2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Self-supervised Multi-frame Monocular Depth Estimation with Pseudo-LiDAR Pose Enhancement

Wenhua Wu, Guangming Wang, Jiquan Zhong, Hesheng Wang, and Zhe Liu

Abstract— **Depth estimation is one of the most important tasks in scene understanding. In the existing joint self-supervised learning approaches of depth-pose estimation, depth estimation and pose estimation networks are independent of each other. They only use the adjacent image frames for pose estimation and lack the use of the estimated geometric information. To enhance the depth-pose association, we propose a monocular multi-frame unsupervised depth estimation framework, named PLPE-Depth. There are a depth estimation network and two pose estimation networks with image input and pseudo-LiDAR input. The main idea of our approach is to use the pseudo-LiDAR reconstructed from the depth map to estimate the pose of adjacent frames. We propose depth re-estimation with a better pose between the image pose and the pseudo-LiDAR pose to improve the accuracy of estimation. Besides, we improve the reconstruction loss and design a pseudo-LiDAR pose enhancement loss to facilitate the joint learning. Our approach enhances the use of the estimated depth information and strengthens the coupling between depth estimation and pose estimation. Experiments on the KITTI dataset show that our depth estimation achieves state-of-the-art performance at low resolution. Our source codes will be released on https://github.com/IRMVLab/PLPE-Depth.**

I. INTRODUCTION

Autonomous systems and intelligent robots have long been people's pursuits [1]–[3]. Depth information is of great significance for the environment perception of them and is widely used in autonomous robot localization and navigation [4], 3-D reconstruction [5], object detection [6], etc. Based on some sensors such as LiDAR and RGB-D cameras, depth information can be directly acquired. However, they are often expensive [7], large in size, and have high power consumption. By contrast, depth estimation from monocular images costs less and is easy to be applied.

Monocular depth estimation is widely studied. Traditional methods, such as structure from motion (SfM) [8] and stereo vision matching [9], rely on feature matching and can only obtain sparse depth maps. With the development of deep learning, pixel-level end-to-end depth estimation of an image becomes a reality [10], [11]. The mainstream methods can be divided into supervised, self-supervised, and semi-supervised. The self-supervised method does not need expensive depth ground-truth and uses neighboring frames as supervised signals, thus realizing the joint learning of depth and pose. Various novel network structures, loss functions, and data augmentation methods have been proposed [12]–[17]. The current results of self-supervised depth estimation can even be comparable to some supervised approaches.

However, in previous works, the pose is usually estimated by independent convolutional neural networks, which only use adjacent image frames as input, and ignore the use of estimated depth information. 3D scene flow estimation studies the motion correlation of adjacent frames [18]–[21]. In addition, a large number of pose estimation algorithms based on 3D point clouds have been proposed, which have achieved comparable results on LiDAR odometry [22]–[24]. Considering that the estimated depth map can be reconstructed to 3D point clouds (pseudo-LiDAR), we can use the point cloud pose estimation network to estimate the pose of the pseudo-LiDAR.

We propose a new self-supervised training framework for joint training of depth and pose. Our framework consists of two pose estimation networks. One is a convolutional neural network with images input, and the other is a point cloud based pose network with pseudo-LiDAR input. A pseudo-LiDAR is an explicit representation of the scene's 3D information, although it may be inaccurate. On the one hand, the pose estimation of the pseudo-point cloud can test the accuracy of the depth estimation. The more accurate the pseudo-point clouds are, the more accurately the pose can be estimated. On the other hand, the pseudo-point clouds contain more three-dimensional information than the images, which are more capable of estimating the pose with higher accuracy. We improve the image reconstruction loss to strengthen the coupling of depth estimation and pose estimation so that they can promote each other to obtain higher precision results.

