Towards Better Generalization: Joint Depth-Pose Learning without PoseNet

Wang Zhao Shaohui Liu Yezhi Shu Yong-Jin Liu*
Department of Computer Science and Technology, Tsinghua University, Beijing, China
zhao-w19@mails.tsinghua.edu.cn, b1ueber2y@gmail.com,
shuyz19@mails.tsinghua.edu.cn, liuyongjin@tsinghua.edu.cn

Abstract

In this work, we tackle the essential problem of scale inconsistency for self-supervised joint depth-pose learning. Most existing methods assume that a consistent scale of depth and pose can be learned across all input samples, which makes the learning problem harder, resulting in degraded performance and limited generalization in indoor environments and long-sequence visual odometry application. To address this issue, we propose a novel system that explicitly disentangles scale from the network estimation. Instead of relying on PoseNet architecture, our method recovers relative pose by directly solving fundamental matrix from dense optical flow correspondence and makes use of a two-view triangulation module to recover an up-to-scale 3D structure. Then, we align the scale of the depth prediction with the triangulated point cloud and use the transformed depth map for depth error computation and dense reprojection check. Our whole system can be jointly trained end-to-end. Extensive experiments show that our system not only reaches state-of-the-art performance on KITTI depth and flow estimation, but also significantly improves the generalization ability of existing self-supervised depth-pose learning methods under a variety of challenging scenarios, and achieves state-of-the-art results among self-supervised learning-based methods on KITTI Odometry and NYUv2 dataset. Furthermore, we present some interesting findings on the limitation of PoseNet-based relative pose estimation methods in terms of generalization ability. Code is available at https://github.com/B1ueber2y/TrianFlow.

1. Introduction

Reconstructing the underlying 3D scenes from a collection of video frames or multi-view images has been a long-standing fundamental topic named structure-from-motion (SfM), which serves as an essential module to many real-world applications such as autonomous vehicles, robotics, augmented reality, etc. While traditional methods are built on the golden rule of feature correspondence and multi-view geometry, a recent trend of deep learning based methods [42, 15, 66] try to jointly learn the prediction of monocular depth and ego-motion in a self-supervised manner, aiming to make use of the great learning ability of deep networks to learn geometric priors from large amount of training data.

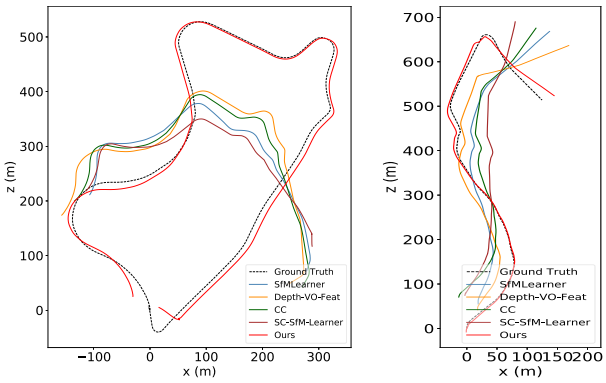

Figure 1. Visual odometry results on sampled sequence 09 and 10 from KITTI Odometry dataset. We sample the original sequences with large stride (stride=3) to simulate fast camera ego-motion that is unseen during training. Surprisingly, all tested PoseNet-based methods get similar failure on trajectory estimation under this challenging scenario. Our system significantly improves the generalization ability and robustness and still works reasonably well on both sequences. See more discussions in Sec 4.4.

The key to those self-supervised learning methods is to build a task consistency for training separated CNN networks, where depth and pose predictions are jointly constrained by depth reprojection and image reconstruction error. While achieving fairly good results, most existing methods assume that a consistent scale of CNN-based monocular depth prediction and relative pose estimation can be learned across all input samples, since relative pose estimation inherently has scale ambiguity. Although several recent proposals manage to mitigate this scale problem [2, 12], this strong hypothesis still makes the learning problem difficult and leads to severely degraded performance, especially in long-sequence visual odometry applications and indoor environments, where the changes of relative pose across sequences are significantly remarkable.

*Corresponding author.

Motivated by those observations, we propose a new self-supervised depth-pose learning system which explicitly disentangles scale from the joint estimation of the depth and relative pose. Instead of using a CNN-based camera pose prediction module (e.g. PoseNet), we directly solve the fundamental matrix from optical flow correspondences and implement a differentiable two-view triangulation module to locally recover an up-to-scale 3D structure. This triangulated point cloud is later used to align the predicted depth map via a scale transformation for depth error computation and reprojection consistency check.

Our system essentially resolves the scale inconsistency problem in design. With two-view triangulation and explicit scale-aware depth adaptation, the scale of the predicted depth always matches that of the estimated pose, enabling us to remove the scale ambiguity for joint depth-pose learning. Likewise, we borrow the advantage of traditional two-view geometry to acquire more direct, accurate and robust depth supervision in a self-supervised end-to-end manner, where the depth and flow prediction can benefit from each other. Moreover, because our relative pose is directly solved from the optical flow, we simplify the learning process and do not require the knowledge of correspondence to be learned from the PoseNet architecture, enabling our system to have better generalization ability in challenging scenarios. See an example in Figure 1.

Experiments show that our unified system significantly improves the robustness of self-supervised learning methods in challenging scenarios such as long video sequences, unseen camera ego-motions, and indoor environments. Specifically, our proposed method achieves significant performance gain on NYU v2 dataset and KITTI Odometry over existing self-supervised learning-based methods, and maintains state-of-the-art performance on KITTI depth and flow estimation. We further test our framework on TUM-RGBD dataset and again demonstrate its much promising generalization ability compared to baselines.

