On the Consistency of Self-Supervised Depth and Motion Estimation

Anonymous CVPR submission

Paper ID 1366

Abstract

The self-supervised learning of depth and pose from monocular sequence provides an attractive solution by using the photometric consistency of nearby frames as it depends much less on the ground-truth data. In this paper, we show that the photometric consistency is often not sufficiently competent to guarantee a good result because of the dynamic nature of real world scenarios. Then, we introduce several new loss terms from indirect self-generated geometric supervision to multi-view consistency inferred from the networks. As demonstrated on commonly used benchmarks, the proposed method achieves by far the best performance among the methods that simultaneously estimate monocular depth and relative poses.

1. Introduction

Along with the great success of deep neural networks on various semantic vision tasks such as recognition and segmentation [22], recent deep methods have interestingly shown promising abilities in modeling geometric relations, such as encoding depth and pose information. These geometric tasks used to have enough good closed-form solutions, yet some learning-based approaches are shown to improve traditional methods [52, 6] and have the potential to solve more challenging geometric tasks such as monocular depth estimation [43].

The joint learning of depth and relative pose from monocular videos [48, 53, 56] has been a heated research area due to its key role in *simultaneous localization and mapping* (SLAM) and *visual odometry* (VO) applications. The simplicity and the unsupervised nature make itself a potential replacement for traditional approaches that involve complicated geometric computations. Given adjacent frames, this approach uses convolutional neural networks (CNNs) to jointly predict the depth map of the target image and the relative poses from the target image to its visible neighboring frames. With the target image depth and relative poses, it minimizes the photometric error between the original target image and the synthesized images formed by bilinear-sampling [23] the adjacent views.

However, several existing problems hinder the performance of this type of methods. First, the photometric loss formulation requires the modeling scene to be static without non-Lambertian surfaces or occlusions. This assumption is often violated in imagery without careful photometric calibration, as well as street-view datasets [8, 16] with moving cars and pedestrians. Second, as the depth inference considers only a single image, there is no guarantee that adjacent frames would result in consistent depth estimation. Third, even though the method works on N-view images, only pairwise information is utilized in the training process. As a result, the scale of estimated $(N-1)$ relative poses is solely determined by the estimated depth of target image.

In this paper, we address the geometric consistency of self-supervised depth and pose estimation, seeking to mitigate or resolve the above limitations. We are mostly interested in the *unsupervised* monocular setting because it does not require stereo pairs or the costly ground-truth depth data. It is sometimes also referred to as the *self-supervised* approach since the supervisory signal inherits from the raw pixel or self-generated intermediate computations in traditional SfM methods [29]. We will use the two terms 'self-supervised' and 'unsupervised' interchangeably throughout the paper. Apart from the self-supervised photometric information that is used as the main source of supervision in the learning, we show that the intermediate geometric computation such as feature matches can also be employed in the training pipeline and greatly improve the performance. Our method is inspired by the several lines of research works. The general architecture is built upon the recent self-supervised learning approaches [56] that make use of photometric loss for depth and pose estimation. Yet, the motion module is inspired by indirect local-feature-based relative pose estimation and epipolar geometry [19, 21, 40]. The proposed formulation emphasizes the geometric consistency of deep interplay between pose and depth.

The consistency in multi-view data, or referred to as the data association problem [7], has been widely applied to and even forms the basis for many sub-steps in SfM, from feature matching [4], view graph construction [45, 54], mo-

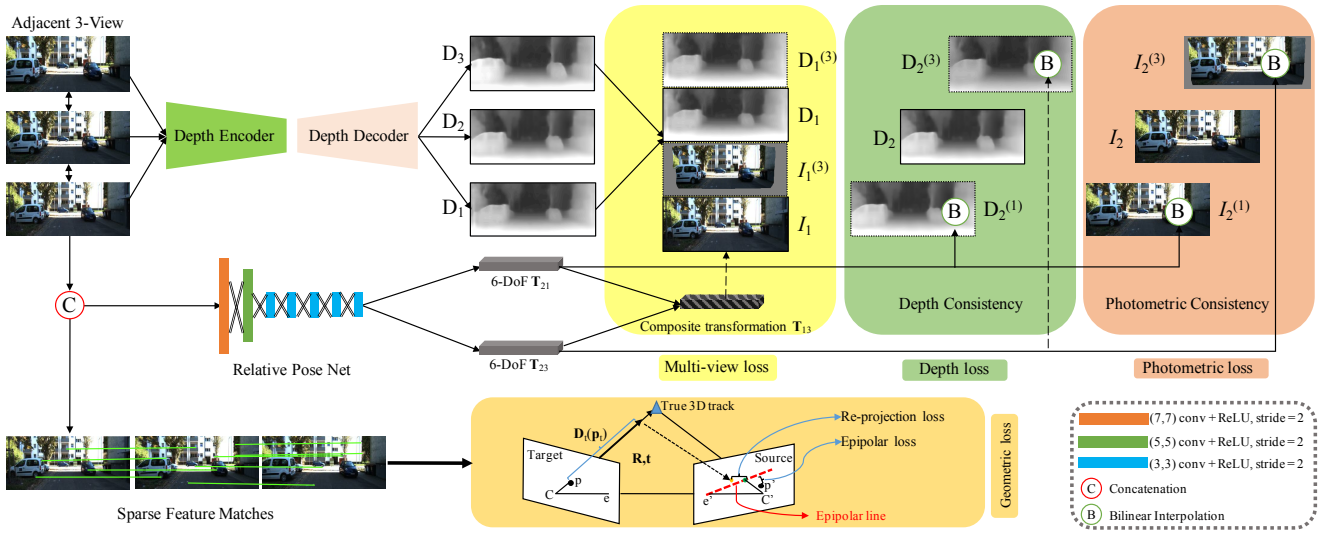


Figure 1. The architecture of our method. Our end-to-end training pipeline involves multiple loss terms to achieve photometric consistency, geometric consistency and multi-view consistency.

tion averaging [18] to bundle adjustment [46]. Yet enforcing the consistency is non-trivial in the learning-based setting. Our contributions are summarized twofold: 1) we incorporate sparse feature matches to the self-supervised learning framework, which introduces epipolar error and reprojection error to complement the noisy photometric loss; 2) we propose a novel depth estimation formulation that aims for consistent outputs of adjacent views. Throughout extensive evaluations on several benchmark datasets [8, 16], we show that the deep interaction of multi-view depths and poses achieves the state-of-the-art performance.

