EMR-MSF: Self-Supervised Recurrent Monocular Scene Flow Exploiting Ego-Motion Rigidity - Supplementary Material -

Zijie Jiang Masatoshi Okutomi
Tokyo Institute of Technology
zjiang@ok.sc.e.titech.ac.jp, mxo@ctrl.titech.ac.jp

1. Additional Implementation Details

Our network is trained using a batch size of 8 on a machine equipped with 4 GTX 3090 GPUs for all experiments. The training takes about two days when the iteration number is equal to 12. Color augmentation, horizontal flipping augmentation, and time-order switching augmentation are applied with a probability of 50% for each during the experiment. For color augmentation, we adopt random gamma adjustments (uniformly sampled from $[0.8, 1.2]$), brightness adjustments (with a multiplication factor uniformly sampled from $[0.5, 2.0]$) and color channel adjustments (with a multiplication factor uniformly sampled from $[0.8, 1.2]$ for each color channel). To ensure stable initialization of the full network during the second stage of training, we disable the use of the non-occlusion mask M_{noc} when calculating the losses L_p and L_g, and remove the mask regularization loss L_m for the first 3k iterations during the second-stage training.

2. Optical Flow Evaluation

Table 1 presents the quantitative comparison of optical flow estimation results of our method with additional self-supervised multi-task methods on the KITTI Scene Flow Training set and Testing set. The training settings of our method are the same as those used in experiments for scene flow evaluation. Our method outperforms all other compared methods on the KITTI Scene Flow Training set. On the KITTI Scene Flow Testing set, our method is slightly surpassed by [7] which requires stereo images during testing, whereas our method only relies on monocular images for testing.

Method	Training set		Testing set
	EPE	F1-all	F1-all
GeoNet [12]	10.81	-	-
DF-Net [13]	8.98	26.01	25.70
Self-Mono-SF [4]	7.51	23.49	23.54
Multi-Mono-SF [5]	-	18.92	19.54
CC-uft [10]	5.66	20.93	25.27
UnOS [11]	5.58	-	18.00
EPC++ [8]	5.43	19.64	20.52
RAFT-MSF [1]	-	17.51	18.37
UnRigidFlow* [7]	5.19	14.68	**11.66**
EffiScene* [6]	4.20	14.31	13.08
Ours	**3.46**	**11.58**	11.93

Table 1: **Quantitative evaluation of the optical flow on the KITTI Scene Flow Training set and Testing set.** The best results are in **bold**. Methods marked with (*) use stereo images for estimation.

3. Visualization of Predictions

In Fig. 1, we provide visualizations of the predictions obtained by our method. We visualize the estimated SE3 motion field $T_{1\to2}$ as the translation field $\tau_{1\to2} \in \mathbb{R}^{H\times W\times 3}$ and rotation field $\phi_{1\to2} \in \mathbb{R}^{H\times W\times 3}$, where $(\tau_{1\to2}, \phi_{1\to2}) = \mathrm{Log}(T_{1\to2})$. We normalize the values in the translation field and rotation field into the range $[0, 1]$ as a color image. We can observe that our method is capable of estimating a constant SE3 motion for pixels in static regions. We attribute this to the effective exploitation of ego-motion rigidity in our method.

4. Failure Cases

Fig. 2 shows some failure cases of our method. Significant estimation errors may still occur in our method for moving objects at the edges of images or textureless regions accompanied by significant motion. Improving the accuracy of estimates in these situations could be a future work for us.