2. Related Work

Monocular Depth Estimation. Recovering 3D depth from a single monocular image is a fundamental problem in computer vision. Early methods [45, 46] use feature vectors along with a probabilistic model to provide monocular clues. Later, with the advent of deep networks, a variety of systems [8, 10, 42] are proposed to learn monocular depth estimation from groundtruth depth maps in a supervised manner. To resolve the data deficiency problem, [36] uses synthetic data to help the disparity training, and several works [30, 26, 29, 27] leverage standard structure-from-motion (SfM) pipeline [47, 48] to generate a psuedo-groundtruth depth map by reprojecting the reconstructed 3D structure. Recently, a bunch of works [11, 15, 66] on self-supervised learning are proposed to jointly estimate other geometric entities that help depth estimation learning via photometric reprojection error. However, although some recent works [54, 12] try to address the scale ambiguity for monocular depth estimation with either normalization or affine adaptation, self-supervised methods still suffer from the problem of scale inconsistency when applied to challenging scenarios. Our work combines the advantages of SfM-based unsupervised methods and self-supervised learning methods, essentially disentangles scale from our learning process and benefits from the more accurate and robust triangulated structure with two-view geometry.

Self-Supervised Depth-Pose Learning. Struction-from-motion (SfM) is a golden standard for depth reconstruction and camera trajectory recovery from videos and image collections. Recently many works [53, 3, 60, 52] try to combine neural networks into SfM pipeline to make use of the learned geometric priors from training data. Building on several unsupervised methods [11, 15], Zhou *et al*. [66] first proposes a joint unsupervised learning framework of depth and camera ego-motion from monocular videos. The core idea is to use photometric error as supervision signal to jointly train depth and ego-motion networks. Along this line, several methods [61, 67, 35, 2, 41, 34, 5, 6] further improve the performance by incorporating better training strategies and additional constraints including ICP regularization [35], collaborative competition [41], dense online bundle adjustment [5, 6], etc. Most related to us, Bian *et al*. [2] introduce geometry consistency loss to enforce the scale-consistent depth learning. Different from them, our method essentially avoids the scale inconsistency in deign by directly solving relative pose from optical flow correspondence. Our system designs and findings are orthogonal to existing depth-pose learning works, significantly improving those methods on both accuracy and generalization.

Two-view Geometry. Establishing pixel-wise correspondences between two images is a long-standing visual problem. Traditional methods utilize hand-crafted descriptors [32, 1, 43] to build rough correspondence for the subsequent fundamental matrix estimation. Recently, building on classic works of optical flow [21, 33], researchers [7, 22, 51] find deep neural networks powerful on feature extraction and dense correspondence estimation between adjacent frames. Likewise, several self-supervised methods [23, 37, 31] are proposed to supervise optical flow training with photometric consistency.

Another line of research is to combine learning-based methods with the fundamental matrix estimation after establishing the correspondence. While some researches [4, 40] focus on making RANSAC [9] differentiable, another alternative is to use an end-to-end pose estimation network [24]. However, some recent findings [44, 65] on image-based localization show that PoseNet design [24] can

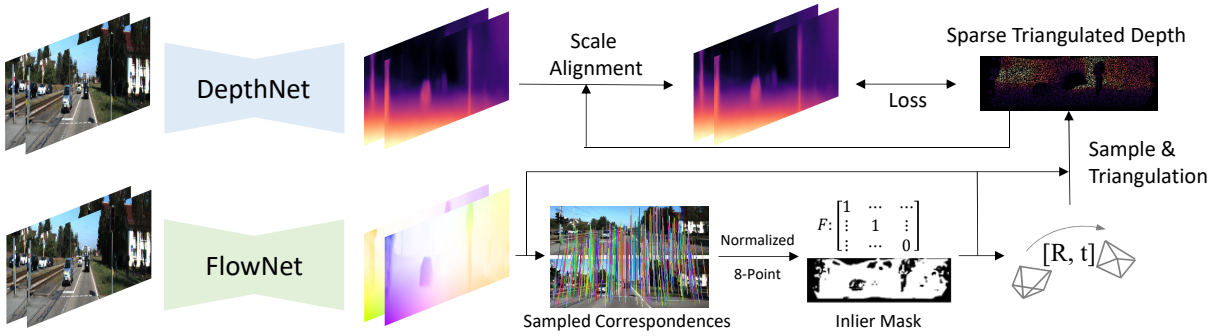

Figure 2. System overview. DepthNet takes each input image and predicts monocular depths respectively. FlowNet take image pairs as input and predict optical flows. The relative pose is recovered by sampling correspondences, solving the fundamental matrix, and cheirality condition check. Accurate pixel matches are re-sampled and used for triangulation. Depth predictions are aligned according to sparse triangulation depth, and then losses are measured respectively, to supervise DepthNet and FlowNet jointly.

degrade the generalization ability compared to geometry-based methods. Also, the inherent problem of scale ambiguity for pose estimation makes it hard to decouple with depth scale during joint training. In our work, we show that by building on conventional two-view geometry, our optical flow estimation module is able to accurately recover relative poses and can benefit from the joint depth-pose learning.

3. Method

3.1. Motivation and System Overview

The central idea of existing self-supervised depth-pose learning methods is to learn two separated networks on the estimation of monocular depth and relative pose by enforcing geometric constraints on image pairs. Specifically, the predicted depth is reprojected onto another image plane using the predicted relative camera pose and then photometric error is measured. However, this class of methods assume a consistent scale of depth and pose across all images, which could make the learning problem difficult and lead to a scale drift when applied to visual odometry applications.