2. Related Works

Structure-from-Motion and Visual SLAM. Structure-from-Motion (SfM) [2] and visual SLAM problems aim to simultaneously recover the camera pose and 3D structures from images. Both problems are well studied and render practical systems [11, 38, 51] by different communities for decades, with the latter emphasizes more on the real-time performance. The self-supervised depth and motion learning framework derives from direct SLAM methods [11, 12, 39]. Different from indirect methods [9, 30, 38] that use reliable sparse intermediate geometric quantities like local features [42], direct methods optimize the geometry using all the pixels in the image. With accurate photometric calibration such as gamma and vignetting correction [27], this formulation does not rely on sparse geometric computation and is able to generate finer-grained geometry. However, this formulation is less robust than indirect ones when the photometric loss is not meaningful, the scene containing moving or non-Lambertian objects.

Supervised Approaches for Learning Depth. Some early monocular depth estimation works rely on information from depth sensors [10, 43] without the aid of geometric relations. Liu *et al*. [33] combine deep CNN and conditional random field for estimating single monocular images. DeMoN [47] is an iterative supervised approach to jointly estimate optical flow, depth and motion. This coarse-to-fine process considers the use of stereopsis and produces good results with both depth and motion supervision.

Unsupervised/Self-Supervised Approaches. The self-supervised approaches for SfM stem from warping-based view synthesis [57], a classical paradigm of which is to composite novel view based on the underlying 3D geometry. Garg *et al*. [15] propose to learn depth using stereo camera pairs with known relative pose, yet the depth learning formulation is unsupervised in nature. Godard *et al*. [17] improve this training paradigm with left-right consistency checking. The above two methods both rely on a calibrated stereo rig and are not applicable to monocular settings with unknown poses. The joint unsupervised optimization of depth and pose starts from Zhou *et al*. [56] and Vijayanarasimhan *et al*. [48]. They propose similar approaches that use two CNNs to estimate depth and pose separately, and constrain the outcome with photometric loss. Later, a series of improvements [29, 36, 49, 53, 55] are proposed. Wang *et al*. [49] discuss the scale ambiguity and combine the estimated depth with direct methods [11]. Zhan *et al*. [55] consider warping deep features from the neural nets instead of the raw pixel values. Klodt *et al*. [29] propose to integrate weak supervision from SfM methods. Mahjourian *et al*. [36] employ geometric constraints of the scene by en-

forcing an approximate ICP based matching loss. We shall discuss and compare these unsupervised joint learning approaches in detail in the experiments.

3. Method

3.1. Problem Formulation

We first formalize the problem and present effective practices employed by previous methods [29, 36, 48, 49, 53, 55, 56]. Given adjacent N-view monocular image sequences (e.g. $\{\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3\}$ for $N = 3$), the unsupervised depth and motion estimation problem aims to simultaneously estimate the depth $\mathbf{D}_t$ of the target (center) image ($\mathcal{I}_2$ in the 3-view case) and the 6-DoF relatives poses $\mathbf{T}_{t\to s} = [\mathbf{R}_{t\to s}|\mathbf{t}_{t\to s}] \in \mathcal{SE}(3)$ to $N-1$ source views ($\mathcal{I}_1$ and $\mathcal{I}_3$), using CNNs with photometric supervision.

For a source-target view pair $(\mathcal{I}_s, \mathcal{I}_t)$, $\mathcal{I}_t$ can be inversely warped to the source frame $\mathcal{I}_t$ given the estimated depth $\mathbf{D}_t$ and the transformation from target to source $\mathbf{T}_{t\to s}$. Formally, given a pixel coordinate p_t in $\mathcal{I}_t$ which is co-visible in $\mathcal{I}_s$, the pixel coordinate p_s in $\mathcal{I}_s$ is given by the following equation which determines the warping transformation

$$p_s \sim \mathbf{K}_s[\mathbf{R}_{t\to s}|\mathbf{t}_{t\to s}]\mathbf{D}_t(p_t)\mathbf{K}_t^{-1}p_t, \tag{1}$$

where $\sim$ denotes 'equal in the homogeneous coordinates', $\mathbf{K}_s$ and $\mathbf{K}_t$ are the intrinsics for the input image pair, and $\mathbf{D}_t(p_t)$ is the depth for this pairticular p_t in $\mathcal{I}_t$.

With this coordinate transformation, synthesized images can be generated from the source view using the differentiable bilinear-sampling method [23]. The unsupervised framework then minimizes the pixel error between the target view and the synthesized image

$$\mathcal{L}_{pixel} = \frac{1}{|\mathcal{M}|}\sum_{\forall p_t \in \mathcal{M}} \left|\widetilde{\mathcal{I}}_t^{(s)}(p_t|\mathbf{R}_{t\to s}, \mathbf{t}_{t\to s}, \mathbf{D}_t) - \mathcal{I}_t(p_t)\right|, \tag{2}$$

where $\widetilde{\mathcal{I}}_t^{(s)}$ represents the synthesized target image from source image. $\mathcal{I}(p)$ is the function that maps the image coordinate p in image $\mathcal{I}$ to pixel value, and the first term $\widetilde{\mathcal{I}}_t^{(s)}$ is the bilear-sampling operation used to acquire the synthesized view given relative motion and depth. $\mathcal{M}$ is a binary mask that determines if the inverse warping fall into a valid region in the source image, and can be computed analytically given the per-pixel depth and relative transformation. $|\mathcal{M}|$ denotes the total number of valid pixels.

In addition to the per-pixel error, structured similarity (SSIM) [50] is shown to improve the performance [17, 53], which is defined on local image patches x and y rather than every single pixel. We follow the previous approaches [36, 53] to compute the *SSIM* loss on 3×3 image patches ($c_1 = 0.01^2, c_2 = 0.03^2$) as follows

$$\mathcal{L}_{SSIM}(\mathcal{I}_s, \mathcal{I}_t) = \frac{1}{2}(1 - \sum_{\forall x \in \mathcal{I}_s, \forall y \in \mathcal{I}_t} \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2\mu_y^2 + c_1)(\sigma_x + \sigma_y + c_2)}). \tag{3}$$

The depth map is further constrained by the smoothness loss to push the gradients to propagate to nearby regions, known as the gradient locality issue [5]. Specifically, we adopt the image-aware smoothness formulation [53] which allows sharper depth changes on edge regions

$$\mathcal{L}_{smooth} = \sum_{\forall p_t \in \mathcal{I}_t} |\nabla\mathbf{D}_t(p_t)|^T \cdot e^{-|\nabla\mathcal{I}_t(p_t)|}, \tag{4}$$

where ∇ denotes the 2D differential operator for computing image gradients. Optimizing a combination of above loss terms wraps up the basic formulation of training objectives, which forms the baseline written as

$$\mathcal{L}_{baseline} = \alpha\mathcal{L}_{pixel} + (1-\alpha)\mathcal{L}_{SSIM} + \beta\mathcal{L}_{smooth}. \tag{5}$$

However, there are drawbacks with the basic formulation. We then describe the key ingredients our contributions.

