Supplementary Material of
Online Depth Learning against Forgetting in Monocular Videos

1. Reproduced Results and Comparison with Original Paper

To implement our statistic adapter, we need to use batch normalization (BN) layer in the encoder of the network. But in the framework of original papers [3, 1], the depth and pose subnetworks have no BN layers. To further validate the reproduced results of our modified framework, we illustrate the comparisons in table 1. Here we just follow the common training protocol used in [3, 1], where we first train our basic network on Kitti Eigen's training split and then test the model on Kitti Eigen's testing split.

Table 1. Comparison of Reproduced Result with Original Paper

		Lower is better				Higher is better		
Method	Training	Abs Rel	Sq Rel	RMSE	RMSE$_{log}$	< 1.25	$< 1.25^2$	$< 1.25^3$
SfM-Learner [3]								
original paper	Kitti	0.208	1.768	6.856	0.283	0.678	0.885	0.957
ours	Kitti	0.202	1.794	6.624	0.286	0.708	0.892	0.955
original paper + Naive	vKitti	0.230	2.132	7.126	0.298	0.655	0.866	0.943
ours + Naive	vKitti	0.224	0.213	7.118	0.297	0.660	0.871	0.949
SC-SfM-Learner [1]								
original paper	Kitti	0.151	1.154	5.716	0.232	0.798	0.930	0.972
ours	Kitti	0.153	1.155	5.601	0.229	0.798	0.933	0.973
original paper + Naive	vKitti	0.179	1.334	5.906	0.249	0.752	0.909	0.960
ours + Naive	vKitti	0.178	1.326	5.941	0.247	0.747	0.902	0.965

We observe that although we use BN layers in our network, the reproduced results of our basic model well match the original paper. Further, we also compare the online adaptation results with the original network, where we pretrain the models on virtual Kitti dataset and simply use naive method to perform online learning. We observe that with naive online learning method, the original framework obtains similar results as our modified network with BN layers. These results further validate that the improvement achieved by our LPF method is not based on simply adding BN layer, but based on the mechanism of our method.

2. Ablation Study on SC-Sfm-Learner

In Section 5.4 of the original paper, we perform the ablation study on the framework of Sfm-Learner. Here we show the corresponding results on the framework of SC-Sfm-Learner [1]. The results are shown in table 2. We observe that the improvement brought by our method is also consistent. Similar conclusions can be obtained compared to the results of table 1 in original paper.

3. More Detailed Comparison with SOTA Method

In section 5.4 of the original paper, we made analysis on basic frameworks and dataset. Here we show more detailed comparison with the state-of-the-art method [2] (L2A) in the same setting. As illustrated in table 3, we find our LPF method outperforms Naive approach and L2A in the two basic frameworks and datasets.

4. More Qualitative Results

In this section we show more visual results of our method, which can be observed in Fig. 1. The prediction of our method

Table 2. Ablation Study on the Framework of SC-Sfm-Learner for Fast Adaptation

		Lower is better				Higher is better		
Method	Training	Abs Rel	Sq Rel	RMSE	RMSE$_{log}$	< 1.25	$< 1.25^2$	$< 1.25^3$
Basic (no adaptation)	standard	0.246	2.641	7.554	0.328	0.642	0.862	0.941
Basic + Naive	standard	0.178	1.526	5.941	0.2674	0.747	0.902	0.950
Basic + SA	Standard	0.170	1.404	5.906	0.258	0.755	0.910	0.959
Basic + SA + WA	Standard	0.172	1.400	5.883	0.255	0.758	0.914	0.962
Basic + SA + WA + $\mathcal{L}_r$	Standard	0.169	1.383	5.851	0.247	0.756	0.914	0.966
Basic + SA	$\mathcal{L}_{meta}$	0.167	1.376	5.802	0.240	0.760	0.917	0.956
Basic + SA + WA	$\mathcal{L}_{meta}$	0.165	1.324	5.702	0.231	0.768	0.920	0.961
Basic + SA + WA + $\mathcal{L}_r$	$\mathcal{L}_{meta}$	**0.162**	**1.297**	**5.658**	**0.224**	**0.776**	**0.923**	**0.970**

