SynDeMo: Synergistic Deep Feature Alignment for Joint Learning of Depth and Ego-Motion

A. Supplementary Material

In this supplementary material, we provide additional quantitative and qualitative results. Firstly, we provide additional odometry evaluation of our SynDeMo for the complete KITTI test sequences [6] in Sec. A.1. In Sec. A.2, we explore the effectiveness of our proposed loss function by comparing the full model with our baseline for odometry evaluation. In addition, we provide ablation study to evaluate the effectiveness of our proposed joint training for depth estimation in Sec. A.3. Moreover, we demonstrate generalization abilities of SynDeMo jointly trained on the KITTI and virtual KITTI (vKITTI) [4] for the Cityscapes Mainz sequence [2] in Sec. A.4.

A.1. Additional Odometry Evaluation

The Absolute Trajectory Error (ATE) [7] metric only evaluates pose error for every 5-frame snippets and considers the first frame pose prediction to be ground truth. Therefore, the small quantitative error can add up in the sequence leading to large performance difference. To better compare the performance, we report the average translational error and average rotational error $(^\circ/100m)$ for the complete KITTI test sequences [6] in Table 1. We split the 11 sequences with the ground truth odometry into two parts: the split in which sequences 00-08 are used for training while 09-10 are used for testing.

As shown in Table 1, our SynDeMo shows superior performance with respect to monocular learning method [9], and is comparable to other SLAM methods e.g., ORB-SLAM [7] (with and without loop closure). Our method also outperforms [8], which uses stereo data for training by a large margin.

Method	Seq. 09		Seq. 10	
	t_{rel} (%)	r_{rel} (°)	t_{rel} (%)	r_{rel} (°)
ORB-SLAM [7] (IEEE T-RO 2015)	15.30	0.26	3.68	0.48
ORB-SLAM-LC [7] (IEEE T-RO 2015)	16.23	1.36	-	-
Zhou et al. [9] (CVPR 2017)	17.84	6.78	37.91	17.78
Zhan et al. [8] (CVPR 2018)	11.92	3.60	12.62	3.43
SynDeMo (Real&Synth. | Full)	3.70	1.49	6.05	2.11

Table 1. Odometry evaluation on two test sequences of the KITTI odometry dataset using the metric of average translational and rotational errors. The results of other baselines are taken from [8]. LC denotes loop closure.

A.2. Ablation Study on Geometric Cue from Synthetic Data

Using multi-view self-supervised loss helps us to leverage the geometric cue from synthetic data and to enhance the depth estimation results. To examine this, we measure the Absolute Pose Error (APE) for our full model and the other baseline without using multi-view self-supervised loss term during training. It is interesting to note that our full SynDeMo model trained with the proposed self-supervised loss yields a notable performance gain in odometry estimation for the KITTI sequence 10 (see Fig. 1) compared to other baseline, SynDeMo (Real&Synth. | Feat Align). These quantitative results align well with the qualitative results in Fig. 2.

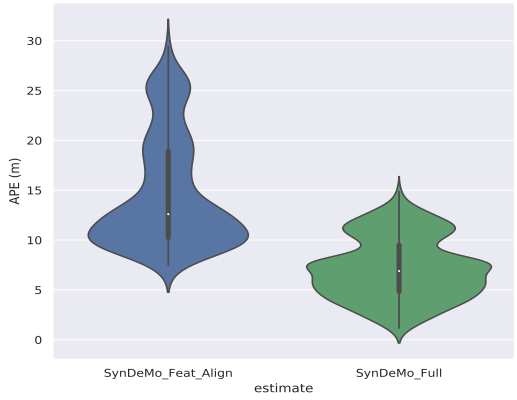

Figure 1. The violin histogram of APE on visual odometry estimation of our SynDeMo (Real&Synth. | Full) compared with SynDeMo (Real&Synth. | Feat Align) for the KITTI testing sequence 10.

A.3. Ablation Study on Joint Training for Depth Estimation

We conduct an ablation study to show how domain adaptation would affect the performance for our depth estimator.

Method	Dataset	Error Metric ↓				Accuracy Metric ↑		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Our Depth Estimator (all-synthetic)	vK	0.265	3.116	6.155	0.314	0.688	0.858	0.937
Our Depth Estimator (real&synthetic)	K+vK	**0.116**	**0.746**	**4.627**	**0.194**	**0.858**	**0.952**	**0.977**

Table 2. Evaluation of depth estimation results for the KITTI test set [3]. For datasets used for training, K is the real KITTI dataset [5] and vK is the virtual KITTI dataset [4]. The best results are shown in **bold**.

Figure 2. Qualitative results on visual odometry of our SynDeMo (Real&Synth. | Full) compared with SynDeMo (Real&Synth. | Feat Align) and the ground truth for the KITTI testing sequence 10.

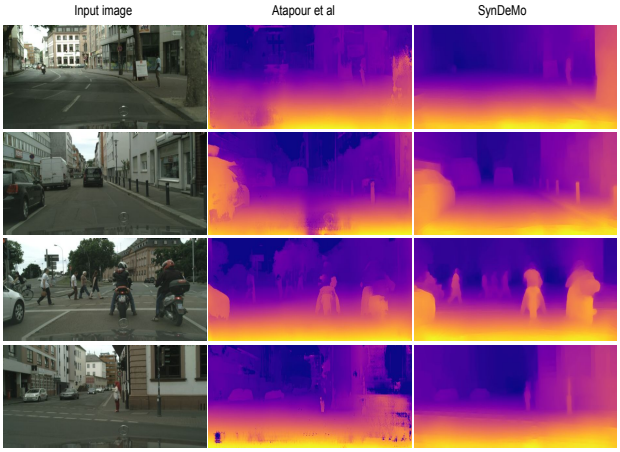

Figure 3. Qualitative results of our depth estimator compared to the most recent cross-domain learning method [1] on the Cityscapes Mainz sequence [2]. Training is done on the KITTI+vKITTI and then we apply our full SynDeMo directly to the Cityscapes without training on the Cityscapes itself. Best viewed in color.

Due to the distribution discrepancy between synthetic images and real monocular images, the learned depth estimator from synthetic images will not achieve the desired performance when the learned model applies to real monocular images. Table 2 demonstrates results of our model on real test images for the KITTI dataset. We can observe that the performance of our depth estimator trained jointly with real and synthetic images has a significant improvement over all metrics compared to our depth estimator trained only on synthetic images.

A.4. Generalization Capabilities

We qualitatively compare our depth estimation results with the most recent cross-domain learning method [1] on the Cityscapes Mainz sequence [2] without training on Cityscapes itself.

As shown in Fig. 3, our SynDeMo is capable to detect more details for objects than [1], with a likely reason being that the scene geometry from the synthetic data is well retained.

References

[1] Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2810, 2018.

[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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[3] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[4] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016.

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

[6] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.

[7] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[8] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 340–349, 2018.

[9] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017.