Self-Supervised Learning of Depth and Ego-Motion From Videos by Alternative Training and Geometric Constraints from 3-D to 2-D

Jiaojiao Fang and Guizhong Liu, Member, IEEE

Abstract—Self-supervised learning of depth and ego-motion from unlabeled monocular videos has acquired promising results and drawn extensive attention. Most of the existing methods jointly train the depth and pose networks by photometric consistency of adjacent views based on the principle of structure-from-motion (SFM). However, the coupled relationship of the depth and pose networks based on the scene reprojection seriously influences the learning performance due to the scale ambiguity of image reconstruction-based geometry learning or the error accumulation between the learning-based method and multiview geometry-based method. In this article, we aim to improve the performance of depth and pose estimation without the auxiliary tasks and reduce the influence of the above problems on algorithm performance by alternatively training each task and geometric constraints from 3-D to 2-D. Distinct from jointly training the depth and pose networks, our key idea is to better utilize the mutual dependency between two tasks by alternatively training each network with respective geometric constraints while fixing the other. To make the optimization process easier, the iterative closest point (ICP)-based 3-D structural consistency-embedded epipolar geometric constraints are further introduced into depth and pose networks learning, which can take full advantage of both geometric methods. Then, a log-scale 3-D structural consistency loss is designed to put more emphasis on the smaller depth values during training. Extensive experiments on various benchmark data sets indicate the superiority of our algorithm over the state-of-the-art self-supervised methods.

Index Terms—Epipolar geometry, iterative closest point (ICP), monocular depth estimation, pose estimation, self-supervised learning.

I. INTRODUCTION

DYNAMIC scenes’ 3-D geometric structure understanding, such as depth and camera motion estimation, is a key yet challenging problem in robotics, augmented reality, and autonomous driving scenarios. Obtaining the accurate 3-D geometric structure of a scene and the camera ego-motion in the real world are essential for motion planning and decision making. Traditional handcrafted feature and multiview geometry-based methods can only obtain the sparse reconstructions. Recently, with the development of deep learning, the supervised learning methods for monocular depth and camera pose estimation are proposed and achieve great advance. But these methods require the use of additional expensive sensors to obtain densely annotated ground-truth information and precise calibration, which are costly and time consuming. Thus, recent works seek to obtain the 3-D scene geometric information and camera ego-motion in a self-supervised manner from either stereo image pairs [10], [32] or video sequences [33].

The self-supervised learning framework that jointly optimizes the camera pose and monocular depth has caught the attention of academics as it much less depends on the ground-truth labels. Previous methods mainly rely on minimizing the photometric consistency loss among adjacent views by reprojecting the backprojected 3-D points of the target view onto the source views, which may contain much systematic error in realistic scenes due to the specular region, repetitive textures, reflective surfaces, and occlusions. Thus, some later works try to explicitly measure the inferred geometry of the whole scene by the multiview 3-D structural alignment [21], [39] or the epipolar geometry [3], [40], which are of great importance for self-supervised learning of monocular depth and camera pose. Although these frameworks have achieved excellent improvement, the following problems still exist.

1) Due to the coupled relationship between depth estimation and pose estimation, the contribution of each task toward the optimization objective is unclear by a united loss function, and it is difficult to ensure that each task is optimized toward a globally optimal solution. Thus, more feasible optimization manner is needed to obtain more reliable results.

2) The smaller depth values may contain richer information and be more important for 3-D scene reconstruction than the larger depth values which have higher fault tolerance in a real application, while the traditional L2-norm in the linear space that treats each element equally would lead to overly emphasize the errors of larger depth values during training.

3) The iterative closest point (ICP)-based geometric constraints can simultaneously constrain the depth and pose network learning and are less affected by the scale ambiguity problem, but it is still subject to error accumulation and is in short of exact linear mathematical
constraint relations, such as epipolar geometry. While the epipolar geometry-based constraints take no account of the multiview 3-D structural consistency. Existing methods consider the 3-D structural consistency loss and epipolar geometric constraints separately, and have not directly established the connection between them.

Inspired by the existing excellent work, in this article, we propose a self-supervised depth and pose learning method to tackle the above issues by incorporating alternative training, geometric constraints based on the epipolar geometry and the 3-D ICP point-clouds registration interacted with each other, and the log-scale 3-D geometric constraint for multiview structural consistency.

Our main contributions are as follows.

Alternatively Training the Depth and Pose Networks With Respective Geometric Constraints: Inspired by the two-step optimization process of the ICP point-clouds registration method, to ensure each network is explicitly optimized toward the direction of the global optimal solution and better utilize the mutual dependency of the depth and pose estimation, alternative training is proposed to train the two networks with respective geometric constraints for aligning depth with pose and aligning pose with depth by turns. In this way, we can optimize one task after another that has already been optimized by one step. The effectiveness of the alternative training with different geometric constraints in the self-supervised depth and pose learning framework is verified.

ICP-Based 3-D Structural Consistency Embedded Epipolar Geometric Constraints: Multiview geometric consistency is vitally important for self-supervised learning of monocular depth and camera pose. The ICP-based geometric constraints can explicitly measure the inferred 3-D structure of the whole scene and are less affected by the scale-ambiguity problem which means that the estimated structure is not in physical scale, but directly optimizing the outputs of the networks toward the geometry-based method would lead to error accumulation and additional exact linear mathematical relation is still needed in self-supervised learning. In this article, to simplify the optimization process and take full advantage of both geometric constraints, we propose a more direct way to construct the epipolar geometric constraint in consideration of the 3-D structural consistency. By establishing a connection between the ICP registration and epipolar geometry, we can find better point correspondences and a more direct way to use the multiview geometry-based method. Furthermore, we also verify the effectiveness of the properties of epipolar geometry, namely, the low rankness and self expression in union-of-subspaces, which could serve as a global regularization and deal with the moving objects, for depth and pose learning.

Log-scale 3-D Structural Consistency Measurement: It would be better to avoid excessive optimization for the relatively larger errors caused by the larger depth values during training, which is with greater error tolerance in practical applications. To this end, we use the log-scale mean squared error to measure the 3-D structural consistency; thus, the importance of the smaller depth value would be improved in the training process.

