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 (°/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 [2] (with and without loop closure). Our method also outperforms [8], which uses stereo data for training by a large margin.

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.

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.
<table>
<thead>
<tr>
<th>Method</th>
<th>Dataset</th>
<th>Error Metric ↓</th>
<th>Accuracy Metric ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Depth Estimator (all-synthetic)</td>
<td>vK</td>
<td>Abs Rel = 0.265, Sq Rel = 3.116, RMSE = 6.155, RMSE log = 0.314</td>
<td>δ &lt; 1.25 = 0.688, δ &lt; 1.25^2 = 0.858, δ &lt; 1.25^3 = 0.937</td>
</tr>
<tr>
<td>Our Depth Estimator (real&amp;synthetic)</td>
<td>K+vK</td>
<td>0.116, 0.746, 4.627, 0.194</td>
<td>0.858, 0.952, 0.977</td>
</tr>
</tbody>
</table>

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**.

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