3.2. Injecting Weak Geometric Supervision

The success of unsupervised joint training requires several important assumptions: 1) The modeling scene should be static without moving objects; 2) The surfaces in the scene should be Lambertian; 3) No occlusion exists between the target view and source view; 4) Cameras should be photometrically calibrated, a technique adopted in the direct SLAM method [11, 12], to compensate for vignetting effect and exposure time. Violation to any of the above criterions would lead to model bias or data bias. The first three assumptions are inevitably violated to some extent because it is hard to capture temporally static images with no occlusion in the real world. The fourth restriction is often neglected by datasets with no photometric calibration parameters provided.

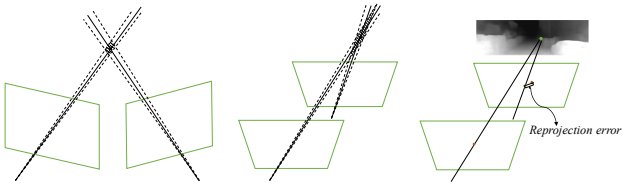
To address these limitations, previous methods [56, 29] additionally train a mask indicating whether the photometric loss is meaningful. Yet, we present a novel approach to tackle this issue by injecting indirect geometric information into the direct learning framework. Different from direct methods that rely on dense photometric consistency, indirect methods for SfM and visual SLAM are based on sparse local descriptors such as SIFT [51] and ORB [38]. Local invariant features are much less likely to be affected by the scale and illumination changes and can be implicitly embedded into the learning framework.

Epipolar error. Assuming the pinhole camera model, the sparse feature matches $\mathcal{S}_{t\leftrightarrow s} = \{\mathbf{p} \leftrightarrow \mathbf{p}'\}$ between the target and source views satisfy the epipolar constraint. Here we assume the calibrated setting where pixel coordinates

are transformed to image coordinate by $\mathbf{K}^{-1}$. The loss with the feature matches and the estimated pose can be quantified using the *symmetric epipolar distance* [19]

$$\mathcal{L}_{epi}(\mathcal{S}|\mathbf{R},\mathbf{t}) = \sum_{\forall(\mathbf{p},\mathbf{p}')\in\mathcal{S}} \left(\frac{\mathbf{p}'^{\mathbf{T}}\mathbf{E}\mathbf{p}}{\sqrt{(\mathbf{E}\mathbf{p})_{(1)}^2 + (\mathbf{E}\mathbf{p})_{(2)}^2}} + \frac{\mathbf{p}^{\mathbf{T}}\mathbf{E}\mathbf{p}'}{\sqrt{(\mathbf{E}\mathbf{p}')_{(1)}^2 + (\mathbf{E}\mathbf{p}')_{(2)}^2}}\right), \tag{6}$$

where $\mathbf{E}$ being the essential matrix computed by $\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$, $[\cdot]_\times$ is the matrix representation of the cross product with $\mathbf{t}$. We simply omit the subindices for conciseness ($\mathcal{S}$ for $\mathcal{S}_{t\leftrightarrow s}$, $\mathbf{R}$ for $\mathbf{R}_{t\to s}$, $\mathbf{t}$ for $\mathbf{t}_{t\to s}$).



(a) Motion with proper baseline (b) Forward motion (c) Re-projection error using depth

Figure 2. (a) For two images with proper motion baseline, the uncertainty (shaded region) is small. (b) For forward motion with narrow baseline, the uncertainty is large. (c) The re-projection error unites estimated depth and pose with sparse features, and does not involve triangulation uncertainty.

Re-projection error. The epipolar constraint does not concern the depth in its formulation. To involve depth optimization using the feature match supervision, there are generally two methods: 1) triangulate the correspondence $\mathbf{p} \leftrightarrow \mathbf{p}'$ using the optimal triangulation method [19] assuming the Gaussian noise model, to obtain the 3D track for depth supervision; 2) back-project 2D features in one image using the estimated depth to compute the 3D track, and re-project the 3D track to another image to compute the re-projection error. For street-view ego-motion scenes [8, 16], points are less precisely triangulated since rays are almost parallel in forward motion (see Figure 2 for illustration). Therefore, we take the second method to weakly supervise the depth estimation by computing the re-projection loss

$$\mathcal{L}_{reproj}(\mathcal{S}|\mathbf{R},\mathbf{t},\mathbf{D}_t) = \sum_{\forall p\leftrightarrow p'\in\mathcal{S}} \left\| [\mathbf{R}|\mathbf{t}]\hat{\mathbf{D}}_t(\mathbf{p})\mathbf{p} - \mathbf{p}' \right\|_2, \tag{7}$$

where $\hat{\mathbf{D}}_t(\mathbf{p})$ is the bilinear-sampling operation [23] in the target depth map as the feature coordinate p is not an integer. Minimizing re-projection error using feature matches can be viewed as creating sparse anchors between the weak geometric supervision and the estimated depth and pose. In contrast, Equation 6 does not involve the estimated depth.

Since outliers may exist if they lie close to the epipolar line, we use the pre-computed pairwise matches co-existed in three views filtered by geometric verification [20]. Minimizing the epipolar error and re-projection error of all matches using CNNs mimics the non-linear pose estimation [3]. Our experiments show that this weak supervisory signal significantly improves the pose estimation and is superior to using other intermediate SfM supervisions such as poses and sparse depth.

3.3. Consistent Depth Estimation

In this section, we describe the depth estimation module. Previous methods, whether it is operated on 3-view or 5-view, are pairwise approaches in essence because loss terms are computed pairwisely from the source frame and target frame. Even though the pose network outputs $N-1$ relative poses at once, it is unknown if these relative poses are aligned to the same scale. We propose the motion-consistent depth estimation formulation to address this issue. Rather than minimizing the loss between the target frame and adjacent source frames, our proposed formulation also considers the depth and motion consistency between adjacent frames themselves.