Some recent proposals [54, 2] introduce additional consistency constraints to mitigate this scale problem. Nonetheless, the scale-inconsistent issue naturally exists because the scales of the estimated depth and pose from neural networks are hard to measure. Also, the photometric error on the image plane supervises the depth in an implicit manner, which could suffer from data noise when large textureless regions exist. Furthermore, similar to two recent findings [44, 65] that CNN-based absolute pose estimation is difficult to generalize beyond image retrieval, the performance of the CNN-based ego-motion estimation also significantly degrades when applied to challenging scenarios.

To address the above challenges, we propose a novel system that explicitly disentangles scale consistency at both training and inference. The overall pipeline of our method is shown in Figure 2. Instead of relying on CNN-based relative pose estimation, we first predict optical flow and solve the fundamental matrix from the dense flow correspondence, thereby recovering relative camera pose. Then, we sample over the inlier regions and use a differentiable triangulation module to reconstruct an up-to-scale 3D structure. Finally, depth error is directly computed after a scale adaptation from the predicted depth to the triangulated structure and reprojection error on depth and flow is measured to further enforce end-to-end joint training. Our training objective L is formulated as follows:

$$L = w_1 L_f + w_2 L_d + w_3 L_p + w_4 L_s. \tag{1}$$

The L_f denotes the unsupervised loss on optical flow, where we follow the photometric error design (pixel + SSIM [56] + smooth) on PWC-Net [51]. Occlusion mask M_o is derived from optical flow by following [55]. We also add a forward-backforward consistency [61] to generate a score map M_s for subsequent fundamental matrix estimation. L_d is the loss between triangulated depth and predicted depth. L_p is the reprojection error for image pairs, which consists of two parts, depth map reconstruction error and flow error between optical flow and rigid flow generated by depth reprojection. L_s is the depth smoothness loss, which follows the same formulation in [2].

In the following parts, we first describe how we recover relative pose via fundamental matrix from optical flow. Then, we show how to use the recovered pose to build up self supervision geometrically without scale ambiguity. Finally, a brief description is given on the inference pipeline of our system when applied to visual odometry applications.

3.2. Fundamental Matrix from Correspondence

We recover camera pose from optical flow correspondence via traditional fundamental matrix computation algorithm. Optical flow offers correspondence for every pixel,

while some of them are noisy and thus not suitable for solving the fundamental matrix. We first select reliable correspondences using the occlusion mask M_o and forward-backward flow consistency score map M_s, which are both generated from our flow network. Specifically, we sample the correspondences that locate in non-occluded regions and have top 20% forward-backward scores. Then we randomly acquire 6k samples out of the selected correspondences and solve the fundamental matrix F via the simple normalized 8-point algorithm [18] in RANSAC [9] loop. Fundamental matrix is then decomposed into camera relative pose, which is denoted as $[R, t]$. Note that there are 4 possible solutions for $[R, t]$ and we adopt cheirality condition check, meaning that the triangulated 3D points must be in front of both cameras, to find the best one solution. In this way, our predicted camera pose fully depends on the optical flow network, which can better generalize across image sequences and under challenging scenarios.

3.3. Two-view Triangulation as Depth Supervision

Recovering the relative camera pose with fundamental matrix estimation from optical flow formulates an easier learning problem and improves the generalization, but cannot enforce scale-consistent prediction on its own. To follow up with this design, we propose to explicitly align the scale of depth and pose. Intuitively two reasonable solutions on scale optimization exist: 1) aligning depth with pose 2) aligning pose with depth. We adopt the former one as it can be formulated as a linear problem using two-view triangulation [19].

Again, instead of using all pixel matches to perform dense triangulation, we first select top accurate correspondences. Specifically, we generate an inlier score map M_r by computing the distance map D_{epi} from each pixel to its corresponding epipolar line, which is helpful for masking out bad matches and non-rigid regions, such as moving objects. Then this inlier score map M_r is combined with occlusion mask M_o, optical flow forward-backward score M_s, to sample rigid, non-occluded and accurate correspondences. Here we also randomly acquire 6k samples out of the top 20% correspondence and perform two-view triangulation to reconstruct an up-to-scale 3D structure. We adopt the midpoint triangulation as it has a linear and robust solution. Its formulation is as follows:

$$x^* = \underset{x}{\operatorname{argmin}} \, [d(L_1, x)]^2 + [d(L_2, x)]^2, \tag{2}$$

where L_1 and L_2 denote two camera rays generated from optical flow correspondence. This problem can be directly solved analytically and the solver is naturally differentiable, enabling our system to perform end-to-end joint training. The derivation of its analytical solution is included in supplementary materials. We use the triangulated 3D structure as the depth supervision. To mitigate the numerical issue, such as triangulation of matches around epipoles, we filter the correspondence online with respect to the angle of the camera rays. Also, we filter the triangulated samples with negative or out-of-bound depth reprojection. Figure 3 visualizes samples for the depth reprojection of the dense triangulated structure. The quality of the depth is much promising and feasible to be used as a psuedo depth groundtruth signal to guide the network learning. This design shares similar spirits with many recent methods [30, 26, 29, 27] on supervising the monocular depth estimation with offline SfM inference where they also use the reconstructed structure as the psuedo groundtruth. Compared to those works, our online robust triangulation module explicitly handles occlusion, moving objects and bad matches, and is successfully integrated into the joint training system where correspondence generation and depth prediction could benefit together.

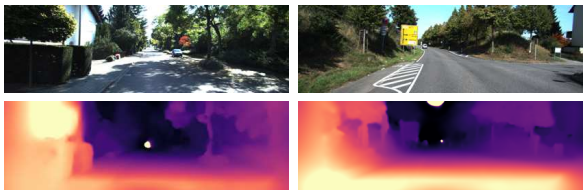

Figure 3. Dense triangulation examples. While most of the triangulated matches are pretty good, the depth values around occluded areas and epipole regions are noisy. In these two examples, epipoles locate near the image center and the nearby triangulated depths are negative or very close to zero. Thus we only use sampled sparse accurate triangulation depth as supervision.