We show that the proposed geometric constraints can be explicitly incorporated into the training process without breaking the simplicity of inference. The proposed framework is extensively evaluated on various data sets and has achieved state-of-the-art performance. The related work is introduced in Section II. The details of the proposed methods are described in Section III. The experimental evaluations are given in Section IV. Finally, the conclusion is drawn in Section V.

II. RELATED WORK

Monocular depth estimation based on deep convolutional neural networks (CNNs) from stereo image pairs or video sequences had achieved great advance. These methods mainly fell into two sorts: 1) the supervised learning methods and 2) the self-supervised learning methods.

A. Supervised Monocular Depth Estimation

Supervised monocular depth estimation refers to the problem set that trains deep neural networks to obtain the 3-D structure of a scene via vast ground-truth data. Due to the superiority of deep learning and the availability of ground-truth data, the supervised learning frameworks had acquired advanced accuracy. Eigen et al. [6] first proposed the depth estimation from a single image by training a network on sparse labels provided by LiDAR scans. Liu et al. [20] used a CNN combined with the conditional random field to learn monocular depth. Karsch et al. [16] proposed a technique that automatically generates plausible depth maps from videos using nonparametric depth sampling. Several works tried to further improve the accuracy of supervised depth estimation by using more robust losses [1], [19]. Fu et al. [7] introduced a spacing-increasing discretization (SID) strategy to discretize depth value and recast depth network learning as an ordinal regression problem for supervised depth estimation. But the superior performance of these supervised methods usually relied on high quality and pixel-aligned ground-truth data for training, which is challenging and expensive to gain in various real-world environments.

B. Self-Supervised Learning of Depth and Pose

Recently, many self-supervised methods for monocular depth and camera pose estimation had been proposed by utilizing the multiview geometric information from multiple cameras or video sequences based on the structure-from-motion (SFM) [29], [40]–[45]. Here, we focus on the most related self-supervised monocular depth and pose estimation from videos.

Zhou et al. [33] first presented a generalized pipeline of self-supervised depth and ego-motion estimation from unlabeled monocular videos with the static scene assumption. Ever since, great advancement has been made to improve the performance of self-supervised learning with regard to powerful loss functions, moving objects discrimination, occlusions removal as well as network architectures. For more robust self-supervision signals, Mahjourian et al. [21] introduced ICP-based differentiable 3-D structural consistency loss as an extra supervisory signal, which directly constrains the outputs of the depth and
pose networks toward the geometry-based method and less affected by the scale ambiguity problem. Bian et al. [39] proposed a relative 3-D geometric consistency loss for scale-consistent predictions and a self-discovered mask for handling moving objects and occlusions. Shen et al. [15], [40] incorporated the epipolar geometry for more robust geometric constraints, while the precomputed correspondences of feature points were needed. To utilize a simple and effective model, Godard et al. [11] proposed an effective minimum photometric loss and a binary automasking method to deal with the occlusions and exclude the invalid regions. Guizilini et al. [12] proposed a novel network architecture for self-supervised monocular depth estimation, called PackNet, to obtain feature representation with more detailed information. Another kind of method leveraged auxiliary tasks to facilitate geometry learning, Yin et al. [31] added a refinement module to the self-supervised depth and pose learning framework for estimating the residual flow to narrow the gap between scene flow and optical flow caused by moving objects and used the forward–backward consistency to account for the moving regions and occlusions. Ranjan et al. [27] jointly learned the motion segmentation, optical flow, camera motion, and monocular depth to obtain the complete geometric structure and moving objects. To obtain semantic-level predictions, Chen et al. [2] proposed a method integrating the geometric information and the semantic information to model the geometric structure of objects for scene understanding. Guizilini et al. [13] proposed a novel architecture using a pretrained semantic segmentation network to guide the learning of geometric representation in the self-supervised learning framework. To learn the reason for the uncertainty of the predicted depth, Poggi et al. [25] focused on the uncertainty estimation for self-supervised learning of monocular depth and showed how this practice improved the accuracy of depth estimation.

C. Epipolar Geometry for Self-Supervised Learning

The epipolar geometric constraints were popular for self-supervised optical flow estimation. Valgaerts et al. [34] introduced a model to simultaneously estimate the fundamental matrix and the optical flow. Wedel et al. [35] used a fundamental matrix as a weak constraint for the optical flow training. These methods, however, assumed that the scene was mostly rigid, and treated the dynamic parts as outliers. Garg et al. [23] used the subspace constraint as a regularization term for multiframe optical flow estimation. Zhong et al. [17] proposed a low-rank constraint as well as a union-of-subspaces constraint based on the epipolar geometry for self-supervised optical flow learning. While Chen et al. [3] captured multiple geometric constraints for relating the optical flow, depth, camera pose, and intrinsic parameters from monocular videos, and used epipolar geometry as a verification of the optical flow correspondences.

This work is based on the previous good practices, with the major distinction that we explore a new training policy to facilitate the training procedure of the depth and pose networks with a log-scale 3-D structural consistency loss and the epipolar geometric constraint embedded with ICP point-clouds registration.

III. PROPOSED METHOD

In this section, the proposed method will be described, including the log-scale 3-D structural consistency loss, the geometric constraints from 3-D to 2-D, the alternative training with different geometric constraints, as well as the properties of the epipolar geometry as regularizations. Fig. 1 illustrates an overview of the proposed method. This section starts with an introduction to the problem formulation.