As far as we know, this paper is the first attempt to use pseudo-LiDAR to enhance depth estimation. The main contributions of this paper include:

- We propose PLPE-Depth, a new self-supervised monocular depth estimation framework that uses pseudo-LiDAR pose estimation to enhance the utilization of estimated depth information.
- We propose depth re-estimation, which strengthens the mutual coupling between depth estimation and pose estimation. We get the optimal pose from the image pose and the pseudo-LiDAR pose, which is the one with minimum reconstruction loss. The enhanced depth map is reestimated with the optimal pose.
- We design a pseudo-LiDAR pose enhancement loss with

*This work was supported in part by the Natural Science Foundation of China under Grant U1913204. The first two authors contributed equally. Corresponding Author: Zhe Liu (liuzhesjtu@sjtu.edu.cn).

W. Wu and Z. Liu are with MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China. G. Wang, J. Zhong, and H. Wang are with Department of Automation, Insititute of Medical Robotics, Key Laboratory of System Control and Information Processing of Ministry of Education, Key Laboratory of Marine Intelligent Equipment and System of Ministry of Education, Shanghai Engineering Research Center of Intelligent Control and Management, Shanghai Jiao Tong University, Shanghai 200240, China.

the enhanced depth map and the pseudo-LiDAR pose. It can facilitate the joint training of depth estimation and pose estimation.
- Our approach achieves state-of-the-art performance at low resolution on the KITTI dataset.

II. RELATED WORK

A. Supervised Depth Estimation

Dense depth ground truth can be obtained using depth sensors such as RGB-D cameras and LiDAR. It can be used as the depth estimation regression target. Eigen et al. [10] first use convolutional neural networks for depth estimation. There are two convolutional neural networks for global depth estimation and local refinement, and the global estimation result is part of the local network input. In addition, they introduce a scale-invariant loss to measure depth relationships. They build a network structure in [25] that can perform three tasks: estimating image depth, surface normals, and semantic labels. A series of scales are used to make depth estimation gradually refined. Laina et al. [26] use ResNet-50 [27] for depth estimation, which reduces network parameters so that the network can operate in real-time. They also propose a new loss called the Reverse Huber loss, which uses different calculation methods for pixels with large and small error estimates. Cao et al. [28] formulate depth estimation as a classification problem, estimating the depth range of each pixel and using cross-entropy loss for training. Bhat et al. [29] divide the depth range into multiple bins, and the final depth result is the weighted sum of the center values of each bin. Their method can accommodate differences in depth distribution between images.

B. Self-supervised Monocular Depth Estimation

The self-supervised monocular depth estimation methods no longer rely on the supervision of depth ground truth and use image reconstruction loss from monocular image sequences to train. However, illumination variation, occlusion, and dynamic objects can affect the accuracy of reconstruction loss. Godard et al. [30] propose a minimal reconstruction loss to solve the occlusion problem, and a multi-scale sampling method to solve the visual tail shadow. They also mask the pixels that have not changed in neighboring frames to solve the interference caused by moving objects. Klingner et al. [31] use a network branch to semantically segment images, distinguish dynamic objects and mask them to solve the interference of dynamic objects. Their semantic segmentation network and depth estimation network can be trained jointly. Jung et al. [32] also explore how to use semantic information to improve depth estimation performance, and they propose a metric learning method. Watson et al. [33] introduce depth hints to prevent the reprojection loss from falling into local minima. Some works build new frameworks. Zhao et al. [34] construct a geometry-aware symmetric domain adaptation framework (GASDA) inspired by CycleGAN [11]. MLDA-Net [35] proposes multi-level feature extraction and a global and local dual attention strategy.

C. Multi-frame Monocular Depth Estimation

Monocular cameras can provide continuous image sequences, and some works have attempted to use image sequences at *test time* to improve depth estimation [36]–[43]. Among them, the test-time refinement method requires a lot of extra computation for a group of test images, the time cost is high [38]. The second method uses a recurrent neural network to extract the features of sequence frames [41], [42]. This does not explicitly use the geometric information of sequence frames. Independent of recurrent neural networks and test time refinement, [43] proposes A new multi-frame depth estimation model. They design an end-to-end adaptive cost volume and introduce new losses to overcome dynamic object effects. Our method is improved based on ManyDepth [43].