3.4. Scale-invariant Design

As aforementioned, we can resolve the scale-inconsistent problem by aligning predicted depth with the triangulated structure. Specifically, we align the monocular depth estimation D with a single scale transformation s to minimize the error between the transformed depth $D_t = sD$ and the psuedo groundtruth depth D_{tri} from triangulation in Eq. (3). Then, the minimized error is used as the depth loss for back-propagation. This online fitting technique was also introduced in a recent work [12].

$$L_d = (\frac{D_{tri} - D_t}{D_{tri}})^2 \tag{3}$$

The transformed depth is explicitly aligned to the triangulated 3D structure, whose scale is decided by relative pose scale, thus scale inconsistency is essentially disentangled from the system. Also, the transformed depth can be further used for computing the dense reprojection error L_p. This error is formulated in Eq. (4):

$$L_p = w_{31} L_{pf} + w_{32} L_{pd}, \tag{4}$$

Given an image pair (I_a, I_b), scale-transformed depth estimations (D_a, D_b), camera intrinsic parameter K, and recovered relative pose $\mathrm{T_{ab}}$ from optical flow F_{ab}, loss L_{pf} is calculated as follows, which measures the 2D error between optical flow and rigid flow generated by depth reprojection.

$$\begin{aligned} p_{bd} &= \phi(\mathrm{K}[\mathrm{T}_{ab} D_a(p_a) \mathrm{K}^{-1}(h(p_a)])) \\ p_{bf} &= p_a + F_{ab}(p_a) \\ L_{pf} &= \frac{1}{|M_r|} \sum_{p_a} M_r(p_a) |p_{bd} - p_{bf}| + |D_{epi}| \end{aligned} \tag{5}$$

where p_a is the pixel coordinate (x, y) in I_a, and $h(p_a)$ indicates the homogeneous coordinates of p_a. Operator $\phi([x, y, z]) = [x/z, y/z]$ gives pixel coordinates. As mentioned in Sec 3.3, D_{epi} is the distance map of each pixel to its corresponding epipolar line and M_r is the inlier score map. $|D_e|$ serves as a geometric regularization term to help improve the correspondences. $|M_r| = \sum_{p_a} M_r(p_a)$ is for normalization. Depth reprojection error L_{pd} is defined as:

$$L_{pd} = \frac{1}{|M_o M_r|} \sum_{p_a} M_o(p_a) M_r(p_a) |1 - \frac{D_b^a(p_{bd})}{D_b^s(p_{bd})}| \tag{6}$$

where D_b^a is the reprojected depth map by D_a and $\mathrm{T_{ab}}$. D_b^s is the interpolated depth map of D_b to align with reprojected pixel coordinates p_{bd}, which is defined in Eq. (5). M_o is the occlusion mask from optical flow.

3.5. Inference Pipeline on Video Sequences

At inference step, we use the same strategy for relative pose estimation via fundamental matrix estimation from optical flow correspondence. Then, the scale of the triangulated structure is aligned as the same with that of monocular depth estimation. When the optical flow magnitude is too small, we use perspective-n-point (PnP) method over the predicted depth directly. In this way, we essentially avoid the scale inconsistency between depth and pose during inference. A recent paper [63] employs similar visual odometry inference strategies to utilize neural network predictions. However, their depth and flow network are pre-trained separately using PoseNet architecture, while our method builds a robust joint learning system to learn better depth, pose and flow predictions in a self-supervised manner.

4. Experiments

4.1. Implementation Details

Dataset. We first validate our design on KITTI dataset [13], then conduct extensive experiments on KITTI Odometry, NYUv2 [49] and TUM-RGBD [50] datasets to demonstrate the robustness and generalization ability of our proposed system. For original KITTI dataset, we use Eigen *et al.*'s split [8] of the raw dataset for training, which is consistent with related works [66, 41, 6, 14]. The images are resized to 832×256. We evaluate the depth network on the Eigen *et al.*'s testing split, and the optical flow network on KITTI 2015 training set. For KITTI Odometry dataset, we follow the standard setting [6, 66, 61] of using sequences 00-08 for training and 09-10 for testing. Since the camera ego-motions in KITTI odometry dataset are relatively regular and steady, we sample the original test sequences to shorter versions, mimicking fast camera motions, for testing the generalization ability of networks on unseen data. NYUv2 [49] and TUM-RGBD [50] are two challenging indoor datasets which consist of large textureless surfaces and more complex camera ego-motions.

Network Architectures. Since our work focuses on an improved self-supervised depth-pose learning scheme, we adopt similar network designs that align with existing self-supervised learning methods. For the depth network, we use the same architecture as [14] which adopts ResNet18 [20] as encoder and DispNet [15] as decoder. The optical flow network is based on PWCNet [51] and handles occlusion using the method described in [55]. Camera pose is calculated from filtered optical flow correspondences in a non-parametric manner.

Training. Our system is implemented in PyTorch [39]. We use Adam [25] optimizer and set learning rate to 10^{-4} and batch size to 8. The whole training schedule consists of three stages. Firstly, we only train optical flow network in an unsupervised manner via image reconstruction loss. After 20 epochs, we freeze optical flow network and train the depth network for another 20 epochs. Finally, we jointly train both networks for 10 epochs.

4.2. Conventional KITTI Setting

Monocular Depth Estimation. We report results on monocular depth estimation on Eigen *et al.*'s testing split on KITTI [13] dataset. The results are summarized in Table 1. Our method achieves comparable or better performance with state-of-the-art methods [14, 16]. The performance gain is benefited from our system design, where the scale is disentangled from training and robust supervision is acquired from two-view triangulation module. We further explore the effects of different loss terms. The performance slightly drops without reprojection loss L_p as shown in Table 1, and the training cannot converge without triangulation supervision loss L_d. Figure 4 shows qualitative results of our depth prediction. Note that our method is orthogonal to many previous works, and could be potentially incorporated with many advanced techniques such as online refinement [5, 6], and more effective architecture [17].