Forward-backward consistency. As shown in Figure 1, our network architecture estimates the depth maps of the target image ($\mathcal{I}_t$), as well as the forward and backward depths. Inspired by [17, 41] that uses left-right consistency on stereo images, we propose the forward-backward consistency for monocular images. In addition to bilinear-sampling from the source image pixel values, we also sample the estimated depth maps of forward and backward images ($\mathcal{I}_s$). This process generates two synthesized depth maps $\widetilde{\mathbf{D}}_t^{(s)}$ that can be used to constrain the estimation of the target image depth $\mathbf{D}_t$.

For the learning setting with stereo images, the images are rectified in advance so the scale ambiguity issue is not considered. While for learning monocular depth, the estimated depth is determined up to scale, therefore the alignment of depth scale is necessary before constraining the depth discrepancy. We first normalize the target depth map using its median $\mathbf{D}_2 :\leftarrow \mathbf{D}_2/\text{median}(\mathbf{D}_2)$ [49], which determines the scale of relative poses. Then we apply a mean alignment to the synthesized depth maps and the normalized target depth map in the corresponding region informed by the analytical mask $\mathcal{M}$ (Equation 2), and further optimize the depth discrepancy

$$\mathcal{L}_{depth} = \frac{1}{|\mathcal{M}|}\sum_{\forall p\in\mathcal{M}} \left| \frac{mean(\mathbf{D}_t \circ \mathcal{M})}{mean(\widetilde{\mathbf{D}}_t^{(s)} \circ \mathcal{M})} \cdot \widetilde{\mathbf{D}}_t^{(s)}(p) - \mathbf{D}_t(p) \right|, \tag{8}$$

where $\circ$ means the element-wise multiplication and the loss is averaged over all the valid pixel p in the mask $\mathcal{M}$.

Multi-view consistency. The above losses are all defined on the single target image (e.g. smoothness loss) or among image pairs, even though the input is N-view ($N \geq 3$) image sequences. The pose network output $N-1$ relative poses between the target and source images, but the $N-1$ relative poses are only weakly connected by the monocular depth. To strengthen the scale consistency for triplet relation, we propose the multi-view consistency loss which penalizes inconsistency of the forward depth and backward depth using the target image as a bridge for scale alignment. Formally, given image sequence $(\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3)$ with $\mathcal{I}_2$ the target image, and corresponding pose and depth predictions $(\mathbf{T}_{2\to1}, \mathbf{T}_{2\to3})$ and $(\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3)$, we again obtained the normalized depth map $\overline{\mathbf{D}}_1 = s_{12} \cdot \mathbf{D}_1$ where the scaling ratio $s_{12} = \frac{mean(\mathbf{D}_2 \circ \mathcal{M}_{12})}{mean(\widetilde{\mathbf{D}}_2^{(1)} \circ \mathcal{M}_{12})}$ as used in Equation 8. The transformation from the backward image $\mathcal{I}_1$ to the forward image $\mathcal{I}_3$ is $\mathbf{T}_{1\to3} = \mathbf{T}_{2\to1}^{-1} \cdot \mathbf{T}_{2\to3}$. The multiview loss minimizes the depth consistency term and photometric consistency term as

$$\begin{aligned}\mathcal{L}_{multi} =& \alpha \mathcal{L}_{pixel}(\mathcal{I}_3|\overline{\mathbf{D}}_1, \mathbf{T}_{1\to3}) + (1-\alpha)\mathcal{L}_{SSIM}(\mathcal{I}_3|\overline{\mathbf{D}}_1, \mathbf{T}_{1\to3}) \\ &+ \frac{1}{|\mathcal{M}_{13}|} \sum_{\forall p \in \mathcal{M}_{13}} \left| \overline{\mathbf{D}}_1(p) - \overline{\mathbf{D}}_1^{(3)}(p) \right|,\end{aligned} \tag{9}$$

where $\overline{\mathbf{D}}_1^{(3)}$ is the synthesized normalized depth for $\overline{\mathbf{D}}_1$ given $\overline{\mathbf{D}}_3$ and $\mathbf{T}_{1\to3}$. The subindices 1 and 3 are interchangable in the above Equation 9. $\mathcal{L}_{multi}$ goes beyond the pairwise loss terms $\mathcal{L}_{pixel}$, $\mathcal{L}_{SSIM}$, $\mathcal{L}_{epi}$ and $\mathcal{L}_{depth}$ because it utilizes the chained relative pose and pushes the two sub-relative poses to be aligned on the same scale. This benefits the monocular SLAM because it facilitates the incremental localization by aligning multiple N-view outputs.

3.4. Modeling the Data Uncertainty

The intrinsic noises (e.g. moving objects, occlusions and non-Lambertian surfaces) in the data pose challenges for the effectiveness of the photometric loss formulation. The proposed indirect supervision is able to mitigate the problem to some extend but still cannot fully solve it. Zhou *et al.* [56] tackle this weakness by training a mask to filter out probable erroneous pixels. We follow this procedure and reiterate it with a more formal representation using the *aleatoric* uncertainty theory. *Aleatoric* uncertainty [26] captures noise inherent in the observations, in contrast to epistemic uncertainty [14] which models the inherent noise in the model parameters. Specifically, heteroscedastic aleatoric uncertainty assume that the observation noise can vary with the input data, which fits our scenario where dynamic objects appear occasionally in the data. We use the Laplace distribution $f(\tilde{\mathcal{I}}|\mathcal{I}, \sigma(p_t)) = \frac{1}{2\sigma(p_t)} \exp(-\frac{|\tilde{\mathcal{I}}-\mathcal{I}|}{\sigma(p_t)})$ to model the heteroscedastic aleartoric uncertainty, whose negative log likelihood corresponds to the $L1$ minimization in Equation 2&3

Algorithm 1 Consistent Depth-Motion Training Loss

Require: $\{\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3\}$: input 3-view image snippet with $\mathcal{I}_2$ ($t=2$) as the target image; $\mathcal{S} = \{\mathbf{p} \leftrightarrow \mathbf{p}'\}$: geometrically verified pairwise matches ($\mathcal{S}_{21}$ and $\mathcal{S}_{23}$)

Ensure: multi-view consistent depth $\mathbf{D}_2$ and relative poses $\mathbf{T}_{21} = \{\mathbf{R}_{21}, \mathbf{t}_{21}\}, \mathbf{T}_{23} = \{\mathbf{R}_{23}, \mathbf{t}_{23}\}$

1: Obtain the estimated depth map $\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3$ and 6DoF pose vectors (Euler angles θ and displacement t), transform pose vector to matrix representation $\mathbf{T}_{21}$ and $\mathbf{T}_{23}$.

