JPerceiver: Joint Perception Network for Depth, Pose and Layout Estimation in Driving Scenes (Supplementary Material)

Haimei Zhao[1], Jing Zhang[1], Sen Zhang[1], and Dacheng Tao[2,1]

[1] The University of Sydney, 6 Cleveland St, Darlington, NSW 2008, Australia
[2] JD Explore Academy, China
{hzha7798,szha2609}@uni.sydney.edu.au
jing.zhang1@sydney.edu.au dacheng.tao@gmail.com

Abstract. This document provides supplementary information on (1) more additional quantitative and qualitative evaluation results and analysis towards BEV layout estimation, depth estimation and visual odometry, (2) the ablation study of the network architecture and loss designing, and (3) the implementation details.

S1 Additional Evaluation Results

S1.1 Layout Estimation

BEV layout estimation on Nuscenes. In nuscenes[2], we compare with more recent methods that either take six-camera [10,14,11,8] or one-camera [9,17] images as input, and evaluate it under two settings, i.e., predicting a BEV layout in a range of $50m \times 50m$ (denoted as Setting 1 in Table 1) and $100m \times 100m$ (denoted as Setting 2 in Table 1). It is noteworthy that we choose to predict in an area of the same size for fairness, which covers the range of Z m in front of the ego vehicle and horizontally covers $\frac{Z}{2}$ m to the left and right. The results of six-camera methods are retrained to predict only two classes (drivable area and car). And the results of PYVA [17] are also trained using their provided codes on the Nuscenes dataset. According to Table 1, while our method takes only one-camera images as input, we have achieved comparable or superior results to the SOTA methods [14,10] that take six-camera input under Setting 1. In addition, our method also surpasses the other methods that take one-camera input in both settings by a large margin.

Ablation of different losses. We conduct an ablation study of different components of Hybrid Loss on the KITTI Object dataset. As shown in Table 2, Hybrid loss achieves superior performance w.r.t. both mIoU and mAP, consisting of CE, IoU and Boundary Losses.

Qualitative results. As shown in Fig. 2, we compare the estimated BEV layouts with the state-of-the-art (SOTA) method and corresponding ground truth in various test scenarios of different datasets, including Argoverse [3] (top), Nuscenes [2] (middle) and KITTI [5] (bottom). Notably, the estimation results

Table 1: Quantitative comparisons on Nuscenes [2]. "c" and "d" in the Input column denote camera images and depth maps, e.g. "6c6d" means six RGB images and six depth maps are taken as input.

Methods	Input	Nuscenes Road				Nuscenes Vehicle			
		Setting 1		Setting 2		Setting 1		Setting 2	
		mIoU(%)	mAP(%)	mIoU(%)	mAP(%)	mIoU(%)	mAP(%)	mIoU(%)	mAP(%)
VED[9]	1c	63.8	-	-	-	15.6	-	-	-
PON[12]	6c	70.5	-	-	-	27.6	-	-	-
VPN[10]	6c6d	69.4	-	-	-	28.3	-	-	-
Lift-Splat[11]	6c	-	-	72.94	-	-	-	32.1	-
Fiery[8]	6c	-	-	-	-	37.7	-	35.8	-
PYVA[17]	1c	77.09	86.19	66.55	80.42	24.34	39.96	20.15	29.29
JPerceiver	1c	**79.02**	**90.73**	68.54	**84.73**	33.01	**49.85**	24.90	**41.12**

Table 2: The experiment results of ablation study of different losses used for layout estimation

KITTI 3D Object		
Loss items	mIoU(%)	mAP(%)
CE	39.45	53.89
IoU	41.11	55.80
Boundary	36.08	60.2
CE+IoU	40.46	56.62
CE+Boundary	39.71	57.47
IoU+Boundary	40.33	56.35
CE+IoU+Boundary	40.85	57.23

of different semantic categories, i.e. road and vehicle layout are estimated simultaneously in our JPerceiver, while manually overlayed from twice inference results for the prior method [17].