Optical Flow Estimation. Table 2 summarizes the results

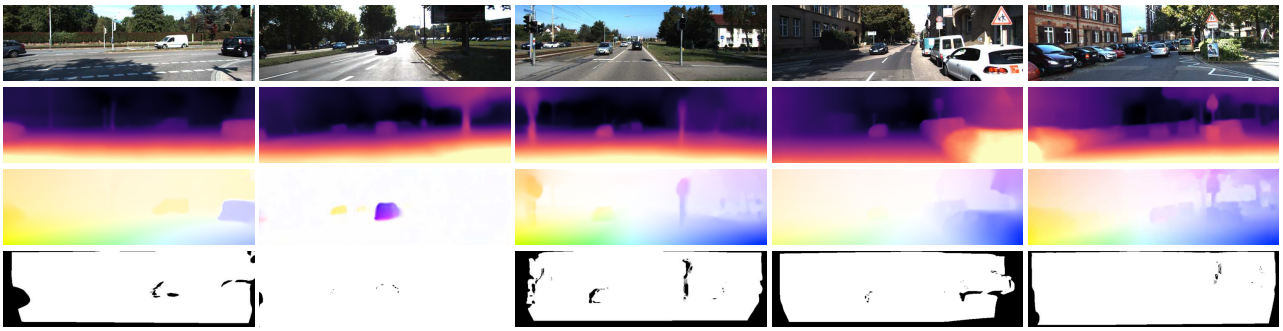

Figure 4. Qualitative results on KITTI dataset. **Top to bottom**: Original image, depth prediction, optical flow prediction and occlusion mask prediction.

Method	Error				Accuracy, δ		
	AbsRel	SqRel	RMS	RMSlog	<1.25	$<1.25^2$	$<1.25^3$
Zhou *et al*. [66]	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Mahjourian *et al*. [35]	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Geonet [61]	0.155	1.296	5.857	0.233	0.793	0.931	0.973
DDVO [54]	0.151	1.257	5.583	0.228	0.810	0.936	0.974
DF-Net [67]	0.150	1.124	5.507	0.223	0.806	0.933	0.973
CC [41]	0.140	1.070	5.326	0.217	0.826	0.941	0.975
EPC++ [34]	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2depth (-ref.) [5]	0.141	1.026	5.291	0.215	0.816	0.945	0.979
GLNet (-ref.) [6]	0.135	1.070	5.230	0.210	0.841	0.948	0.980
SC-SfMLearner [2]	0.137	1.089	5.439	0.217	0.830	0.942	0.975
Gordon *et al*. [16]	0.128	0.959	5.230	0.212	0.845	0.947	0.976
Monodepth2 (w/o pretrain) [14]	0.132	1.044	5.142	0.210	0.845	0.948	0.977
Monodepth2† [14]	0.115	0.882	4.701	0.190	**0.879**	**0.961**	0.982
Ours (w/o pretrain and L_p)	0.135	0.932	5.128	0.208	0.830	0.943	0.978
Ours (w/o pretrain)	0.130	0.893	5.062	0.205	0.832	0.949	0.981
Ours†	**0.113**	**0.704**	**4.581**	**0.184**	0.871	**0.961**	**0.984**

Table 1. Quantitative comparison between our proposed system and state-of-the-art depth-pose learning methods (without post-processing) for monocular depth Estimation on KITTI [13] dataset. † indicates ImageNet pretraining.

of optical flow estimation on KITTI 2015 training set. We also report the performance of only training our optical flow network, denoted as FlowNet-only. Results show that the optical flow module can benefit from joint depth-pose learning process and therefore outperforms most previous unsupervised flow estimation methods and joint learning methods. Figure 4 shows some qualitative results.

4.3. Generalization on Long Sequences

We further extend our system for visual odometry applications. Most of current depth-pose learning methods suffer from error drift when applied on long sequences since the pose network is trained to predict relative pose in short snippets. Recently, Bian *et al*. [2] propose a geometric consistency loss to enforce the long-term consistency of pose prediction and show better results. We test our system with their method and other state-of-the-art depth-pose learning methods on KITTI Odometry datatset. Since monocular systems lack real world scale factor, we align all the predicted trajectory to groundtruth by applying 7DoF (scale + 6DoF) transformation. Table 3 shows the results. Because our method essentially mitigates the scale drift of existing

Method	Noc	All	Fl
FlowNetS [22]	8.12	14.19	-
FlowNet2 [51]	4.93	10.06	30.37%
UnFlow [37]	-	8.10	23.27%
Back2Future [23]	-	7.04	24.21%
Geonet [61]	8.05	10.81	-
DF-Net [67]	-	8.98	26.01%
EPC++ [34]	-	5.84	-
CC [41]	-	**5.66**	20.93%
GLNet [6]	4.86	8.35	-
Ours (FlowNet-only)	4.96	8.97	25.84%
Ours	**3.60**	5.72	**18.05%**

Table 2. Optical flow estimation results. We report the average end-point-error (EPE) on non-occluded regions and overall regions, and Fl score on KITTI 2015 training set, following [61, 6]. Top 2 rows: supervised methods which are trained on synthetic data only. Middle 2 rows: unsupervised optical flow learning methods. Bottom rows: joint depth-pose learning methods.