A. Problem Formulation

The problem of the self-supervised learning of the depth and pose networks from monocular video sequences can be formulated as follows. Given a target RGB image $I_t \in \mathbb{R}^{H \times W \times 3}$ and the source images $I_s \in \mathbb{R}^{H \times W \times 3}$ where $s \in \{t-1, t+1\}$, collected by a potentially moving camera with an intrinsic matrix $K$. For simplicity of notation, here we assume that the camera intrinsic matrixes of all the views are the same. The 3-D rigid
transformation of a camera’s ego-motion can be represented as a matrix $T_{s \rightarrow t} = [R_{s \rightarrow t}, t_{s \rightarrow t}]$, where $R_{s \rightarrow t}$ is the rotation matrix and $t_{s \rightarrow t}$ is the translation vector from time $s$ to $t$, which would be estimated by a pose network. Assuming a pixel $p_t = [u_t, v_t]$ in the target frame $I_t$ and the depth value $D_0(p_t)$ of point $p_t$ would be estimated by a depth network, then the corresponding 3-D locations $Q_t(p_t)$ in the camera’s world coordinates system at time $t$ is the backprojection of the pixel point $p_t$

$$Q_t(p_t) = \begin{bmatrix} X_t \\ Y_t \\ Z_t \\ 1 \end{bmatrix} = D_0(p_t)^{-1} \begin{bmatrix} u_t \\ v_t \\ 1 \end{bmatrix}$$ (1)

where $D_0$ represents the output of the depth network and $\theta$ refers to the parameters of the depth network. Similarly, the backprojected 3-D point cloud $Q_s(p_s)$ of a point $p_s$ in the source views can be computed by (1) with the estimated depth $D_s(p_s)$.

Assuming $Q_t(p_t)$ is transformed rigidly from time $t$ to $s$, then its corresponding point $p'_s$ in the source view $I_s$ can be calculated by reprojection of the backprojected 3-D points $Q_t(p_t)$

$$[p'_s, 1]^T = D(p_t)^{-1} KT_{s \rightarrow t} Q_t(p_t)$$ (2)

where $D(p_t)$ is the depth of the transformed 3-D point in the camera’s world coordinates at time $s$ calculated by the transformation of the backprojected 3-D point $Q_t(p_t)$ with the matrix $T_{s \rightarrow t}$. Note that the transformation applied to the 3-D point is the inverse of the camera movement from time $s$ to $t$. Thus, the reconstructed target view can be easily constructed from the source views by the displacement in (2) to compute the commonly used photometric loss between the reconstructed image and the real captured one. Following the existing work [10], the bilinear sampling [37] is used for the noninteger reprojection coordinate points to reconstruct the target image.

In this article, the minimum photometric error over all source views [11] is used as the photometric loss $L_{ph}$

$$L_{ph} = \min_s [ph(I_t, I_{s \rightarrow t})]$$ (3)

where $I_{s \rightarrow t}$ is the reconstructed target image from the source images by the reprojection relations in (2), and the photometric error $ph(\cdot, \cdot)$ is the convex combination of $L1$-norm and structural similarity (SSIM) [39] over all pixels

$$ph(\cdot, \cdot) = a(1 - SSIM(\cdot, \cdot))/2 + (1 - a) \| \cdot - \cdot \|_1$$ (4)

where $a = 0.85$ is the combination coefficient. As in [10], the edge-aware smoothness loss $L_s$ is used to constrain the estimated depth maps

$$L_s = |\partial_x D_0(p_t)^* e^{-|\partial_x I_t|} + |\partial_y D_0(p_t)^* e^{-|\partial_y I_t|}$$ (5)

where $D_0(p_t)^* = D_0(p_t)/\|D_0(p_t)\|$ is the mean-normalized inverse depth to deal with the shrinking of the predicted depth values [11], $\partial_x I_t$ and $\partial_y I_t$ are the gradients of the image, $\partial_x D_0(p_t)^*$ and $\partial_y D_0(p_t)^*$ are the horizontal and vertical gradients of estimated depth map.

A per-pixel binary mask $\mu$ proposed by Godard et al. [11] is used to exclude the relatively stationary regions with the camera which are harmful to the photometric loss. Thus, the method proposed in [11] that trained on the monocular video sequences by the losses $\mu L_{ph} + L_s$ is used as the baseline of our method.

B. Log-Scale 3-D Structural Consistency in Adjacent Views

The geometric structure of a scene obtained by the estimated depth of adjacent multiviews should be consistent [3], [21]. Thus, enforcing the 3-D structural consistency to consider the entire structure of a scene is necessary for the self-supervised learning of depth and camera pose, which is highly related to the 3-D point-cloud registration process. In this article, to improve the importance of the smaller depth values and to avoid excessive optimization for the larger depth values during training, a log-scale 3-D geometric consistency loss is introduced to penalize the structural variations of the estimated depth maps for the same scene in multiple views and enforce the scale consistency of the self-supervised learning of monocular depth.

Given the reprojection relation between a pixel $p_t$ in the target image, and its corresponding point $p'_s$ in the source images by (2), the structural inconsistency between these two points $Q_s(p_s)$ and $Q_t(p_t)$ in the world coordinate system can be penalized by the Euclidean distance in the log-scale space instead of in the linear space by the bilinear sampling [37]. In consideration of the occlusions, inspired by the photometric loss proposed in [11], we use the minimum error instead of the average error over all source views as the 3-D geometric loss

$$L_{3-D} = \min_s (\log(Q_s(p_s') - \log(T_{s \rightarrow t}Q_t(p_t)))$$ (6)

In this way, the gradients of the loss in (6) in regard to the error of the larger depth value would be smaller during training. Similar to the geometric consistency loss in [39] that computes the relative error instead of the absolute error, the point clouds are first mapped into the log scale before computing the loss function to focus more on the errors of the smaller depth values. The experiments prove the effectiveness of the simple logarithm operation.

C. Epipolar Geometry Embedded With 3-D Structural Consistency

The epipolar geometric constraint is less affected by the depth estimation as it does not concern the depth in its formulation. Existing methods use either the precomputed sparse feature matching [15] or the optical flow [3] combined with the estimated camera pose to construct the epipolar geometric constraint on the image planes. The ICP point-cloud registration based method is proposed by Majhouri et al. [21] to compute the transformation and matching error between two sets of point clouds and directly constrains the outputs of the networks toward the computed result of the ICP-based method.