S1.2 Depth Estimation

In Fig. 3, we demonstrate more visualization results of depth estimation and the comparison with our baseline method Monodepth2 [6] on the validation or test sets of three datasets, i.e. Argoverse [3], Nuscenes [2] and KITTI [5].

Analysis of CGT loss and depth estimation. The CGT scale loss does not use all pixels in the road plane but chooses the region with vehicle occupancy removed to impose scale constraint. Due to the flat-ground assumption and the few-pixel constraint, CGT does not act as strict supervision but only provides scale at a limited range. Thus, we analyze the relevance between depth metrics and distances. As shown in Fig. 1, the metrics and scale factor fluctuate with regard to distance but just slightly. This is expected since long-distance depth estimation is more difficult than that in close range.

S1.3 Visual Odometry

The visual odometry trajectories on KITTI test sets 07 and 10 are visualized in Fig. 4. The left part shows the trajectories without scaling, which demonstrate the absolute scale can be learned in our method. Scaled trajectories using the scaling ratio obtained from ground truth are listed in the right part, showing our predicted trajectory is much closer to the ground truth.

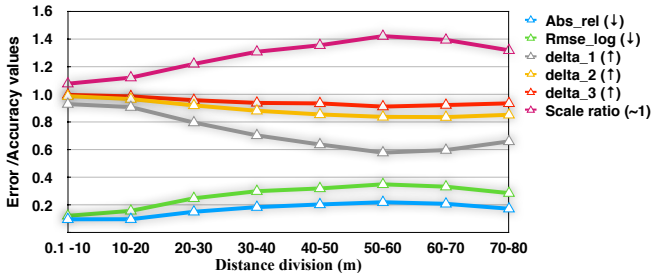

Fig. 1: The relevance between the depth metrics and distances.

Table 3: The comparison of Visual Odometry. t_{err} is the average translational RMSE drift (%) on length from 100, 200 to 800 m, and r_{err} is average rotational RMSE drift ($^\circ/100m$) on length from 100, 200 to 800 m.

Methods	Scaling	Sequence 09		Sequence 10	
		t_{err}	r_{err}	t_{err}	r_{err}
SfMLearner[19]	GT	11.32	4.07	15.25	4.06
GeoNet[18]	GT	28.72	9.8	20.73	9.04
Monodepth2[6]	GT	11.47	3.2	11.60	5.72
SC-Sfmlearner[1]	GT	7.64	2.19	10.74	4.58
Dnet[16]	Camera height	7.23	1.91	13.98	4.07
LSR[15]	None	**5.93**	1.67	10.54	4.03
JPerveiver	None	6.81	**1.18**	**6.92**	**1.47**

Comparison with more methods on test sequences 09 and 10. Following the commonly used protocol in self-supervised depth estimation and visual odometry, we retrain our models using sequences 00-08 as training set and 09-10 as the test set. For the data without layout labels, we use the models pretrained on those sequences with the layout labels to generate pseudo labels, which is demonstrated feasible to provide absolute scale for self-supervised depth estimation and visual odometry. Due to the promising ability of generalization, the pretrained model can be used to generate labels for more data set to help complete scale-aware perception in future work. The quantitative comparison of self-supervised visual odometry is shown in Table 3.

S2 Network Architecture

S2.1 Ablation Study for Network Architecture

To explore the effectiveness of the joint learning architecture, we complete an additional ablation study via using one encoder as the feature extractor for depth network and layout network on the KITTI Odometry dataset. As shown in Table 4 for layout estimation, Table 6 for depth estimation and Table 5 for visual odometry, the effectiveness of all three tasks have deteriorated using a shared encoder, especially for depth estimation and visual odometry.

Table 6: Ablation study of network architecture for depth estimation. "Scale factor" is calculated during inference. "w" and "w/o" denote evaluation results with or without rescaling by the scale factor. "E_S" denotes the variant using shared encoder.