Methods	Seq. 09		Seq. 10	
	t_{err} (%)	r_{err} (°/100m)	t_{err} (%)	r_{err} (°/100m)
ORB-SLAM2† [38]	9.31	0.26	2.66	0.39
ORB-SLAM2 [38]	**2.84**	**0.25**	**2.67**	**0.38**
Zhou *et al.* [66]	11.34	4.08	15.26	4.08
Deep-VO-Feat [62]	9.07	3.80	9.60	3.41
CC [41]	7.71	2.32	9.87	4.47
SC-SfMLearner [2]	7.60	2.19	10.77	4.63
Ours	**6.93**	**0.44**	**4.66**	**0.62**

Table 3. Visual odometry results on KITTI Odometry dataset. The average translation and rotation errors are reported. ORB-SLAM2† indicates that the loop closure is disabled.

depth-pose learning methods with scale inconsistency, we achieve significant performance improvement over state-of-the-art depth-pose learning systems. Although our dense correspondence is learned in an unsupervised manner and no local BA and mapping are used at inference, we achieve comparable results with conventional SLAM systems [38]. Figure 5 shows the recovered trajectories on two tested sequences respectively.

4.4. Generalization on Unseen Ego-motions

To verify the robustness of our method, we design an experiment to test visual odometry application with unseen camera ego-motions. Original sequences in KITTI Odometry dataset are recorded by driving cars with relatively steady velocity, therefore there are nearly no abrupt motions. Meanwhile, the data distributions of relative poses on testing sequences are quite similar to those on training set. We sample the sequences 09 and 10 with different strides to mimic the velocity changes of cameras, and directly test our methods and other depth-pose learning methods, which are all trained on original KITTI Odometry training split, and tested on these new sequences. Table 4 shows the results on sequences 09 and 10 which are sampled with stride 3. It is clearly shown that our method is robust and generalize much better on this unseen data distribution, even compared to ORB-SLAM2 [38], which frequently fails and re-initializes under fast motion. More surprisingly, as shown in Figure 1, all existing depth-pose learning methods relying on PoseNet fail to predict reasonable and consistent poses, and produce relatively similar trajectories, which drift far away from the groundtruth trajectory. This might be due to the fact that CNN-based pose estimation acts more like a retrieval method and cannot generalize to unseen data. This interesting finding shares similar spirits with recent works [44, 65], where the generalization ability of CNN-based absolute pose estimation is studied in depth. With our scale-agnostic system design and the use of conventional two-view geometry, we achieve significantly more robust performance on videos with unseen per-frame ego-motions.

4.5. Generalization on Indoor Datasets

To further test our generalization ability, we evaluate our method on two indoor datasets: NYUv2 [49] and TUM-

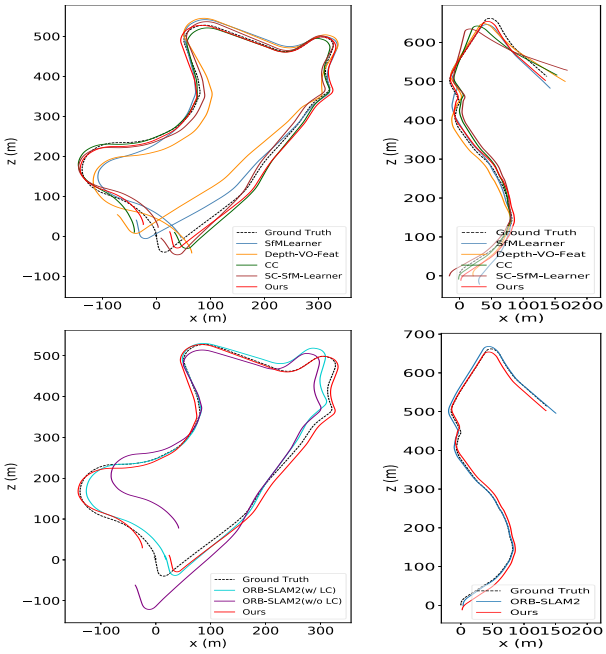

Figure 5. Visual odometry results on sequence 09 and 10.

Methods	Seq. 09		Seq. 10	
	t_{err} (%)	r_{err} (°/100m)	t_{err} (%)	r_{err} (°/100m)
ORB-SLAM2 [38]	X	X	X	X
Zhou *et al.* [66]	49.62	13.69	33.55	16.21
Deep-VO-Feat [62]	41.24	10.80	24.17	11.31
CC [41]	41.99	11.47	30.08	14.68
SC-SfMLearner [2]	52.05	14.39	37.22	18.91
Ours	**7.21**	**0.56**	**11.43**	**2.57**

Table 4. Visual odometry results on KITTI Odometry dataset with large sample stride (stride=3). While ORB-SLAM2 is hard to initialize and keeps losing tracking in this case, our method can produce fairly good prediction. See Figure 1 for plotted trajectories.

RGBD [50] benchmark. Indoor environments are challenging due to the existence of large texture-less regions and much more complex ego-motion (compared to relatively consistent ego-motion on KITTI [13]), making the training of most existing self-supervised depth-pose learning method collapse, as shown in Figure 7. We train our network on NYUv2 raw training set and evaluate the depth prediction on labeled test set. Training images are resized to 192×256 by default. Quantitative results are shown in Table 5. Our method achieves state-of-the-art performance among unsupervised learning baselines. To further study the effects on our system design, we introduce two baseline methods in Table 5: *PoseNet* baseline is built by substituting our optical flow and two-view triangulation module with a PoseNet-like architecture, where relative pose is directly predicted with a convolutional neural network, and *PoseNet-Flow* baseline uses optical flow as input for PoseNet branch to predict relative pose. See supplementary material for more details about these two baselines. Our proposed system achieves a large performance gain, indicating the effectiveness and robustness of our system design.

