maps. As shown in Figure 1, we can find that each view in DDAD dataset has self-occlusion areas caused by the current vehicle. Following [3], we annotate occlusion masks manually and use them to reweight photometric loss. The nuScenes [2] dataset consists of 1000 sequences of various scenes captured in both Boston and Singapore, where each sequence is approximately 20 seconds long. The dataset is officially partitioned into training, validation, and testing subsets with 700, 150, and 150 sequences, respectively. For each sample, we have access to the six surrounding cameras as well as the camera calibrations. We filter out static frames in nuScenes dataset.

Implementation Details: We use Adam [4] as our optimizer with $\beta 1 = 0.9$ and $\beta 2 = 0.999$. The initial learning rate is set as 1e-4. For both spatial and temporal photometric loss, we utilize 2 neighboring frames as source frames. Specifically, I^i_{t-1}, I^i_{t+1} and I^{i-1}_t, I^{i+1}_t frames are used as temporal and spatial contexts. Following Monodepth2 [5], we set SSIM weight as 0.85 and depth smoothness weight as 1e-3.

B Evaluation Metrics

The detailed evaluation metrics of self-supervised depth estimation can be described as follows:

- Abs Rel: $\frac{1}{|T|}\sum_{d\in T}|d-d^*|/d^*$

*Equal contribution.
†Corresponding author.

6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.

Figure 2: Scale-aware SfM pretraining. Due to the small overlap and large view changes, conventional two-frame Structure-from-Motion will produce many wrong correspondences. By introducing region masks, we reduce the correspondences searching scope and improve the quality. With the camera extrinsic matrices, we can further filter outliers leveraging epipolar geometry. **Better viewed when zoomed in.**

Method	Abs Rel	Sq Rel	$\delta < 1.25$
pretrain	0.286	4.698	0.550
joint train	0.259	4.177	0.619
finetune	**0.208**	**3.371**	**0.693**

Table 1: The ablation study for training method. 'pretrain' indicates the performance of SfM pretrained model. 'joint train' means that we simultaneously use SfM pseudo depths and temporal-spatial photometric loss to supervise the networks.

- Sq Rel: $\frac{1}{|T|}\sum_{d\in T}||d-d^*||^2/d^*$
- RMSE: $\sqrt{\frac{1}{|T|}\sum_{d\in T}||d-d^*||^2}$
- RMSE log: $\sqrt{\frac{1}{|T|}\sum_{d\in T}||\log d-\log d^*||^2}$
- $\delta < t$: % of d s.t. $\max(\frac{d}{d^*},\frac{d^*}{d}) = \delta < t$

where d and d^* indicate predicted and groundtruth depths respectively, and T indicates all pixels on the depth image D. During evaluation, conventional self-supervised monocular depth estimation methods use the factor $\frac{median(D^*)}{median(D)}$ to align the scale. However, in scale-aware experiments, we do not need to perform this scale alignment.

C Scale-aware Structure-from-Motion Pretraining

We provide an example to show the effectiveness of our scale-aware pretraining in Figure 2. As an alternative way to leverage SfM points, we simultaneously use SfM pseudo depths and temporal-spatial photometric loss to supervise the model. From Table 1, we find that since SfM points are sparse and non-uniform and cannot provide strong supervision, this method will get worse results. Further, the pretrained model is not good enough and we need to finetune them with spatial and temporal photometric loss.

D Visualization

Qualitative Results Fig. 3 shows qualitative results on DDAD validation set. Our SurroundDepth can predict visually appealing results on all six views. For the occlusion areas, our method can