Methods	Scaling	Abs Rel (↓)	Sq Rel(↓)	RMSE(↓)	RMSE log(↓)	Scale factor
JPerceiver	w	0.116	**0.517**	**3.573**	**0.180**	1.065 ± 0.071
	w/o	**0.112**	**0.559**	**3.817**	**0.196**	–
Jperceiver $-E_S$	w	0.133	0.646	3.853	0.195	0.990 ± 0.082
	w/o	0.137	0.666	3.867	0.200	–

Table 4: Road layout estimation results on KITTI Odometry. "E_S" denotes the variant using shared encoder.

KITTI Odometry Road		
Methods	mIoU(%)	mAP(%)
JPerceiver	**78.13**	**89.57**
JPerceiver$-E_S$	77.53	88.16

Table 5: The comparison of Visual Odometry. t_{err} is the average translational RMSE drift (%) on length from 100, 200 to 800 m, and r_{err} is average rotational RMSE drift ($^\circ/100m$) on length from 100, 200 to 800 m. "E_S" denotes the variant using shared encoder.

Methods	Scaling	Sequence 07		Sequence 10	
		t_{err}	r_{err}	t_{err}	r_{err}
JPerveiver	None	**4.57**	**2.94**	**7.52**	**3.83**
JPerveiver$-E_S$	None	9.73	5.72	15.75	6.9

S2.2 Network Details

Input and Output. We take in the RGB images of size 1024×1024 as input and output depth map, BEV layout and poses simultaneously. For layout network, the estimated BEV layouts of size 256×256 represent a specific region in the BEV plane, such as $40m \times 40m$ in Argoverse [3] and KITTI [5], and $50m \times 50m$ or $100m \times 100m$ in Nuscenes [2].

Encoder. We take ResNet-18 [7] as the backbone of our feature extractor for three tasks. Following the baseline method [6], we start training with weights pretrained on ImageNet [13]. The encoder of the pose network is modified to take two-frame pair as input.

Task-specific Decoder. The decoder of depth network is similar to [6], using sigmoid activation functions in multi-scale side outputs and ELU nonlinear functions [4] otherwise. While the decoder of the pose network consists of three convolution layers to predict a 6-DoF relative pose. The decoder of the layout network is composed of four deconvolution blocks to upsample the feature maps and decrease the number of feature channels, which finally arrive at the size of $256 \times 256 \times 2$ and then processed by a non-linear layer to obtain the layout.