In this article, instead of using the optical flow, the sparse local feature descriptors matching, or the reprojection of the backprojected 3-D structure from the estimated depth of an image to build the pixel point correspondences between
adjacent views, we use the precise ICP-based point-cloud registration methods [46] to establish the dense correspondences of the two groups of point clouds $Q_t = (Q_t(p_1), \ldots, Q_t(p_N))$ and $Q_s = (Q_s(p_1), \ldots, Q_s(p_N))$ from adjacent views, where $N$ is the total number of points. The ICP-based method alternatively computes correspondences between two groups of 3-D point clouds by the closest point searching and a best-fit transformation between two sets of point clouds with the given corresponding points. The next iteration recomputes the minimum distance-based correspondences with the transformation of the previous iteration. Thus, the optimization objectives of the $n$th iteration are as follows:

$$
\min_{s'' \in \{1, \ldots, N\}} \left\| Q_s(p_{s''}) - T_{s \rightarrow t}^{-1} Q_t(p_t) \right\|, \quad t \in \{1, \ldots, N\} 
$$

(7)

where $p_t$ and $p_{s''}$ denote the coordinates of the corresponding points of adjacent views found by the closest distances in the two groups of point clouds, $T_{s \rightarrow t}^{-1}$ represents the estimated pose of the $(n - 1)$th iteration. With the given correspondences of the point clouds, the best-fitted transformation between two adjacent views can be optimized by

$$
\min_{T_{s \rightarrow t}} \sum_{t=1}^{N} \left\| Q_s(p_{s''}) - T_{s \rightarrow t}^{-1} Q_t(p_t) \right\|^2. 
$$

(8)

A transformation is optimized to minimize the distances between the corresponding point pairs. The above-mentioned two-steps point-cloud registration method has been used to construct a differentiable 3-D loss for constraining the geometry consistency among adjacent views [21]. The differentiable 3-D loss uses both the computed transformation and the final residual registration error after ICP’s minimization which is an indirect optimization objective and hard to achieve the global optimal solution due to the error accumulation of the deep learning and the traditional optimization method. In this article, we combine the ICP and the epipolar geometry to constrain the self-supervised learning of depth and pose in a more feasible way. To better utilize the mutual dependency of the depth and pose networks and make the optimization process easier, we incorporate the epipolar geometry to constrain the pose network learning while taking the multiviews 3-D geometric structural consistency into account, which could be obtained by the optimized depth network. Instead of directly constructing a loss between the transformations computed by the ICP and the outputs of the pose network, we obtain a corresponding between coordinates of adjacent views by the two groups of 3-D point-cloud registration in (7) and construct an epipolar geometry with the 3-D structural consistency embedded correspondences. The optimization aims to transfer the consistency of depth to pose by the epipolar geometry.

The correspondences between pixel coordinates of adjacent views based on the ICP alignment are

$$
p_{s''} = p_t + [\delta_x, \delta_y] 
$$

(9)

where $\delta_x$ and $\delta_y$ are the displacements in the horizontal and vertical directions, respectively. Assuming a pinhole imaging model, the correspondences $p_t$ and $p_{s''}$ of adjacent views should satisfy the epipolar geometry with a fundamental matrix $F = K^{-1} [I_{N \times d}] \cdot R_{s \rightarrow t} \cdot K^{-1}$, which are the cross product of the rotation matrix and the translation vector multiplied with the inverse of the camera intrinsic matrix

$$
\arg \min_{T_{s \rightarrow t}} \sum_{t=1}^{N} \left\| Q_s(p_{s''}) - T_{s \rightarrow t}^{-1} Q_t(p_t) \right\|^2, \quad t \in \{1, \ldots, N\} 
$$

(10)

where $[\cdot]_x$ is the skew-symmetric matrix of a 3-D vector [30]. The correspondences in (9) are integer values with no need for bilinear sampling. The roles of this loss can be twofold, one is a validation for the multiviews 3-D structural consistency by the exact linear mathematical relation, and the other is for the pose network optimization. In this way, the error accumulation between the geometry-based method and the learning-based method can be alleviated to some extent. The main difference between the existing multiview geometry-based method and the proposed method is establishing the connection between the ICP-based point-cloud registration and epipolar geometry.

The basic epipolar geometry is an overconstrained formulation, which is not robust to outliers or noises. To improve the robustness of the epipolar geometric constraint, in this article we incorporate the low-rankness loss $L_{lr}$ and self-expression in union-of-subspaces proposed in [17], into the self-supervised depth and pose learning.

To introduce the properties of the epipolar geometry, we rewrite the epipolar geometric constraint in (10) as

$$
f^T \text{vec} \left( [p'_{t}, 1]^T [p_t, 1] \right) = 0 
$$

(11)

where $f$ is the vectorized fundamental matrix $F$ along the column direction, and $e_t = \text{vec}((p'_{t}, 1)^T [p_t, 1])$ is the vectorized product of the homogeneous coordinate vectors of the two points, which is a 9-D column vector lying in the epipolar subspace [4]. Then we can form a matrix $E = [e_1, \ldots, e_N]$ by all the vectors $e_t$, where $t \in \{1, \ldots, N\}$, $N \gg 9$ is the number of pixel points. Thus, the low-rankness loss can be formulated as

$$
L_{lr} = \| E \|_s 
$$

(12)

where $\| \cdot \|_s$ is the nuclear norm, which can be optimized by the singular value decomposition (SVD). By using this constraint, we do not have to explicitly compute the fundamental matrix. So it can be applied in the degenerated cases where a fundamental matrix is unknown. Although the low-rankness constraint is too loose, it still can improve the generalization ability to some extent [17].