Method	Error			Accuracy, δ		
	rel	log10	rms	<1.25	$<1.25^2$	$<1.25^3$
Make3D [46]	0.349	-	1.214	0.447	0.745	0.897
Li *et al*. [28]	0.232	0.094	0.821	0.621	0.886	0.968
MS-CRF [58]	0.121	0.052	0.586	0.811	0.954	0.987
DORN [10]	0.115	0.051	0.509	0.828	0.965	0.992
Zhou *et al*. [64]	0.208	0.086	0.712	0.674	0.900	0.968
PoseNet	0.283	0.122	0.867	0.567	0.818	0.912
PoseNet-Flow	0.221	0.091	0.764	0.659	0.883	0.959
Ours	0.201	0.085	0.708	0.687	0.903	0.968
Ours (448×576)	**0.189**	**0.079**	**0.686**	**0.701**	**0.912**	**0.978**

Table 5. Results on NYUv2 depth estimation. Supervised methods are shown in the first rows. *PoseNet* indicates replacing flow and triangulation module with PoseNet in our system. *PoseNet-Flow* indicates using optical flow as input for PoseNet.

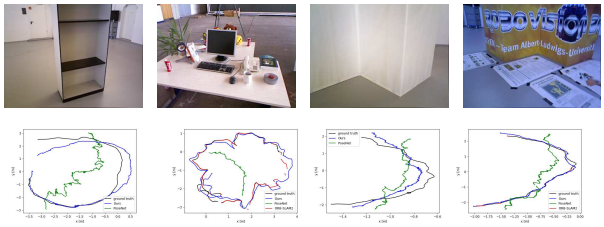

Figure 6. Visual odometry results on TUM RGBD dataset. Our proposed system can still work well with large textureless regions (the 1st and the 3rd cases), complex camera motions (the 2nd case) and different lighting conditions (the 4th case), demonstrating improved robustness compared to the baseline. **Better viewed when zoomed in.**

In addition, we test our method on TUM-RGBD [50] dataset, which is widely used for evaluating visual odometry and SLAM systems [38, 57]. This dataset is collected mainly by hand-held cameras in indoor environments, and consists of various challenging conditions such as extreme textureless regions, moving objects, and abrupt motions, etc. We follow the same train/test setting as [59]. Figure 6 shows four trajectory results. The PoseNet-like baseline fails to generalize under this setting and produce poor results. Conventional SLAM system like ORB-SLAM2 works well if there exists rich textures but tends to fail when large textureless region occurs, such as the first and the third cases shown in Figure 6. In most cases, thanks to joint dense correspondence learning, our method can establish accurate pixel associations to recover camera ego-motions and produce reasonably well trajectories, again demonstrating our improved generalization.

4.6. Discussion

Our experiments show that in addition to that our method maintains on par or even better performance on the widely tested KITTI benchmark, we achieve significant improvement on robustness and generalization from a variety of different aspects. This gain on generalization comes from our two novel designs as follows: 1) direct camera ego-motion prediction from optical flow, and 2) explicit scale alignment between the depth and the triangulated 3D structure. Our findings suggest that optical flow, which does not suffer from scale ambiguity naturally, is a more robust visual clue compared to relative pose estimation for deep learning models, especially under challenging scenarios. Likewise, explicitly handling the scale of depth and pose is still crucial for deep learning based visual SLAM. However, our current system cannot handle multi-view images where the motion magnitude is beyond the cost volume of optical flow, and pure rotation cannot be handled online with the two-view triangulation module.

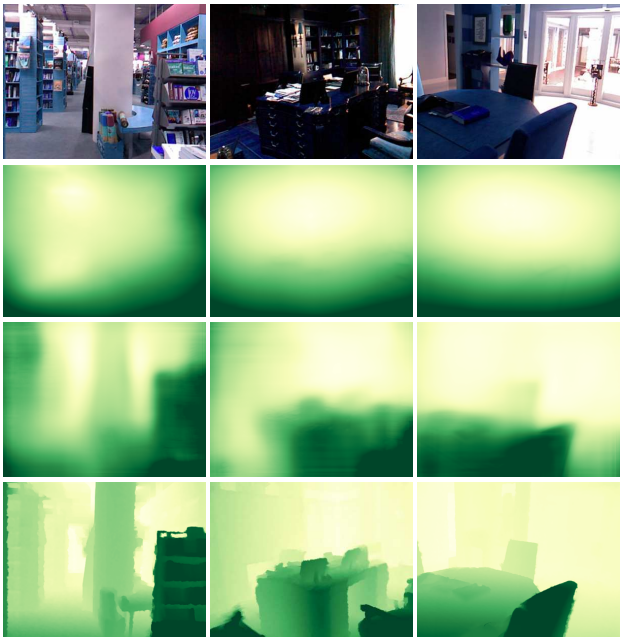

Figure 7. Depth estimation results on NYUv2 test data. **Top to bottom**: Input image, PoseNet baseline prediction, our prediction and depth groundtruth. PoseNet baseline fails to generalize for this indoor environment, which is also reported in [64].

5. Conclusion

In this paper, we propose a novel system which tackles the scale inconsistency for self-supervised joint depth-pose learning, by (1) directly recovering relative pose from optical flow and (2) explicit scale alignment between depth and pose via triangulation. Experiments demonstrate that our method achieves significant improvement on both accuracy and generalization ability over existing methods. Handling the above mentioned failure cases, developing general correspondence prediction and integration with back-end optimization could be interesting future directions.

Acknowledgements

This work was partially supported by NSFC (61725204, 61521002), BNRist and MOE-Key Laboratory of Pervasive Computing.