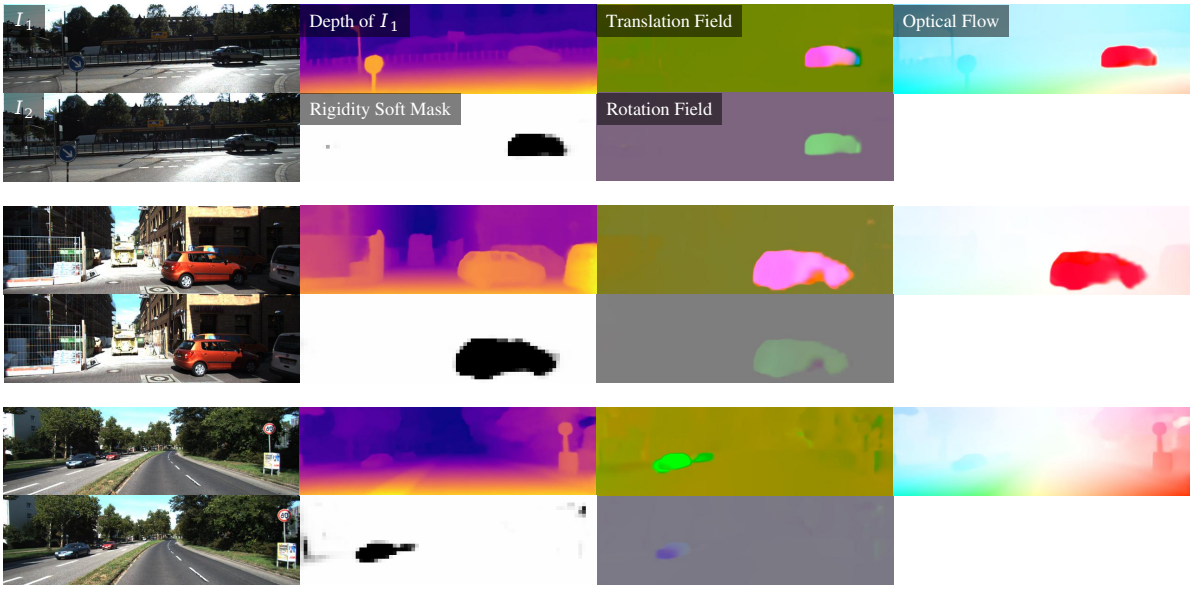

Figure 1: **Visualization of predictions by our method on KITTI Scene Flow Testing set.** We visualize the estimated SE3 motion field as the translation field and the rotation field. Pixels with the same color have the same translation/rotation.

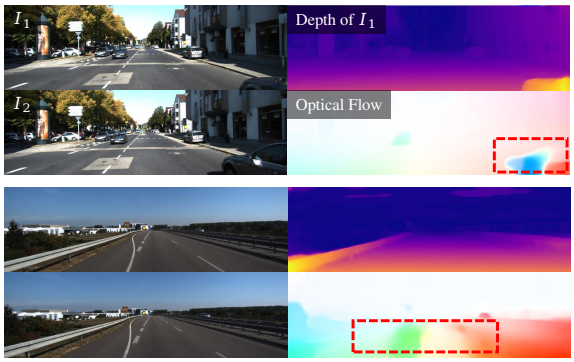

Figure 2: **Failure cases of our method.** The erroneous estiamtions are highlighted in red boxes.

5. Additional Qualitative Comparisons

We provide additional qualitative comparison results of scene flow estimation in Fig. 3 and Fig. 4.

6. Additional Generalization Examples

In Fig. 5, we present additional generalization results of our model originally trained on the KITTI [3] dataset, to the Cityscapes [2] dataset. Moreover, we compare the visual results of our model with those of the model trained on the same data from Self-Mono-SF [4]. Our model exhibits superior generalization capabilities, particularly in static regions such as planar roads and walls.

References

[1] Bayram Bayramli, Junhwa Hur, and Hongtao Lu. Raft-msf: Self-supervised monocular scene flow using recurrent optimizer. *arXiv preprint arXiv:2205.01568*, 2022. 1, 3, 4

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. CVPR*, 2016. 2, 5

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *Intl. J. of Robotics Research*, 32(11):1231–1237, 2013. 2

[4] Junhwa Hur and Stefan Roth. Self-supervised monocular scene flow estimation. In *Proc. CVPR*, 2020. 1, 2, 3, 4, 5

[5] Junhwa Hur and Stefan Roth. Self-supervised multi-frame monocular scene flow. In *Proc. CVPR*, 2021. 1, 3, 4

[6] Yang Jiao, Trac D Tran, and Guangming Shi. Effiscene: Efficient per-pixel rigidity inference for unsupervised joint learning of optical flow, depth, camera pose and motion segmentation. In *Proc. CVPR*, 2021. 1

[7] Liang Liu, Guangyao Zhai, Wenlong Ye, and Yong Liu. Unsupervised learning of scene flow estimation fusing with local rigidity. In *Proc. IJCAI*, 2019. 1

[8] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE PAMI*, 42(10):2624–2641, 2019. 1