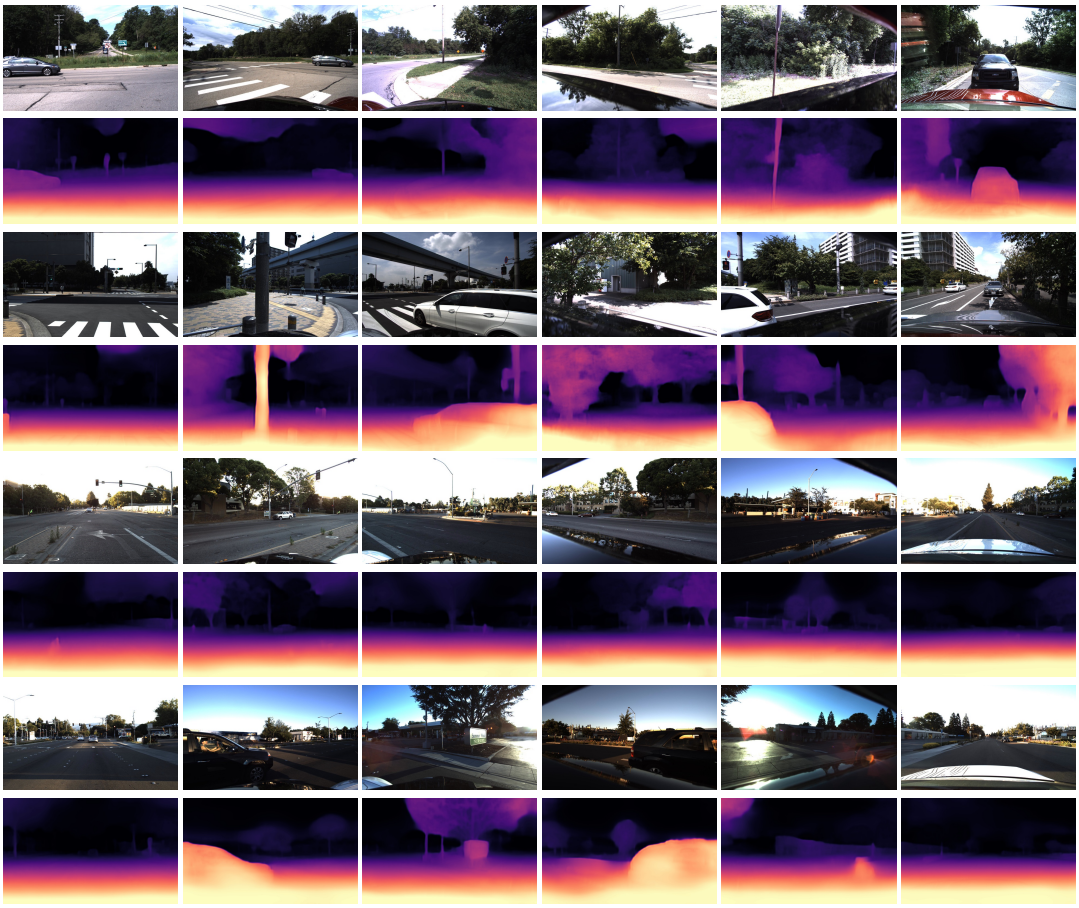

Figure 3: Qualitative results on DDAD [1] dataset. Our method can predict visually appealing depth maps with texture details.**Better viewed when zoomed in.**

surprisingly inpaint them and predict correct depths. Moreover, the generated depth maps preserve the texture details of corresponding RGB images.

Attention Distribution We visualize cross-view attention maps at the smallest scale in Fig. 4. For a set of query points, the cross-view attention maps can highlight the feature map at the corresponding locations in other views, demonstrating that our cross-attentions are able to entangle multi-view features to predict depths jointly.

References

[1] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, pages 2485–2494, 2020.

[2] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020.

[3] V. Guizilini, I. Vasiljevic, R. Ambrus, G. Shakhnarovich, and A. Gaidon. Full surround monodepth from multiple cameras. *arXiv preprint arXiv:2104.00152*, 2021.

[4] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[5] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3828–3838, 2019.

Figure 4: The visualization of cross-view attention maps. Given a set of query points, our cross-view attention maps will highlight those features at the corresponding locations in other views. The peaks of attention maps tend to appear in the overlapping areas of adjacent images, which proves that our CVT can accurately entangle the features from multiple views to achieve joint prediction. The predictions are made on the validation set at the smallest scale. **Better viewed when zoomed in.**