To deal with moving objects in the scene, another constraint called the self-expression in union-of-subspaces is further introduced. This constraint implies that all vectors lying in the union-of-subspaces can be characterized by the self-expression property [17], i.e., each vector can be sparsely represented as a linear combination of the other vectors. The coefficients would be nonzero among the vectors in the same subspace while keeping zero among the vectors in others. As in [17],
the mathematical expression is as follows:

\[
L_{\text{sub}} = \| (I + \lambda E^T E)^{-1} \lambda E^T E \|_F^2 \\
+ \lambda / 2 \| E (I + \lambda E^T E)^{-1} \lambda E^T - E \|_F^2
\] (13)

where \( C = (I + \lambda E^T E)^{-1} \lambda E^T E \) is the coefficients matrix of the self-expression in the union-of-subspaces, \( \lambda = 0.05 \) is the relaxation factor. In consideration of the GPU memory and the computational efficiency, we randomly sample 2000 point pairs to compute this loss, similar to in [17]. Even though these epipolar subspaces would be disjoint, they can still serve as a global regularization. No specific scale alignment method is used to solve the scale ambiguity problem in the proposed method, but the combination of the ICP point-cloud registration and epipolar geometry can minimize the negative impact of the scale ambiguity on algorithm performance.

D. Alternative Training by Different Optimization Objectives

In this section, we describe the proposed training policy and losses used in the self-supervised learning of depth and pose networks. We embed the traditional two-steps optimization idea into the self-supervised depth and camera pose learning framework, which aligns the 3-D structure of the adjacent views with the estimated pose and learns the transformation between adjacent views by the pose network with the correspondences from two groups of 3-D point-cloud registration. To ensure that each of the networks is directly optimized toward the gradient descent direction and obtain the global optimal solution, we use an alternative training policy with different geometric constraints, the log-scale 3-D geometric structural consistency loss, the 3-D structural consistency embedded epipolar geometric constraint, and the properties of the epipolar geometry as regularizations for depth and pose networks learning, respectively. For the first trained network, it is similar to the joint training, while another network will be trained on top of an already optimized network.

**Pose Network Optimization:** As the epipolar geometric constraint is less influenced by the depth estimation, it is a better way for camera pose learning. Thus, in this article, we use the epipolar geometric constraint in (10) for the pose network optimization. To deal with the occlusions and out of views, we still use the minimum error over all source views to construct the epipolar geometric loss \( L_{\text{ep}} \)

\[
L_{\text{ep}} = \min \left( p_{p'}^c, \| F[p_r, 1]^T \right) .
\] (14)

The loss \( L_{\text{ep}} \) can be easily obtained by computing the distance map from each pixel to its corresponding epipolar lines, as in [40]. As the camera pose estimation is more vulnerable to the moving objects in the static scene, the properties of self-expression in the union-of-subspaces \( L_{\text{sub}} \) in (13) is further used to regularize the pose network optimization by the relation of the scene reprojected in (2). Hence, the consistency between the two kinds of correspondences \( p_{p'}^c \) and \( p_{p''}^c \) would be guaranteed. Thus, the total loss of the pose network training is as follows:

\[
L_{\text{pose}} = \mu L_{\text{ph}} + \lambda_c L_{\text{ep}} + \lambda_{su} L_{\text{sub}}
\] (15)

where \( \lambda_c \) and \( \lambda_{su} \) are the weights for the different losses. When optimizing one of the depth and pose networks, the outputs of the other network are treated as known volume.

**Depth Network Optimization:** The 3-D structural consistency loss in (6) is more concentrated on the whole structure of a scene and would be a better way to directly optimize the depth network. In consideration of the effectiveness of the low-rankness constraints \( L_{\text{lr}} \) reported in [17], we also use it to regularize the depth network learning as it has little to do with pose estimation. Thus, the total loss for the depth network training can be expressed as

\[
L_{\text{depth}} = \mu L_{\text{ph}} + \lambda_s L_s + \lambda_{3d} L_{3D} + \lambda_l L_{\text{lr}}
\] (16)

where \( \lambda_s \), \( \lambda_{3d} \), and \( \lambda_l \) are the weights for the different losses. These losses average over all the pixels, scales, and batches to train the networks in an end-to-end manner. The respective losses are used to alternatively train one network while fixing the other, which means that only the feedforward propagation is performed for the other network without backpropagation. The role of the photometric loss here is as the raw data term and as a global verification to balance the photometric and geometric errors when the initial values of the depth and pose are not accurate enough. We only train one network at a time while the other network is optimized based on the already optimized one, and the first trained network is similar to the joint training method at the beginning of the training.

IV. EXPERIMENTS

In this section, we evaluate the performance of the proposed model and compare it with the published state-of-the-art self-supervised methods on the KITTI 2015 stereo data set [9]. We also use the Make3D data set [26] to evaluate the generalization ability on the cross data set.

A. Training Data Sets

**KITTI Raw Data Set:** We mainly used the raw KITTI data set [9] for training and evaluation. The data set contains 42,382 rectified stereo pairs from 61 scenes, with a typical resolution of 1242 \( \times \) 375 pixels. We trained our model with the Eigen split [6] that excluded 679 images for testing and removed static frames following Zhou et al. [33]. This led to a total of 44,234 monocular frames, where 39,810 for training and 4424 for validation. To facilitate the training process and provide a fair evaluation, all the input images were resized to 640 \( \times \) 192.

**KITTI Visual Odometry (VO) Data Set:** The KITTI odometry data set contains 11 driving sequences with ground-truth labels and 11 sequences without ground-truth labels. As in the standard setting, we used sequences 00–08 for training and sequences 09 and 10 for testing.

**Cityscapes Data Set:** Since starting from a pretrained model boosted the performance [33], we also tried to pretrain the model on the Cityscapes [5] data set, where 88,084 images for training and 9659 images for validation.
TABLE I
RESULTS OF COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE KITTI DATA SET [9] WITH THE SPLIT OF EIGEN ET AL. [6], WHERE THE ERROR METRIC IS LOWER THE BETTER, AND THE ACCURACY METRIC IS HIGHER THE BETTER. THE BEST RESULTS ARE IN BOLD; THE SECOND BEST IS UNDERLINED. “K” REPRESENTS KITTI RAW DATA SET AND “CS” REPRESENTS CITYSCAPES TRAINING DATA SET. M REFERS TO METHODS THAT TRAIN USING MONOCULAR SEQUENCES, S REFERS TO METHODS THAT TRAIN USING STEREO PAIRS, D REFERS TO METHODS THAT USE GROUND-TRUTH DEPTH SUPERVISION, “SEM” REFERS TO METHODS THAT INCLUDE SEMANTIC INFORMATION. “w/o p-p” REFERS TO THE RESULTS OBTAINED WITHOUT POSTPROCESSING.