2: $\mathbf{D}_2 :\leftarrow \mathbf{D}_2/\text{median}(\mathbf{D}_2)$. Normalize the target depth $\mathbf{D}_2$ using its median, a critical step [49] to fix the scale of the image triplet.

3: Compute the $\mathcal{L}_{pixel}$, $\mathcal{L}_{SSIM}$ pairwisely between $(\mathcal{I}_1, \mathcal{I}_2)$ and $(\mathcal{I}_2, \mathcal{I}_3)$, and $\mathcal{L}_{smooth}$ on $\mathcal{I}_2$ given estimated depth and relative poses.

4: Compute $\mathcal{I}_{epi}$ and $\mathcal{I}_{reproj}$ pairwisely between $(\mathcal{I}_1, \mathcal{I}_2)$ and $(\mathcal{I}_2, \mathcal{I}_3)$ given feature matches $\mathcal{S}$.

5: First align the scales of $\mathbf{D}_1$ and $\mathbf{D}_3$ using Equation 8 and compute $\mathcal{L}_{depth}$ pairwisely.

6: First compute the chained motion and normalized depth map $\overline{\mathbf{D}}_1$ and $\overline{\mathbf{D}}_3$, then compute $\mathcal{L}_{multi}$ which is the only loss term that involves all the estimated quantities.

7: Losses involving per-pixel depth map are weighted by the data-dependent variance term $\sigma(p_t)$ as in Equation 10 and the total loss is given by Equation 11.

and the variance $\sigma(p_t)$ varies with respect to every pixel p_t. With the data-dependent variance term, we can for example rewrite the pixel error (Equation 2) using negative log likelihood as

$$\begin{aligned}\mathcal{L}_{pixel}^{\star} \propto& \frac{1}{|\mathcal{M}|} \sum_{\forall p_t \in \mathcal{M}} \left(\frac{1}{\sigma(p_t)} \left| \tilde{\mathcal{I}}_t^{(s)}(p_t) - \mathcal{I}_t(p_t) \right| + \log \sigma(p_t) \right) \\ =& \frac{1}{|\mathcal{M}|} \sum_{\forall p_t \in \mathcal{M}} \left(\exp(-s(p_t)) \left| \tilde{\mathcal{I}}_t^{(s)}(p_t) - \mathcal{I}_t(p_t) \right| + s(p_t) \right).\end{aligned} \tag{10}$$

In practice, we compute the log variance $s(p_t) = \log \sigma(p_t)$ as an additional regression split from the last convolutional layer of the depth estimation network. We train the network to predict $\log \sigma(p_t)$ because it is claimed to be more numerically stable [26]. Our final formulation takes into account the basic losses in Equation 5, the feature matching terms, as well as the consistency terms, written as

$$\begin{aligned}\mathcal{L}_{total} =& \alpha \mathcal{L}_{pixel} + (1-\alpha)\mathcal{L}_{pixel} + \beta \mathcal{L}_{smooth} + \\ & \gamma_1 \mathcal{L}_{epi} + \gamma_2 \mathcal{L}_{reproj} + \mu_1 \mathcal{L}_{depth} + \mu_2 \mathcal{L}_{multi}.\end{aligned} \tag{11}$$

For loss terms concerning per-pixel depth map, we can employ the data uncertainty similar to the one in Equation 10. We compare the total loss with and without the aleatoric uncertainty mask in the experiments. The weighting for different losses are set empirically given hyperparameters in previous methods and our attempts ($\alpha =$

Table 1. Single-view depth estimation performance. The statistics for the compared methods are excerpted from corresponding papers, except that the results marked with *'updated'* are captured from the websites. 'K' represents KITTI raw dataset (Eigen split) and CS represents cityscapes training dataset. The method [17] marked with ⋆ are trained and tested on larger scale (256×512) images, whereas others use 128×416 images. The metrics marked by red means 'the lower the better' and the ones marked by green means 'the higher the better'. The best results for each catergory are **bolded**.

Method	Supervision	Dataset	Cap (m)	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen *et al.* [10] Fine	Depth	K	80	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu *et al.* [33]	Depth	K	80	0.202	1.614	6.523	0.275	0.678	0.895	0.965
Godard *et al.* [17]⋆	Stereo/Pose	K	80	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Godard *et al.* [17]⋆	Stereo/Pose	K + CS	80	**0.114**	**0.898**	**4.935**	**0.206**	**0.861**	**0.949**	**0.976**
Zhou *et al.* [56] updated	No	K	80	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Zhou *et al.* [56] updated	No	K	-	0.185	2.170	6.999	0.271	0.734	0.901	0.959
Klodt *et al.* [29]	No	K	80	0.166	1.490	5.998	-	0.778	0.919	0.966
Mahjourian *et al.* [36]	No	K	80	0.163	1.24	6.22	0.25	0.762	0.916	0.968
Wang *et al.* [49]	No	K	80	0.151	1.257	5.583	0.228	0.810	0.936	0.974
Yin *et al.* [53]	No	K	80	0.155	1.296	5.857	0.233	0.793	0.931	0.973
Yin *et al.* [53]	No	K	-	0.156	1.470	6.197	0.235	0.793	0.931	0.972
Yin *et al.* [53] updated	No	K + CS	80	0.149	1.060	5.567	0.226	0.796	0.935	0.975
Ours	No	K	80	0.140	1.014	5.473	0.222	0.816	0.937	0.974
Ours	No	K	-	0.140	1.014	5.473	0.222	0.816	0.937	0.974
Ours	No	K + CS	80	**0.139**	**0.964**	**5.309**	**0.215**	**0.818**	**0.941**	**0.977**
Garg *et al.* [15]	Stereo/Pose	K	50	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Zhou *et al.* [56]	No	K	50	0.201	1.391	5.181	0.264	0.696	0.900	0.966
Yin *et al.* [53]	No	K + CS	50	0.147	0.936	4.348	0.218	0.810	0.941	0.977
Ours	No	K	50	**0.133**	**0.778**	**4.069**	**0.207**	**0.834**	**0.947**	**0.978**

$0.15, \beta = 0.1, \gamma_1 = \gamma_2 = 0.001, \mu_1 = \mu_2 = 0.1$). Though we can also learn the optimal weighting using homoscedastic uncertainty [25], we find that it achieves no better result than empirically setting the weights. To avoid confusions for the proposed loss terms, we summarize the training loss composition in Algorithm 1.