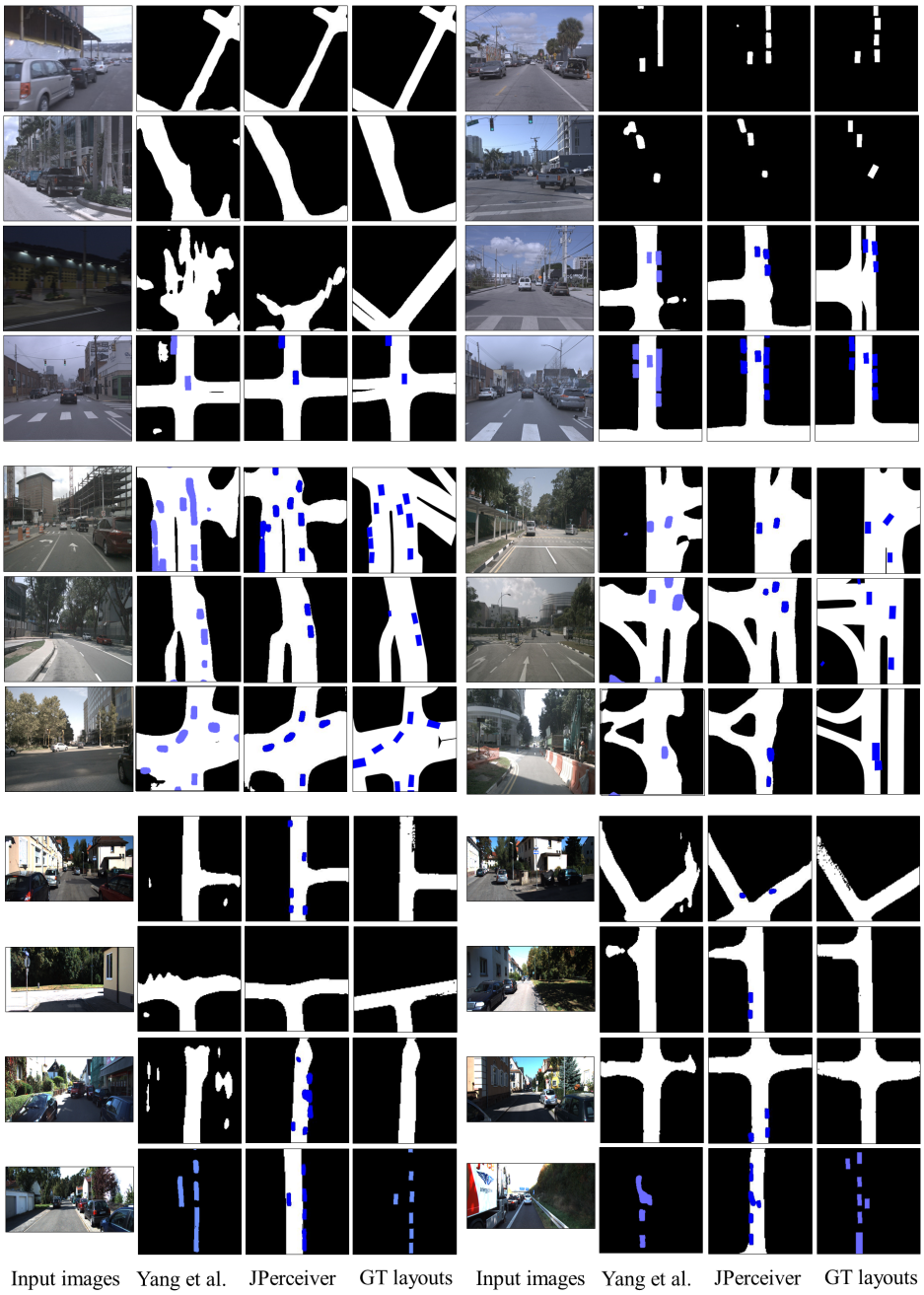

Fig. 2: Visualization examples of the road and vehicle BEV layouts on Argoverse [3], Nuscenes [2] and KITTI [5], compared with the SOTA method [17]. In the KITTI dataset, the Odometry and Raw split only provide road layout label while the Object split only provide vehicle layout label. However, our JPerceiver can predict the layouts of roads and vehicles simultaneously after training different branches on different splits, which demonstrates its superior ability of generalization in unseen scenarios. More visualization results are shown in the supplemental videos.