D. Pseudo Point Cloud and Deep LiDAR Odometry

LidAR is one of the main sensors for autonomous driving. It can provide high-precision 3D scene information, but the cost is expensive. Recently, pseudo-liDAR point clouds reconstructed by depth estimation have been widely used in object detection [6], [7], [44], scene flow estimation [45], and visual odometry [46], etc. Deep LiDAR Odometry is a hot research topic. The main idea is to estimate the pose transformation of continuous point clouds through the network. Some works have designed innovative network structures and achieved good estimation results [22], [47]–[50]. The pseudo-LiDAR has the same data type as the point cloud acquired by LiDAR. These networks can also be applied to pose estimation of pseudo-LiDAR.

III. MAIN APPROACH

A. System Overview

We propose a novel self-supervised multi-frame monocular depth estimation framework. The input to the system contains a continuous RGB image sequence $(I_{t-N}, I_{t-N+1}, ..., I_t)$ and camera intrinsics matrix K. We aim to obtain a high-accuracy multi-frame monocular depth estimation network θ_{depth}. The estimated depth map is expressed as $D_t = \theta_{depth}(I_{t-N}, I_{t-N+1}, ..., I_t)$. The framework of our method is shown in Fig. 1. We introduce a pseudo-LiDAR point cloud pose estimation network θ_{pc_pose}. The estimated depth map is reconstructed into a pseudo-LiDAR. After preprocessing of the pseudo-LiDAR, the pose estimation of the continuous frame pseudo-LiDAR is carried out. Pseudo-LiDAR contains more 3D information than images so that it can achieve more accurate pose estimation. In the training process, we can obtain two kinds of poses: the image-input network pose estimation $T_{t\rightarrow t'\_image}$ and the pseudo-LiDAR-input network pose estimation $T_{t\rightarrow t'\_pc}$. We will calculate the photometric reconstruction error separately and choose the optimal pose to perform depth re-estimation. We design a Pseudo-LiDAR pose enhancement loss to facilitate the joint learning of depth estimation and pseudo-LiDAR pose estimation.

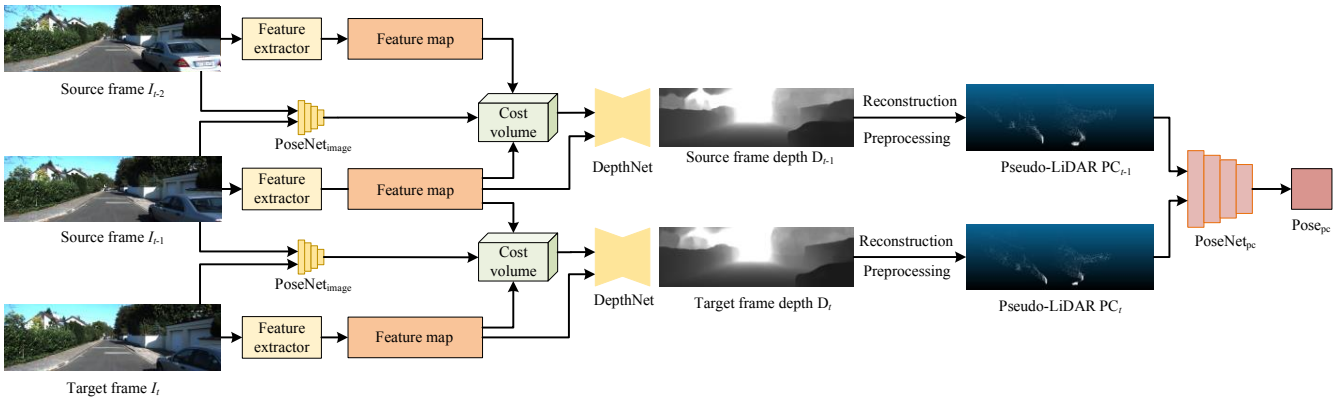

Fig. 1. Our multi-frame depth estimation and pose estimation framework. It consists of a feature extractor, a depth generator, an image-input pose estimation network, and a pseudo-LiDAR pose estimation network. The image-input pose network estimates the pose transformation of adjacent images, and then the cost volume is built. The depth generator generates depth maps from cost volume and image features. The pseudo-LiDAR is obtained by reconstruction and preprocessing. Finally, The pseudo-LiDAR pose estimation network estimates the pose transformation of adjacent frames.