| Method       | Supervision | Dataset | Error metric  | Accuracy metric |
|--------------|-------------|---------|---------------|-----------------|
|              |             |         | Abs Rel | Sq Rel | RMSE  | RMSE log | δ < 1.25 | δ < 1.25^2 | δ < 1.25^3 |
| Eigen et al. Fine [6] | D  | K       | 0.203  | 1.548  | 6.307  | 0.282   | 0.702  | 0.890  | 0.958  |
| Liu et al. [20]       | D  | K       | 0.201  | 1.584  | 6.471  | 0.273   | 0.678  | 0.898  | 0.967  |
| Zhou et al. [33]      | M  | CS+K    | 0.198  | 1.836  | 5.665  | 0.275   | 0.718  | 0.901  | 0.960  |
| Godard et al. [10]    | S  | CS+K    | 0.141  | 1.186  | 5.677  | 0.238   | 0.809  | 0.928  | 0.969  |
| Vidi2dep. [21]        | M  | CS+K    | 0.159  | 1.213  | 5.912  | 0.243   | 0.784  | 0.923  | 0.970  |
| Yin et al. [31]       | M  | CS+K    | 0.149  | 1.060  | 5.677  | 0.226   | 0.796  | 0.935  | 0.975  |
| Shen et al. [15]      | M  | CS+K    | 0.139  | 0.964  | 5.309  | 0.215   | 0.818  | 0.941  | 0.977  |
| Zhan et al. [32]      | S  | K       | 0.135  | 1.135  | 5.585  | 0.229   | 0.820  | 0.933  | 0.971  |
| Monodepth2 [11]       | M  | K       | 0.112  | 0.851  | 4.754  | 0.190   | 0.881  | 0.960  | 0.981  |
| Monodepth2, w/o p-p [11] | M  | K       | 0.115  | 0.903  | 4.863  | 0.193   | 0.877  | 0.959  | 0.981  |
| Chen et al. [3]       | M  | K       | 0.135  | 1.070  | 5.230  | 0.210   | 0.841  | 0.948  | 0.980  |
| Guizilini et al. [13] | M+Sem | CS+K   | 0.117  | 0.854  | 4.714  | 0.191   | 0.873  | 0.963  | 0.981  |
| Xue et al. [36]       | M  | K       | 0.113  | 0.864  | 4.812  | 0.191   | 0.877  | 0.960  | 0.981  |
| Klinger et al. [44]   | M+Sem | (CS)+K | 0.113  | 0.835  | 4.693  | 0.191   | 0.879  | 0.961  | 0.981  |
| Ours w/o p-p          | M  | K       | 0.112  | 0.835  | 4.748  | 0.189   | 0.878  | 0.960  | 0.981  |
| Ours                  | M  | CS+K    | 0.110  | 0.793  | 4.674  | 0.186   | 0.884  | 0.963  | 0.982  |
| Ours                  | M  | K       | 0.110  | 0.806  | 4.681  | 0.187   | 0.881  | 0.961  | 0.982  |

Depth capped at 80m

| Method       | Supervision | Dataset | Error metric  | Accuracy metric |
|--------------|-------------|---------|---------------|-----------------|
|              |             |         | Abs Rel | Sq Rel | RMSE  | RMSE log | δ < 1.25 | δ < 1.25^2 | δ < 1.25^3 |
| Vidi2dep. [21] | M  | CS+K    | 0.151  | 0.949  | 4.838  | 0.227   | 0.802  | 0.935  | 0.974  |
| Yin et al. [31] | M  | CS+K    | 0.147  | 0.936  | 4.384  | 0.218   | 0.810  | 0.941  | 0.977  |
| Shen et al. [15] | M  | K       | 0.133  | 0.778  | 4.069  | 0.207   | 0.834  | 0.947  | 0.978  |
| Godard et al. [11] | M  | K       | 0.108  | 0.661  | 3.659  | 0.181   | 0.890  | 0.965  | 0.983  |
| Ours          | M  | K       | 0.105  | 0.616  | 3.602  | 0.179   | 0.890  | 0.966  | 0.984  |

Depth capped at 50m

B. Implementation Details

Depth Network: The depth network was a fully convolutional encoder-decoder structure with skip connections, which was similar to the DispNetS [22]. The ResNet18 [14] was used as the encoder if there was no otherwise specification. The decoder module had five deconvolution layers. Networks had the outputs at four different spatial resolutions. The lower resolution depth maps were up-sampled to the input resolution for the loss function calculation as in [11].

Pose Network: The pose network took two adjacent views as input and outputted the relative motions between the target view and source views. The network consisted of seven convolutional layers followed by a 1 × 1 convolutional layer with six output channels, corresponding to rotation angles about three coordinate axes and translations in three coordinates.

Parameters Setting and Processing: For all the experiments, the weights of the different loss components were empirically set as $\lambda_s = 0.001$, $\lambda_r = 0.002$, $\lambda_d = 0.02$, $\lambda_i = 0.001$ and $\lambda_{su} = 0.0001$, where the value of $\lambda_s$ is the same as in [11] for a fair comparison. The proposed model was trained for 20 epochs using the Adam [18] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, Gaussian random initialization, ResNet18 with pretrained weight on the ImageNet [28], and minibatch size of 4. The learning rate was originally set to 0.0001 and halved after every ten epochs until the end. We used the same data augmentations as in [11]. For disparity maps, we followed a similar postprocessing technique as in [11] and capped depth values at 80 m as per standard practice during evaluation [10]. The per-image median ground-truth scaling introduced by [33] was used in the report results. For the networks trained on a pretrained model, the depth network was first trained, and then the pose network was trained based on the optimized depth network by the same minibatch during each iteration.