[9] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proc. CVPR*, 2015. 3, 4

[10] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive

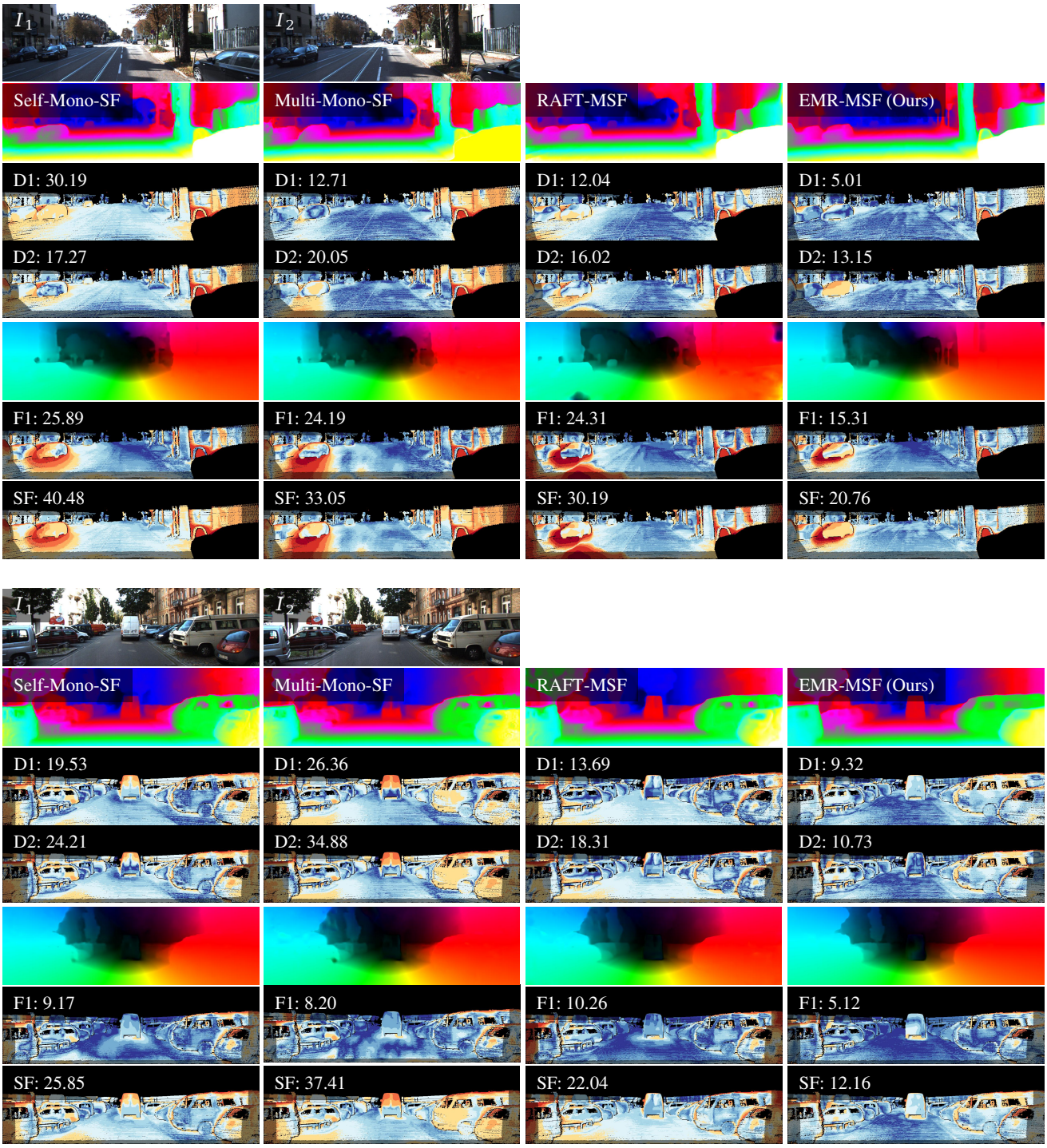

Figure 3: **Qualitative evaluation on KITTI Scene Flow Testing set (1).** We compare our method with Self-Mono-SF [4], Multi-Mono-SF [5] and RAFT-MSF [1] for two scenes using the visualizations provided by the KITTI benchmark [9]. From top to bottom: input images, disparity visualization of I_t, $D1$ error plot, $D2$ error plot, optical flow visualization, corresponding $F1$ error plot and combined SF error plot. The outlier rates are shown on each error plot.

collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proc. CVPR*, 2019. 1

[11] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi

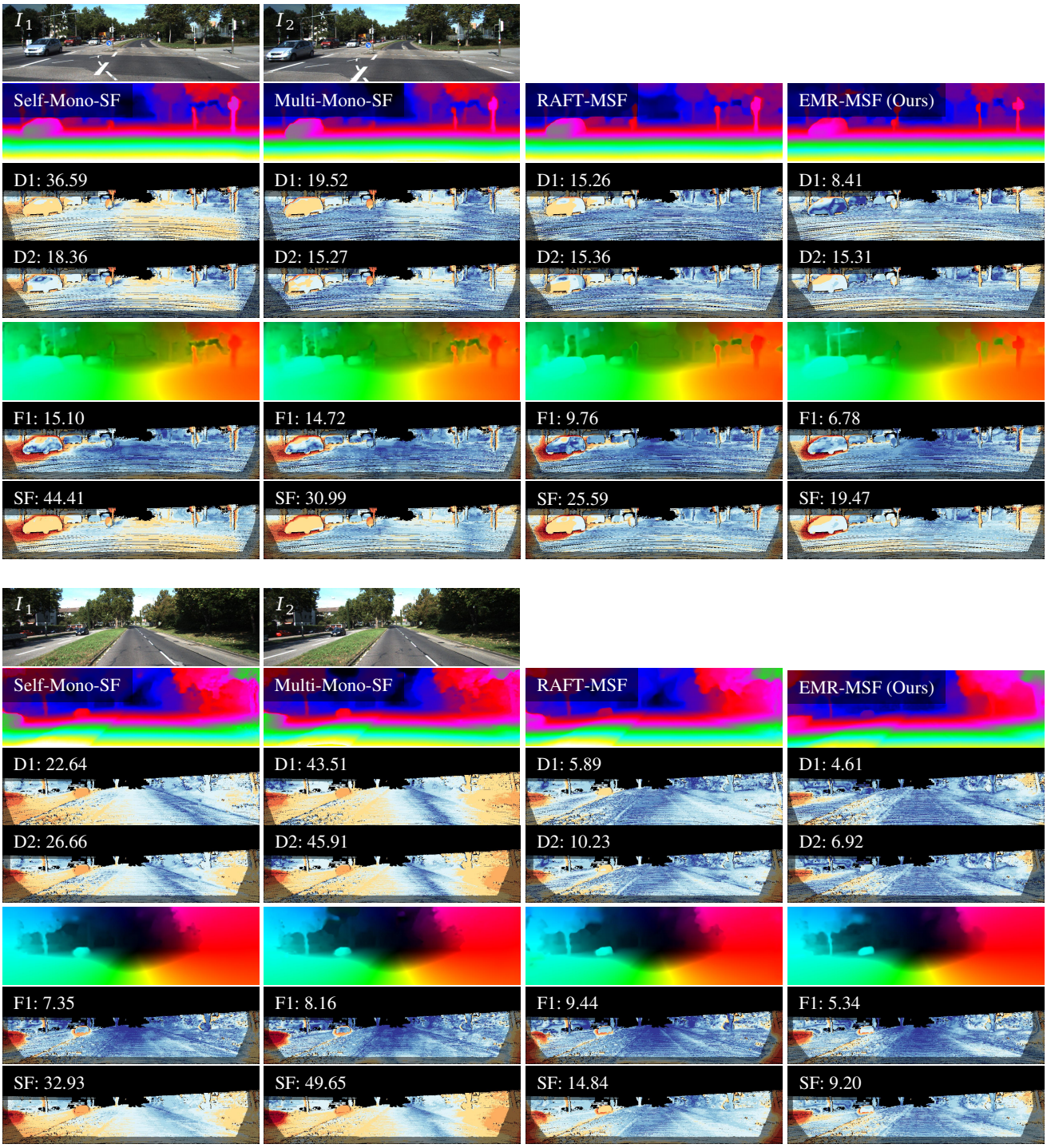

Figure 4: **Qualitative evaluation on KITTI Scene Flow Testing set (2).** We compare our method with Self-Mono-SF [4], Multi-Mono-SF [5] and RAFT-MSF [1] for two scenes using the visualizations provided by the KITTI benchmark [9]. From top to bottom: input images, disparity visualization of I_t, $D1$ error plot, $D2$ error plot, optical flow visualization, corresponding $F1$ error plot and combined SF error plot. The outlier rates on shown on each error plot.

Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proc. CVPR*, 2019. 1

[12] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learn-

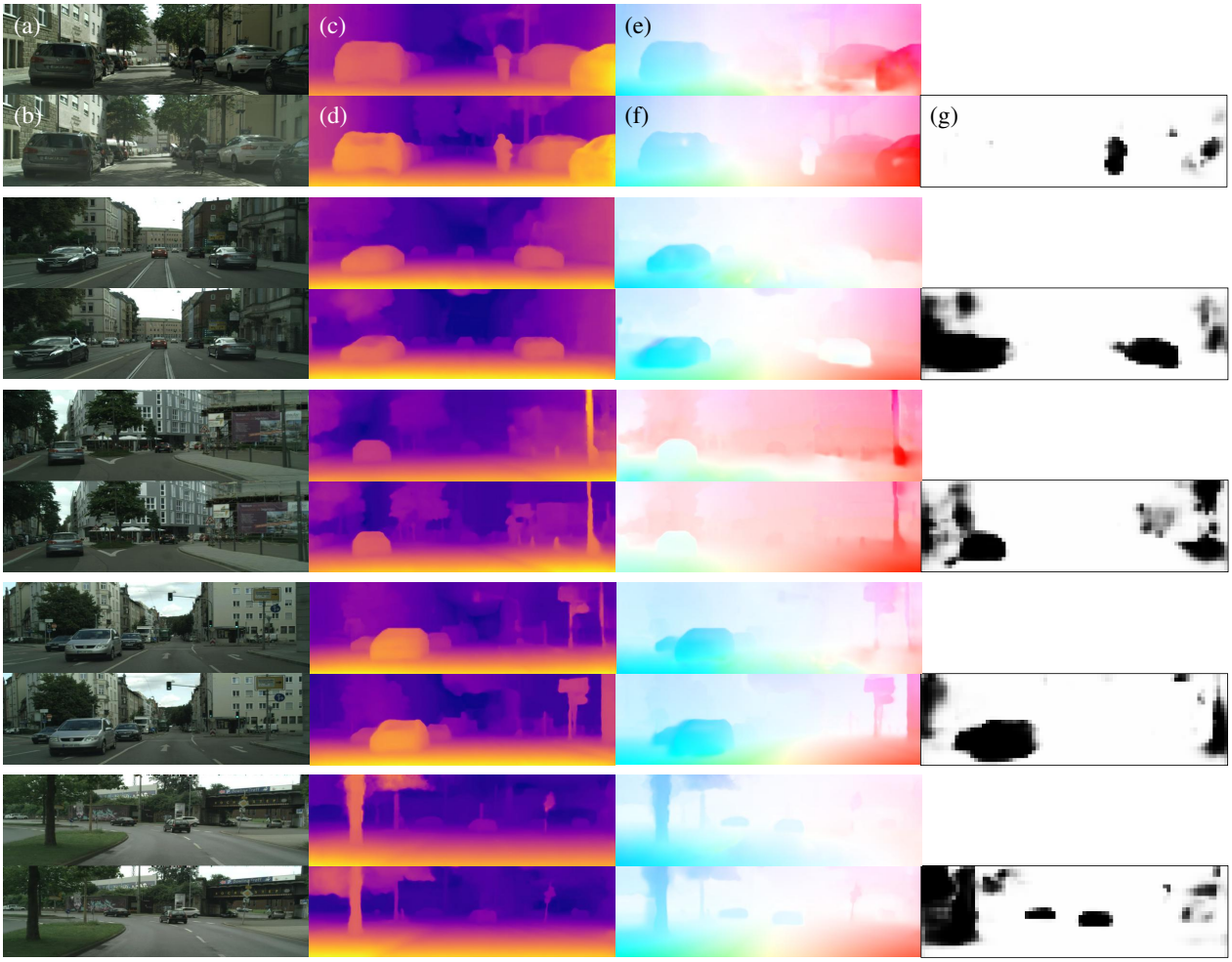

Figure 5: **Comparison of generalization ability between our method and [4] on Cityscapes dataset [2].** (a) input first frame, (b) input second frame, (c) predicted depth of the first frame by [4], (d) predicted depth of the first frame by our method, (e) synthesized optical flow by [4], (f) synthesized optical flow by our method, (g) predicted rigidity soft mask by our method. Our method shows a better generalization ability than [4], especially for the predictions in static regions.

ing of dense depth, optical flow and camera pose. In *Proc. CVPR*, 2018. 1

[13] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proc. ECCV*, 2018. 1