4. Experiments

4.1. Training Dataset

KITTI. We evaluate our method on the widely-used KITTI datasets [16, 37]. To conduct fair comparisons with previous methods, we use the KITTI raw dataset with Eigen split [10] for evaluating depth estimation, and the KITTI odometry dataset for pose estimation. Training and testing images are down-sampled to 128×416 to facilitate the training and provide a fair evaluation setting. For Eigen split, we use 20129 images for training and 2214 images for validation. The 697 testing images are selected by [10] from 28 scenes. The images in these scenes are excluded from the training set. The KITTI odometry dataset contains 11 street-view scenes with ground-truth poses. We follow the previous convention [53, 56] to train the model on sequence 00-08 and test on sequence 09-10. We further split sequence 00-08 to 18361 images for training and 2030 for validation. Note that the two KITTI datasets may overlap with each other, the models for evaluating depth and pose are different even though trained with the same pipeline.

Cityscapes. We also try pre-training the model on the Cityscapes [8] dataset since it is shown that starting from a pre-trained model boosts the performance [56]. The process is conducted without adding feature matches for 60k steps. 88084 images are used for training and 9659 images for validation.

4.2. Implementation Details

Data preparation. Since our method utilizes the feature matches as a weak geometric supervision, we extract SIFT feature [35] and conduct pairwise matching offline using SiftGPU [51]. The putative matches are further filtered by geometric verification using fundamental matrix [21] with RANSAC [13]. We select 100 pairs of matches for each image pairs and use them for training. Matches are only used for training and not necessary for inference.

Learning. We implement our learning pipeline using Tensorflow [1]. We use training images of size 128×416, which is consistent with most of the previous unsupervised learning approaches [53, 56] to set up a comparable setting. We use ResNet-50 [22] as the depth encoder as it is proved to be a more effective model [53]. The depth decoder is a symmetric architecture that upsamples the encoded depth feature using transpose convolutions. The relative pose net is composed of 7 convolutional layers (one (7×7)-kernel layer, one (5×5)-kernel layer and five (3×3)-kernel layers), with the lengths of feature maps reduced by half and the number of feature channels multiplied by two from each previous layer. If not explicitly specified, we train the neural

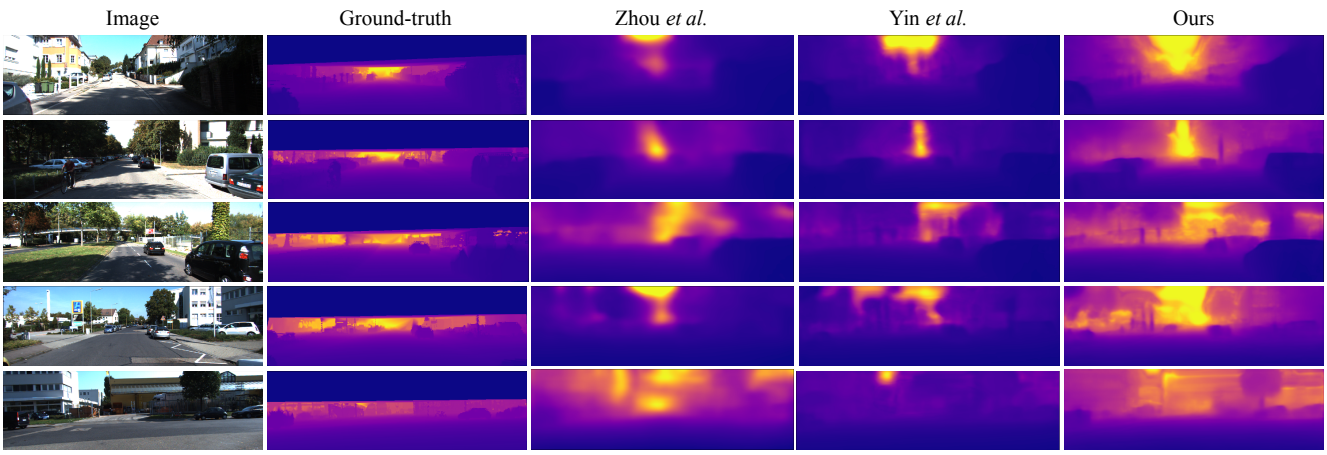


Figure 3. Qualitative comparison for depth estimation on the Eigen split. The predicted depth maps are first aligned with the ground-truth using medians. Then the depth values larger than 80m are set to 80m to ensure a consistent color scale. It shows that our result best reflects the ground-truth depth range and contains richer details (best view in color).

networks using 3-view image sequences as the photometric error would accumulate for longer input sequences. We use the Adam [28] solver with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a learning rate of 0.0001 and a batch size of 4. We adopt ResNet-50 [22] as the depth encoder and the same architecture for pose network as [56].

4.3. Depth Estimation

We compare our depth estimation performance with state-of-the-art supervised and unsupervised methods. The evaluation metrics protocol follows a series of previous works [36, 53, 56]. As shown in Table 1, our method achieves the best performance among all the methods that jointly learn depth and pose. Previous methods often filter the predicted depth map by setting a maximum depth at 50m or 80m (the ground-truth depth range is within 80m) before computing depth error, since distant pixels may have larger prediction error. We also evaluate the performance without filtering the predicted depth maps. Table 1 shows that without capping the maximum depth [53, 56] become worse while our result seldomly changes, meaning our consistent training renders depth predictions with little noise and this filtering process is unnecessary. Figure 3 provides a qualitative comparison of the predictions. We show the depth value (the nearer the darker) instead of the inverse depth (disparity) parameterization, which makes the distant region stand out.

It is noted that our method achieves worse performance than [17] which is trained and tested on larger images using the rectified stereo images pairs without estimating the relative pose. It is also noteworthy that [29] and ours both use self-supervised weak supervisions. This implies that the raw feature matches are more robust as the supervisory signal, whereas using pose and depth computed from classical SfM [29] is possible to introduce additional bias inherited from the PnP [32] or triangulation algorithms.