Equipment and Efficiency: The algorithm was deployed in the PyTorch [24] framework which was compiled with CUDA 9.0 and CuDNN 7.0 on a computer with an Intel Xeon E5-1660v4 HP-Z440 8-Core 3.2-GHz CPU and a Titan Xp GPU. With a single Titan Xp GPU, the network took almost 3.1 h per epoch compared with 0.8 h of the baseline method. While the runtime of our model for testing was the same as the baseline.

If there was no additional specification, the models were trained by the conditions mentioned above.

C. Main Results of Depth and Pose Evaluations

Depth Evaluation: The evaluation of depth estimation followed the previous works [11], [21], [33]. Here, we provided a comparison of the depth estimation with the state-of-the-art self-supervised methods [3], [10], [11], [13], [15], [21], [32], [33], [36] and the classical supervised methods [6], [20]. To be fair for all methods, we used the same crop manner as [33] and evaluated the prediction with the same resolution as the input image. The measure criterion conformed to the one used in [11]. As shown in Table I, with the same underlying network structure, the proposed method outperformed state-of-the-art methods by a large margin. The network first pretrained on the larger Cityscapes data set [5], and then fine-tuned on the KITTI data set [9], which would result in slight
Fig. 2. Qualitative results of our proposed architecture on the KITTI data set with the Eigen split [6]. The columns from left to right show, respectively, input images, the state-of-the-art predicted depth maps (Godard et al. [11]; Xue et al. [36]), and the depth maps obtained by our proposed architecture. Our method recovers more subtle details such as trees, trunks, and advertising boards and has clearer contours and better internal smoothness.

performance improvement. The final postprocessing step led to an accuracy increase and fewer visual artifacts at the expense of doubling the test time. To prove the performance of the close-range depth estimation, we also provided separate results for a depth capped at 50 m, which was also shown the advantage of our method. Qualitative results compared with the predictions of Godard et al. [11] and Xue et al. [36]) could be seen in Fig. 2. It was shown that our method could reduce artifacts in low-texture regions of the image and improve the accuracy of close-range objects. The performance improvements mainly owe to the 3-D geometric consistency is further verified by the 2-D geometric constraints on the image planes, by alternative aligning depth with pose and aligning pose with depth can correct each other and the minimum loss instead of average loss is also useful for geometric constraints.

Pose Evaluation: Although our method mainly concentrated on better depth estimation, we also evaluated the performance of relative pose estimation with competing methods on the official KITTI odometry benchmark using the absolute trajectory error (ATE) metric over $N$-frame snippets ($N = 3$ or $5$), as in [11]. The pose estimation results in Table II showed the improvement over existing methods. We had observed that with the epipolar geometric constraints, the result of pose estimation would be notably improved, which is consistent with the report in [15].

Generalization Ability on the Make3D Data Set: To illustrate the generalization ability of the proposed model on other unseen data sets during training, we evaluated the depth estimation network on the Make3D data set using the same assessment criteria as in [10], which was only trained on the KITTI data set. As shown in Table III, our methods achieved competitive advantages over the self-supervised methods. Furthermore, we observed that the low-rankness constraint $L_{lr}$ could improve the generalization ability of the depth estimation. Qualitative results were shown in Fig. 3. Despite the dissimilarities of the data sets, both in contents and camera parameters, we still achieved reasonable results.

D. KITTI Ablation Study

Performance of Different Losses: To analyze the individual impact of each loss, we provided an ablation study over different combinations of losses. Models were trained only on
TABLE IV

| Loss Configuration and Alternative Training Manner | Error metric | Accuracy metric |
|---|---|---|
| | | | Abs Rel | Sq Rel | RMSE | RMSE log | δ < 1.25 | δ < 1.25^2 | δ < 1.25^3 |
| Baseline | L_{3D} | L_{lp} | L_{lp} | L_{sub} | Alternate | 0.115 | 0.903 | 4.863 | 0.193 | 0.877 | 0.959 | 0.981 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.114 | 0.885 | 4.845 | 0.192 | 0.877 | 0.960 | 0.980 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.115 | 0.896 | 4.844 | 0.193 | 0.876 | 0.959 | 0.981 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.114 | 0.884 | 4.834 | 0.192 | 0.877 | 0.959 | 0.982 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.113 | 0.879 | 4.828 | 0.191 | 0.876 | 0.960 | 0.981 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.113 | 0.872 | 4.825 | 0.189 | 0.882 | 0.960 | 0.981 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.114 | 0.870 | 4.812 | 0.192 | 0.876 | 0.960 | 0.981 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.113 | 0.868 | 4.809 | 0.192 | 0.877 | 0.959 | 0.982 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.112 | 0.835 | 4.748 | 0.189 | 0.878 | 0.960 | 0.982 |

Table IV: Ablation Studies on Depth Estimation. Evaluation of Different Training Loss Configurations and Training Policies. All Models Are Either Solely Trained on the KITTI Raw Data Set Without Pretraining on Cityscapes [5] and Postprocessing [10]. The Depth Estimation Performance Is Evaluated With Maximum Depth Capped at 80 m, Where the Error Metric Is Lower the Better, and the Accuracy Metric Is Higher the Better. The Best Results of Each Metric Are Bold.

Fig. 3. Illustration of examples of depth predictions on the Make3D data set [26]. Note that our model is only trained on the KITTI data set and directly tested on Make3D.

KITTI raw data set. We chose an incremental order for the proposed techniques to avoid too many loss combinations. As shown in Tables IV and V, we had the following observations.

1) The 3-D structure consistency was essential for improving the performance of depth estimation, while the log-scale 3-D structural consistency loss could further improve the depth estimation. Our method was more stable for all metrics, especially noticeable on metrics that were especially noticeable on metrics that sensitive to large depth errors, e.g., the square relative error and RMSE.

2) The epipolar geometric constraint could improve the performance of both depth and pose estimation, especially for the pose estimation, which was consistent with the report in [15]. Although the properties of epipolar geometry, low-rankness constraints, and union-of-subspaces constraints, were suitable for self-supervised optical flow estimation, the improvements for self-supervised depth and pose estimation were limited.

In summary, the epipolar geometry helped both the pose and depth estimation, while the log-scale 3-D geometric consistency terms also could improve the performance of depth estimation.