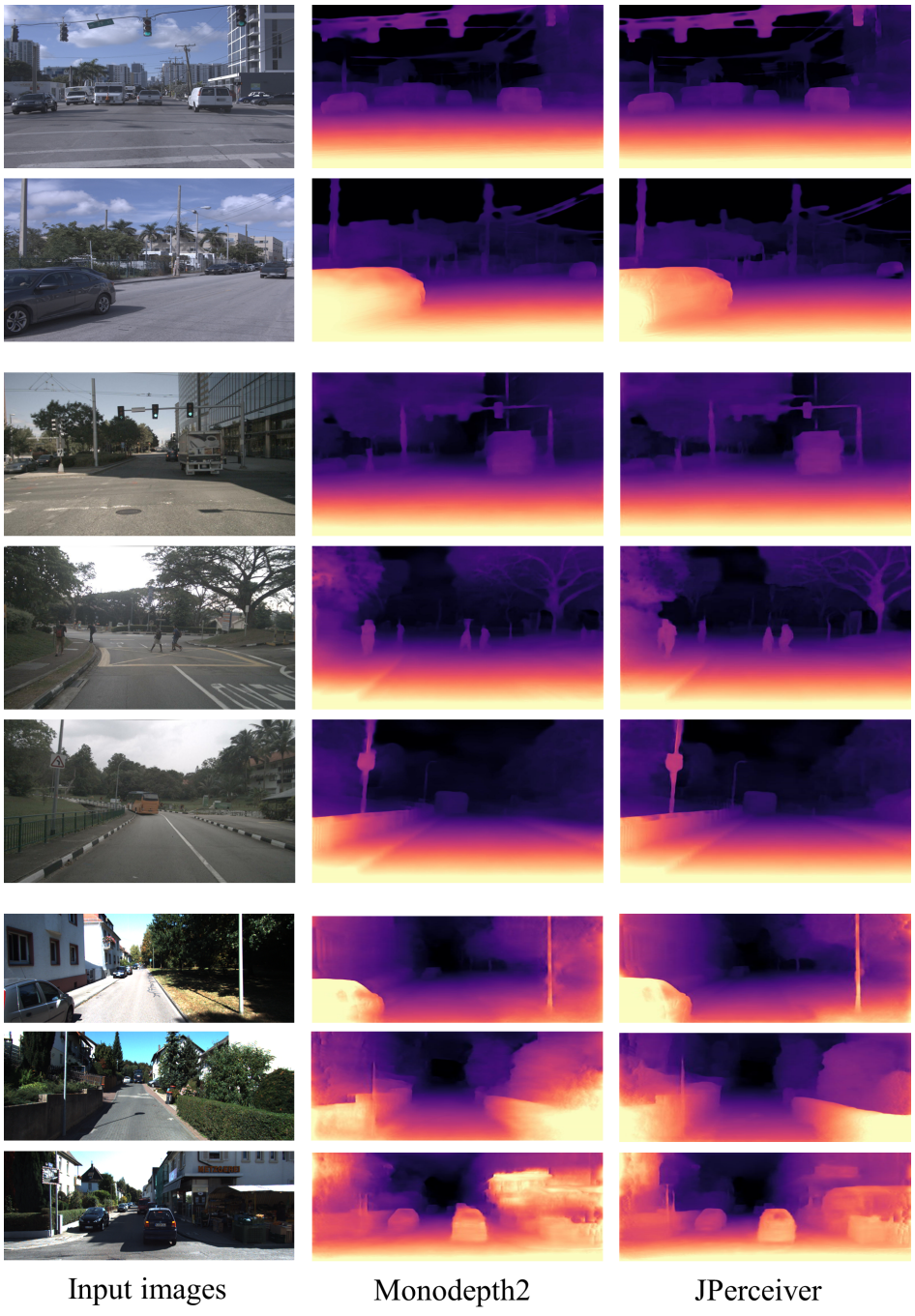

Fig. 3: Visualization of predicted depth maps and qualitative comparison with our baseline method [6] on the Argoverse [3] (top part), Nuscenes [2] (middle part) and KITTI [5] (bottom part) datasets.