Table 2. Pose estimation evaluation. All the learning-based methods are trained and tested on 128 $\times$ 416 images, while ORB-SLAM2 are tested on full-sized (370 $\times$ 1226) images.

Method	Seq 09	Seq 10
ORB-SLAM2 [38]	0.014 ± 0.008	0.012 ± 0.011
Zhou *et al.* [56] updated (5-frame)	0.016 ± 0.009	0.013 ± 0.009
Yin *et al.* [53] (5-frame)	0.012 ± 0.007	0.012 ± 0.009
Mahjourian *et al.* [36] , no ICP (3-frame)	0.014 ± 0.010	0.013 ± 0.011
Mahjourian *et al.* [36] , with ICP (3-frame)	0.013 ± 0.010	0.012 ± 0.011
Klodt *et al.* [29] (5-frame)	0.014 ± 0.007	0.013 ± 0.009
Ours *et al.* (3-frame)	$\mathbf{0.009 \pm 0.005}$	$\mathbf{0.008 \pm 0.007}$

4.4. Pose Estimation

We evaluate the performance of relative pose estimation on the KITTI odometry datasets. We have observed that with the pairwise matching supervision, the result for motion estimation has been extensively improved. We measure the Absolute Trajectory Error (ATE) over N-frame snippets. The mean error and variance is averaged from the full sequence. As shown in Table 2, with the same underlying network structure, the proposed method outperforms state-of-the-art methods by a large margin.

4.5. Ablation Study

Performance with different modules. After horizontally comparing with different architectures, we also evaluate our method vertically to show the effects of different modules. The benchmarks are conducted in terms of both depth and pose trained solely on KITTI raw dataset and KITTI odometry dataset correspondingly. Since there are too many combinations of using different loss terms, we choose an incremental order to add the proposed techniques. When the

Table 3. Evaluation of different training loss configurations. All models are either solely trained on KITTI raw dataset (for depth) or KITTI odometry dataset (for pose) without pre-training on Cityscapes. The depth estimation performance is evaluated with maximum depth set/capped at 80m. All models except the last two are trained on 3-view image sequences. The best result for each metric is **bolded**.

Loss Configuration						**Depth** (KITTI raw Eigen split)							**Pose** (KITTI odometry)	
Baseline	Epipolar	Re-projection	Forward-backward	Multi-view	Uncertainty	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25^3$	$\delta < 1.25^2$	$\delta < 1.25^3$	Seq 09	Seq 10
✓	-	-	-	-	-	0.163	1.371	6.275	0.249	0.773	0.918	0.966	0.014 ± 0.009	0.012 ± 0.012
✓	✓	-	-	-	-	0.159	1.287	5.725	0.239	0.791	0.927	0.969	0.010 ± 0.005	0.009 ± 0.008
✓	✓	✓	-	-	-	0.152	1.205	5.56	0.227	0.800	0.935	0.973	0.009 ± 0.005	0.009 ± 0.008
✓	✓	✓	✓	-	-	0.146	1.391	5.791	0.229	0.814	0.936	0.972	0.009 ± 0.005	0.008 ± 0.007
✓	✓	✓	✓	✓	-	**0.140**	1.014	5.473	0.222	**0.816**	0.937	0.974	$\mathbf{0.009 \pm 0.005}$	$\mathbf{0.008 \pm 0.007}$
✓	✓	✓	✓	✓	✓	0.143	**1.003**	**5.238**	**0.216**	0.812	**0.942**	**0.978**	0.009 ± 0.005	0.008 ± 0.007
✓ (5-view)	-	-	-	-	-	0.169	1.607	6.129	0.255	0.779	0.917	0.963	0.014 ± 0.009	0.013 ± 0.009
✓ (5-view)	✓	✓	-	-	-	0.157	1.449	5.796	0.239	0.803	0.929	0.970	0.012 ± 0.008	0.010 ± 0.007

configuration changes, we restart the training from scratch and the statistics are recorded when the training process is converged. Based on the results in Table 3, we have the following observations:

- The re-implemented baseline model (first line in Table 3), using the total loss from Equation 5, has already surpassed several models [29, 36, 56]. The reasons can be attributed to 1) the strong depth encoder ResNet-50 [22], the same as in [53]; 2) the *SSIM* loss which [56] lacks.
- Combined with the epipolar loss term $\mathcal{L}_{epi}$, the result for the relative pose estimation is greatly improved. It shows the efficacy of using raw feature matches as the weakly supervised signal. However, the improvement for depth estimation is not as significant as pose estimation, because $\mathcal{L}_{epi}$ does not concern the estimated depth.
- Re-projection loss inferred from sparse feature matches further improves the monocular depth inference. The improvements for pose estimation brought by ingredients other than the epipolar loss are marginal.
- The forward-back consistency and multi-view consistency are the essential parts for the improvement in depth estimation.
- Adding probabilistic uncertainty mask only improves the result slightly on some metrics, though it completes the theoretical foundation by taking data uncertainty into consideration. This observation is consistent with [56] that it is possible to be determined by the characteristics of the training dataset (KITTI).

In summary, the proposed weak geometric supervision helps most for the pose estimation, while the consisteny terms proposed in Section 3.3 essentially improve monocular depth estimation.

Sequence Length. The multi-view depth consistency loss brings boosts for depth estimation. However, the performance boost can be also attributed to using longer image snippets, since similar second-order relations can be exploited by using 5-view image sequences for training. Therefore, we further evaluate the performance of using 5-view training images with different configurations. As shown in Table 3, training on longer image sequences would deteriorate the performance, because long sequences also contain larger photometric noises. We argue that the proposed formulation elevates the results not from more data terms, but the consistency embedded in geometric relations.

Table 4. Generalization experiments on Make3D. The evaluation metrics are the same as the ones used in Table 1 except the last one (RMSE $\log_{10}$) to conform with [24]. The methods marked with † are supervised by Make3D training depth. The depth estimation performance is evaluated with maximum depth set/capped at 70m. We use the center-cropped images as in [17] and resize them to 128×416 for inference.