Alternative Training Policy: We also conducted an ablation study over the proposed alternative training compared with the joint training. It showed that the alternative training policy was as effective as the jointly training policy and more effective by training with different geometric constraints.

Different Depth Network Structures: For the sake of completeness, similar to [13], we also provided an ablative analysis of its generalization ability to different depth networks. To this end, we further considered two variations on the well-performed structures, the ResNet50 [14] as the encoder and the PackNet [12]. The estimation results of this consideration were shown in Table VI, where we could see that the proposed method could consistently improve the performance with different depth networks for all considered metrics.

Training Time and Inference Time: We compared with Monodepth2 [11], and both methods were trained on a single 12GB Titan XP GPU. We measured the time taken for each training iteration consisting of both forward and backward propagation. The resolution of the input images was 640 × 192. Due to computing the epipolar geometric constraints, the training time was three times longer than the Monodepth2 [11], while the inference times were identical because the network structure was not changed in the proposed method. The training time and the inference time cost by the proposed methods were shown in Table VII.

Different Choices of the Weights of the Losses: The weights of different losses were vitally important for the proposed methods. In this article, the weights of different losses were chosen by the empirical value as well as by trial and error. To prove the rationality of choosing these parameters for different weights, we given the experiments on changing some of these parameters in Table VIII, the “Basic” means that the model was trained with the weights given in the parameters.
TABLE VI
ABLATIVE ANALYSIS OF THE GENERALIZATION OF OUR PROPOSED NETWORK ON VARIANT NETWORK STRUCTURES. ALL MODELS ARE SOLELY TRAINED ON THE MONOCULAR IMAGES OF THE KITTI RAW DATA SET WITHOUT PRETRAINING ON CITYSCAPES [5] AND POSTPROCESSING [10].

| Methods | Error metric | RMSE | RMSE log | Accuracy metric | δ < 1.25 | δ < 1.25² | δ < 1.25³ |
|---------|--------------|------|----------|----------------|----------|-----------|-----------|
|         | Abs Rel      | Sq Rel |          |                |          |           |           |
| Monodepth2 [11] | 0.112  | 0.831 | 4.662    | 0.187          | 0.883    | 0.962     | 0.982     |
| Ours    | 0.107        | 0.792 | 4.661    | 0.182          | 0.887    | 0.962     | 0.983     |
| PackNet [12]    | 0.111  | 0.785 | 4.601    | 0.189          | 0.878    | 0.960     | 0.982     |
| Ours    | 0.107        | 0.725 | 4.344    | 0.179          | 0.882    | 0.962     | 0.983     |
| ResNet-50 [14]|              |       |          |                |          |           |           |

TABLE VII
TRAINING TIME OF PER ITERATION AND INFEERENCE TIME ON PER IMAGE OR IMAGE PAIRS.

| Models     | Training | Inference |
|------------|----------|-----------|
| Monodepth2 [11] | DepthNet | 0.43 s | 4.9 ms |
| Ours       | PoseNet  | 0.6 ms   |        |
|            | DepthNet | 1.125 s | 4.9 ms |
| Ours       | PoseNet  | 0.25 s  | 0.6 ms |

TABLE VIII
INFLUENCE OF PARAMETERS SETTING ON THE WEIGHTS OF DIFFERENT LOSSES ON THE KITTI DATA SET.

| Parameters | Abs Rel | Sq Rel | RMSE | RMSE log10 |
|------------|---------|--------|------|------------|
| Basic      | 0.112   | 0.835  | 4.748| 0.189      |
| λ_s=0.003  | 0.113   | 0.850  | 4.783| 0.191      |
| λ_r=0.03   | 0.113   | 0.851  | 4.767| 0.191      |
| λ_o=0.002  | 0.113   | 0.840  | 4.764| 0.191      |
| λ_m=0.0002 | 0.112   | 0.850  | 4.784| 0.190      |

TABLE IX
INFLUENCE OF THE MODEL TRAINED WITHOUT PRETRAINING ON THE KITTI DATA SET.

| Methods     | Type | Abs Rel | Sq Rel | RMSE | RMSE log10 |
|-------------|------|---------|--------|------|------------|
| Monodepth2 [11] | M    | 0.132   | 1.044  | 5.142| 0.210      |
| Ours-Pose first | M    | 0.129   | 0.948  | 4.969| 0.206      |
| Ours-Depth first | M    | 0.126   | 0.943  | 4.927| 0.199      |

V. CONCLUSION
In this article, we put forward a self-supervised depth and pose estimation architecture that incorporates both geometric principles and photometric-based learning metrics. Our main contribution is to better utilize the mutual dependency of the depth and pose learning by alternatively training each network with different geometric constraints and simplifying the ICP registration-based optimization objectives by incorporating the epipolar geometry. Specifically, the log-scale 3-D structural consistency loss and the epipolar geometry embedded with the ICP registration are adopted in the respective tasks. To make the result more robust and reliable, we incorporate novel ingredients by the properties of epipolar geometry, namely, the low-rankness and self-expression in union-of-subspaces, for depth and pose networks learning, respectively, and try to eliminate the negative effects of the moving objects by self-expression in the union of epipolar subspaces. The experimental results demonstrate that our method can obtain depth maps with better contour for the foreground target and can generalize well to the unseen data sets. Further explorations include applying to the uncalibrated data set and adaptively setting the weights of different loss functions.

REFERENCES
[1] Y. Cao, Z. Wu, and C. Shen, “Estimating depth from monocular images as classification using deep fully convolutional residual networks,” IEEE Trans. Circuits Syst. Video Technol., vol. 28, no. 11, pp. 3174–3182, Nov. 2018.
[2] P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y.-C. F. Wang, “Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2624–2632.
[3] Y. Chen, C. Schmid, and C. Sminchisescu, “Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 7063–7072.
[4] P. Ji, H. Li, M. Salzmann, and Y. Zhong, “Robust multi-body feature tracker: A segmentation-free approach,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 3843–3851.
[5] M. Cordts et al., “The cityscapes dataset for semantic urban scene understanding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 3213–3223.