References

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *ECCV*, pages 404–417. Springer, 2006. 2

[2] Jiawang 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. In *NeurIPS*, pages 35–45, 2019. 1, 2, 3, 6, 7

[3] Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J Davison. Codeslam—learning a compact, optimisable representation for dense visual slam. In *CVPR*, pages 2560–2568, 2018. 2

[4] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *CVPR*, pages 6684–6692, 2017. 2

[5] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *AAAI*, volume 33, pages 8001–8008, 2019. 2, 5, 6

[6] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *ICCV*, pages 7063–7072, 2019. 2, 5, 6

[7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015. 2

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, pages 2366–2374, 2014. 2, 5

[9] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2, 4

[10] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, pages 2002–2011, 2018. 2, 8

[11] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, pages 740–756. Springer, 2016. 2

[12] Rahul Garg, Neal Wadhwa, Sameer Ansari, and Jonathan T Barron. Learning single camera depth estimation using dual-pixels. In *ICCV*, pages 7628–7637, 2019. 1, 2, 4

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE, 2012. 5, 6, 7

[14] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *CVPR*, pages 3828–3838, 2019. 5, 6

[15] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 270–279, 2017. 1, 2, 5

[16] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8977–8986, 2019. 5, 6

[17] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, and Adrien Gaidon. Packnet-sfm: 3d packing for self-supervised monocular depth estimation. *arXiv preprint arXiv:1905.02693*, 2019. 5

[18] Richard I Hartley. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):580–593, 1997. 4

[19] Richard I Hartley and Peter Sturm. Triangulation. *Computer Vision and Image Understanding*, 68(2):146–157, 1997. 4

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[21] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981. 2

[22] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, pages 2462–2470, 2017. 2, 6

[23] Joel Janai, Fatma Guney, Anurag Ranjan, Michael Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *ECCV*, pages 690–706, 2018. 2, 6

[24] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*, pages 2938–2946, 2015. 2

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[26] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: learning sfm from sfm. In *ECCV*, pages 698–713, 2018. 2, 4

[27] Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019. 2, 4

[28] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *CVPR*, pages 1119–1127, 2015. 8

[29] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, pages 4521–4530, 2019. 2, 4

[30] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pages 2041–2050, 2018. 2, 4

[31] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *CVPR*, pages 4571–4580, 2019. 2

[32] David G Lowe et al. Object recognition from local scale-invariant features. In *ICCV*, volume 99, pages 1150–1157, 1999. 2

[33] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981. 2

[34] 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. *arXiv preprint arXiv:1810.06125*, 2018. 2, 6

[35] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, pages 5667–5675, 2018. 2, 6

[36] Nikolaus Mayer, Eddy Ilg, Philipp Fischer, Caner Hazirbas, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. What makes good synthetic training data for learning disparity and optical flow estimation? *IJCV*, 126(9):942–960, 2018. 2

[37] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI*, 2018. 2, 6

[38] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 7, 8

[39] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS 2017 Autodiff Workshop: The Future of Gradient-based Machine Learning Software and Techniques*, 2017. 5

[40] René Ranftl and Vladlen Koltun. Deep fundamental matrix estimation. In *ECCV*, pages 284–299, 2018. 2

[41] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *CVPR*, pages 12240–12249, 2019. 2, 5, 6, 7

[42] Anirban Roy and Sinisa Todorovic. Monocular depth estimation using neural regression forest. In *CVPR*, pages 5506–5514, 2016. 1, 2

[43] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, volume 11, page 2. Citeseer, 2011. 2

[44] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. Understanding the limitations of cnn-based absolute camera pose regression. In *CVPR*, pages 3302–3312, 2019. 2, 3, 7

[45] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *NeurIPS*, pages 1161–1168, 2006. 2

[46] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840, 2008. 2, 8

[47] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2

[48] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 2

[49] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer, 2012. 5, 7

[50] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 573–580. IEEE, 2012. 5, 7, 8

[51] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018. 2, 3, 5, 6

[52] Chengzhou Tang and Ping Tan. Ba-net: Dense bundle adjustment network. In *ICLR*, 2019. 2

[53] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *CVPR*, pages 6243–6252, 2017. 2

[54] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, pages 2022–2030, 2018. 2, 3, 6

[55] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *CVPR*, pages 4884–4893, 2018. 3, 5

[56] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 3

[57] Thomas Whelan, Stefan Leutenegger, R Salas-Moreno, Ben Glocker, and Andrew Davison. Elasticfusion: Dense slam without a pose graph. Robotics: Science and Systems, 2015. 8

[58] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *CVPR*, pages 5354–5362, 2017. 8

[59] Fei Xue, Xin Wang, Shunkai Li, Qiuyuan Wang, Junqiu Wang, and Hongbin Zha. Beyond tracking: Selecting memory and refining poses for deep visual odometry. In *CVPR*, pages 8575–8583, 2019. 8

[60] Nan Yang, Rui Wang, Jorg Stuckler, and Daniel Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *ECCV*, pages 817–833, 2018. 2

[61] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, pages 1983–1992, 2018. 2, 3, 5, 6

[62] 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 *CVPR*, pages 340–349, 2018. 7

[63] Huangying Zhan, Chamara Saroj Weerasekera, Jiawang Bian, and Ian Reid. Visual odometry revisited: What should be learnt? *arXiv preprint arXiv:1909.09803*, 2019. 5

[64] Junsheng Zhou, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. Moving indoor: Unsupervised video depth learning in challenging environments. In *ICCV*, pages 8618–8627, 2019. 8

[65] Qunjie Zhou, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe. To learn or not to learn: Visual localization from essential matrices. *arXiv preprint arXiv:1908.01293*, 2019. 2, 3, 7

[66] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, pages 1851–1858, 2017. 1, 2, 5, 6, 7

[67] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *ECCV*, pages 36–53, 2018. 2, 6