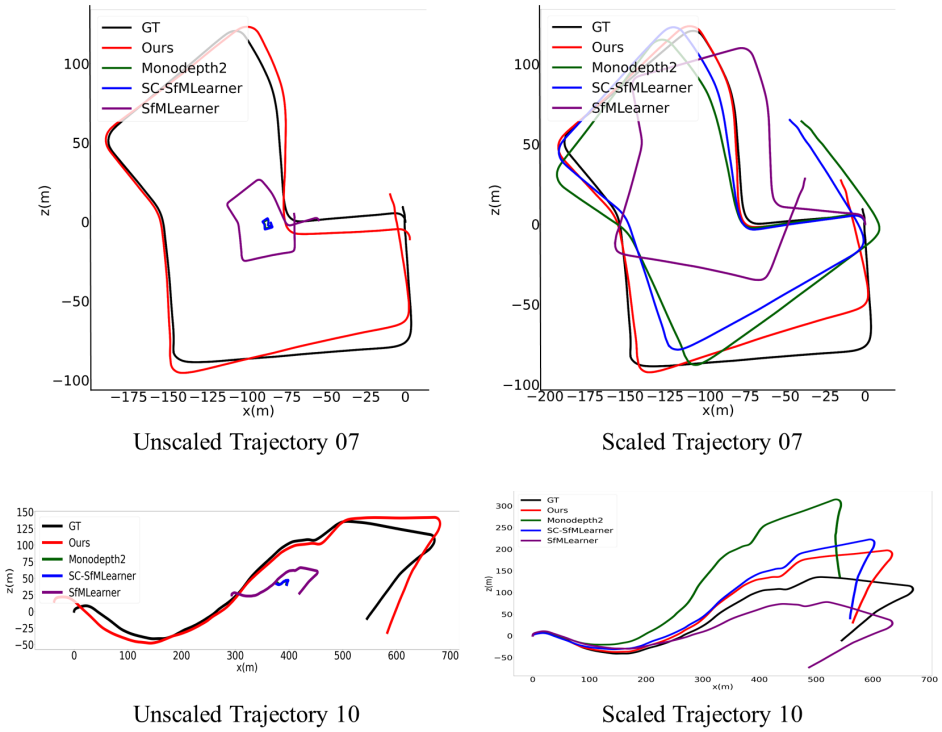

Fig. 4: We demonstrate the visual odometry trajectory before and after scaling on KITTI Odometry sequences 07 and 10. More visualization results are shown in the supplemental videos.

References

1. Bian, J.W., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M.M., Reid, I.: Unsupervised scale-consistent depth and ego-motion learning from monocular video. Advances in Neural Information Processing Systems (2019)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
3. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., et al.: Argoverse: 3d tracking and forecasting with rich maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8748–8757 (2019)
4. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 (2015)
5. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361. IEEE (2012)
6. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth prediction (October 2019)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
8. Hu, A., Murez, Z., Mohan, N., Dudas, S., Hawke, J., Badrinarayanan, V., Cipolla, R., Kendall, A.: FIERY: Future instance segmentation in bird's-eye view from surround monocular cameras. In: Proceedings of the International Conference on Computer Vision (ICCV) (2021)
9. Lu, C., van de Molengraft, M.J.G., Dubbelman, G.: Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks. IEEE Robotics and Automation Letters **4**(2), 445–452 (2019)
10. Pan, B., Sun, J., Leung, H.Y.T., Andonian, A., Zhou, B.: Cross-view semantic segmentation for sensing surroundings. IEEE Robotics and Automation Letters **5**(3), 4867–4873 (2020)
11. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: European Conference on Computer Vision. pp. 194–210. Springer (2020)
12. Roddick, T., Cipolla, R.: Predicting semantic map representations from images using pyramid occupancy networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11138–11147 (2020)
13. 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. International journal of computer vision **115**(3), 211–252 (2015)
14. Saha, A., Mendez, O., Russell, C., Bowden, R.: Enabling spatio-temporal aggregation in birds-eye-view vehicle estimation. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 5133–5139. IEEE (2021)
15. Wagstaff, B., Kelly, J.: Self-supervised scale recovery for monocular depth and egomotion estimation. arXiv preprint arXiv:2009.03787 (2020)
16. Xue, F., Zhuo, G., Huang, Z., Fu, W., Wu, Z., Ang, M.H.: Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2330–2337. IEEE (2020)

17. Yang, W., Li, Q., Liu, W., Yu, Y., Ma, Y., He, S., Pan, J.: Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15536–15545 (2021)
18. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1983–1992 (2018)
19. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1851–1858 (2017)