Table 3. Comparison of SOTA Method on Different Frameworks and Datasets

		Lower is better				Higher is better		
Method	Dataset	Abs Rel	Sq Rel	RMSE	RMSE$_{log}$	< 1.25	$< 1.25^2$	$< 1.25^3$
SfM-Learner [3]								
Naive	vKitti	0.224	2.131	7.118	0.299	0.656	0.871	0.949
	Cityscapes	0.202	1.885	6.793	0.281	0.717	0.898	0.953
L2A [2]	vKitti	0.217	1.743	6.802	0.285	0.676	0.876	0.941
	Cityscapes	0.190	1.662	6.325	0.258	0.739	0.903	0.960
LPF	vKitti	0.203	1.608	6.561	0.278	0.694	0.897	0.962
	Cityscapes	0.175	1.492	6.068	0.249	0.750	0.919	0.980
SC-SfM-Learner [1]								
Naive	vKitti	0.178	1.526	5.941	0.267	0.747	0.902	0.950
	Cityscapes	0.168	1.476	5.819	0.251	0.775	0.916	0.963
L2A [2]	vKitti	0.169	0.141	5.672	0.238	0.760	0.914	0.967
	Cityscapes	0.152	0.124	5.453	0.228	0.797	0.921	0.975
LPF	vKitti	0.162	1.297	5.658	0.224	0.776	0.923	0.970
	Cityscapes	0.138	1.059	5.348	0.206	0.819	0.930	0.984

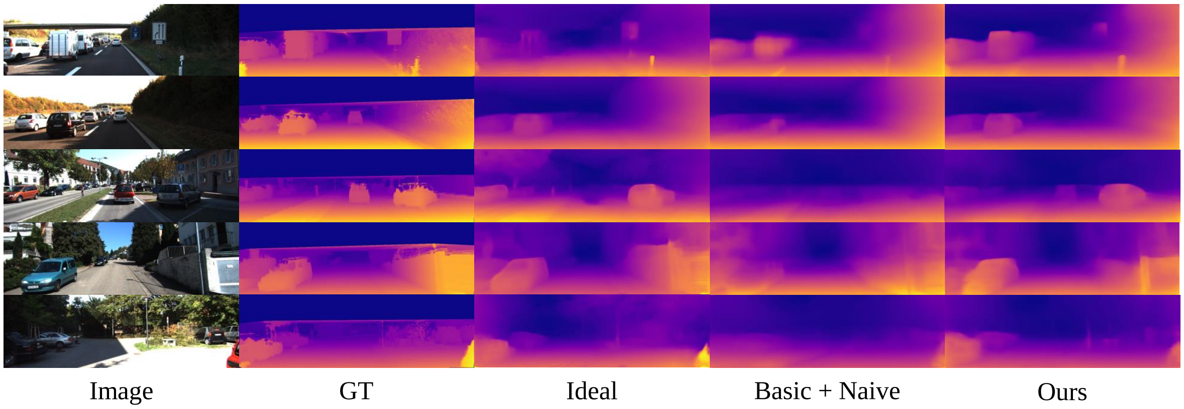

Figure 1. Visual results of ideal or online adaptation approaches. The ideal model is pre-trained on Kitti, while other models are pre-trained on vKitti and directly online adapted to Kitti videos. Results of our method are superior than naive baseline, and close to or even competitive with those of ideal method.

are much better than those of naive online learning, and very close to the ideal offline model. As the ideal model is pre-trained offline on Kitti, it can be seen as a upper bound of online learning approach. Hence, these qualitative results further demonstrate the effectiveness of our method.

References

[1] Jia-Wang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *NeurIPS*, 2019.

[2] Alessio Tonioni, Oscar Rahnama, Thomas Joy, Luigi Di Stefano, Thalaiyasingam Ajanthan, and Philip HS Torr. Learning to adapt for stereo. In *CVPR*, pages 9661–9670, 2019.

[3] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.