Method	Supervision		Metrics			
	depth	pose	Abs Rel	Sq Rel	RMSE	RMSE $\log_{10}$
Karsch *et al.* [24]†	✓	-	0.417	4.894	8.172	0.144
Liu *et al.* [34]†	✓	-	0.462	6.625	9.972	0.161
Laina *et al.* [31] †	✓	-	0.198	1.665	5.461	0.082
Godard *et al.* [17]	-	✓	0.443	7.112	8.860	0.142
Zhou *et al.* [56]	-	-	0.392	4.473	8.307	0.194
Ours	-	-	0.378	4.348	7.901	0.183

4.6. Generalization on Make3D

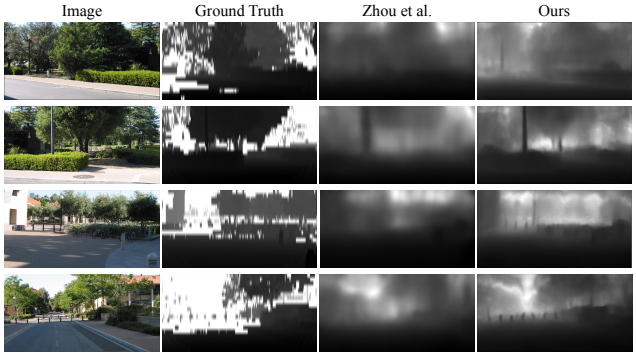


Figure 4. Sample depth predictions on the Make3D dataset. Both our method and SfMlearner [56] are trained on Cityscapes+KITTI datasets.

To illustrate that the proposed method is able to generalize to other datasets unseen in the training, we compare to several supervised/self-supervised methods on the Make3D dataset [44], whose test set contains 134 RGB-D images. We use the same evaluation protocol in [17] which crops the

central region of images for testing. As shown in Table 4, our best model, pre-trained on Cityscapes and finetuned on KITTI raw dataset, achieves reasonable generalization ability and even beats several supervised methods on some metrics. A qualitative comparison is shown in Figure 4.

5. Conclusion

We have presented an unsupervised pose and depth estimation pipeline that absorbs both the geometric principles and learning-based metrics. We emphasize on the consistency issues during the training process and propose novel ingredients to make the result more robust and reliable. Our method achieves the best performance in terms of depth and motion estimation among recent state-of-the-art methods. The current methods are still far from solving the SfM problem in an end-to-end fashion. Further investigations include enforcing consistency across the whole unordered dataset, such as loop closure and bundle adjustment techniques, using the learning-based methods.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, 2016. 6

[2] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *ICCV*, 2009. 2

[3] A. Bartoli and P. Sturm. Non-linear estimation of the fundamental matrix with minimal parameters. *PAMI*, 2004. 4

[4] A. Baumberg. Reliable feature matching across widely separated views. In *CVPR*, 2000. 1

[5] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *ECCV*, 1992. 3

[6] M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. J. Davison. Codeslam-learning a compact, optimisable representation for dense visual slam. In *CVPR*, 2018. 1

[7] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas. Probabilistic data association for semantic slam. In *ICRA*, 2017. 1

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1, 2, 4, 6

[9] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *PAMI*, 2007. 2

[10] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014. 2, 6

[11] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *PAMI*, 2017. 2, 3

[12] J. Engel, T. Schöps, and D. Cremers. Lsd-slam: Large-scale direct monocular slam. In *ECCV*, 2014. 2, 3

[13] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981. 6

[14] Y. Gal. Uncertainty in deep learning. *University of Cambridge*, 2016. 5

[15] R. Garg, V. K. BG, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016. 2, 6

[16] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 1, 2, 4, 6

[17] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 2, 3, 4, 6, 7, 8

[18] V. M. Govindu. Robustness in motion averaging. In *ACCV*, 2006. 2

[19] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1, 4

[20] R. I. Hartley. In defence of the 8-point algorithm. In *ICCV*, 1995. 4

[21] R. I. Hartley. In defense of the eight-point algorithm. *PAMI*, 1997. 1, 6

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 6, 7, 8

[23] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015. 1, 3, 4

[24] K. Karsch, C. Liu, and S. B. Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *PAMI*, 2014. 8

[25] A. Kendall, R. Cipolla, et al. Geometric loss functions for camera pose regression with deep learning. In *CVPR*, 2017. 6

[26] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, 2017. 5

[27] S. J. Kim and M. Pollefeys. Robust radiometric calibration and vignetting correction. *PAMI*, 2008. 2

[28] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[29] M. Klodt and A. Vedaldi. Supervising the new with the old: learning sfm from sfm. In *ECCV*, 2018. 1, 2, 3, 6, 7, 8

[30] K. Konolige and M. Agrawal. Frameslam: From bundle adjustment to real-time visual mapping. *IEEE Transactions on Robotics*, 2008. 2

[31] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016. 8

[32] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o (n) solution to the pnp problem. *IJCV*, 2009. 7

[33] F. Liu, C. Shen, G. Lin, and I. D. Reid. Learning depth from single monocular images using deep convolutional neural fields. *PAMI*, 2016. 2, 6

[34] M. Liu, M. Salzmann, and X. He. Discrete-continuous depth estimation from a single image. In *CVPR*, 2014. 8

[35] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 6

[36] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 2018. 2, 3, 6, 7, 8

[37] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 6

[38] R. Mur-Artal and J. D. Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 2017. 2, 3, 7

[39] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. Dtam: Dense tracking and mapping in real-time. In *ICCV*, 2011. 2

[40] D. Nistér. An efficient solution to the five-point relative pose problem. *PAMI*, 2004. 1

[41] M. Poggi, F. Tosi, and S. Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *3DV*, 2018. 4

[42] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, 2011. 2

[43] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *NIPS*, 2006. 1, 2

[44] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *PAMI*, 2009. 8

[45] T. Shen, S. Zhu, T. Fang, R. Zhang, and L. Quan. Graph-based consistent matching for structure-from-motion. In *ECCV*, 2016. 1

[46] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustmenta modern synthesis. In *International workshop on vision algorithms*, 1999. 2

[47] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. Demon: Depth and motion network for learning monocular stereo. In *CVPR*, 2017. 2

[48] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017. 1, 2, 3

[49] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, 2018. 2, 3, 4, 5, 6

[50] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 3

[51] C. Wu et al. Visualsfm: A visual structure from motion system. 2011. 2, 3, 6

[52] N. Yang, R. Wang, J. Stückler, and D. Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *ECCV*, 2018. 1

[53] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018. 1, 2, 3, 6, 7, 8

[54] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. In *CVPR*, 2010. 1

[55] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, 2018. 2, 3

[56] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 1, 2, 3, 5, 6, 7, 8

[57] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. In *TOG*, 2004. 2