B. Multi-frame Monocular Depth Estimation

Our multi-frame monocular depth estimation network is inspired by ManyDepth [43]. It contains a feature extraction module, a cost volume, and a depth generator. The input of the network is consecutive frames of monocular video (I_{t-1}, I_t). I_{t-1} is the source image, and I_t is the target image. The feature extraction module extracts image features. Image-input pose estimation network estimates pose transformation. The estimated pose is used to warp the source image feature map, and then the cost volume is constructed from the warped source image feature map and the target image feature map. The depth generator generates a dense depth map of I_t from the cost volume and the feature map of I_t. For the depth estimation of a static single frame image, we make $I_{t-1} = I_t$ and then do the same as above.

C. Pseudo-LiDAR Pose Estimation

In this section, we introduce one of our main innovations, pseudo-LiDAR pose estimation. Different from the previous pose estimation, pseudo-LiDAR pose estimation depends on the result of depth estimation.

1) Pseudo-LiDAR: With the dense depth map of depth estimation and the camera parameter matrix, the coordinates of each pixel in 3D space (x, y, z) can be projected. The calculation method is given by the following formula:

$$\begin{cases} x = \frac{(u-c_U)}{f_U} d, \\ y = \frac{(v-c_V)}{f_V} d, \\ z = d, \end{cases} \tag{1}$$

where d is the depth of the pixel, (u, v) is the pixel coordinates, and (c_U, c_V) is the pixel coordinates of the image center. f_U and f_V are the horizontal and vertical focal lengths of the camera.

We get the pseudo-LiDAR point cloud PC_p through the above calculation. Compared with depth map, pseudo-LiDAR is a more direct representation of 3D scenes and can provide richer geometry, shape, and scale information. A richer representation of 3D information is the prerequisite for more accurate pose estimation.

2) Data Preprocessing of Pseudo-LiDAR Point Cloud: The original dense point cloud has a total of $H \times W$ points, which contains lots of large error points. To improve the quality of pseudo-LiDAR, points higher than d_h meters and points lower than d_l meters are removed as sky points and ground points, and distant points with a depth over d_f are also removed. Ground points contain less feature information, and high points and far points tend to have greater depth estimation errors. In order to reduce the computational burden and ensure the same number of points in each point cloud, we randomly sample N points to obtain sparse high-quality Pseudo-LiDAR point cloud.

3) Pseudo-LiDAR Pose Estimation Network: We use our previous work PWCLO-Net [22] as the pseudo-LiDAR pose estimation network. It contains a feature pyramid encoder, an attentive cost volume for generating embedding features, an embedding mask to weight local motion, a pose generation network to generate the original pose, and three pose refinement networks to refine the pose from coarse to fine. The inputs of the network are adjacent pseudo-LiDAR PC_t and $PC_{t'}$, and the outputs are four poses from coarse to fine. We select the most refined one among them as the final pseudo-LiDAR pose estimation result.

D. Self-supervised Monocular Depth Estimation

Similar to [43], we only select the before and after frames of the target image I_t as the source images $I_{t'}$ during training. We use the depth map of the target image D_t and the pose transformation $T_{t\to t'}$ to reconstruct the image with the same perspective as the target image. The reconstructed image is:

$$I_{t'\to t} = I_{t'} \left\langle proj(D_t, T_{t\to t'}, K) \right\rangle, \tag{2}$$

where $\langle\rangle$ is the sampling operator. $proj()$ returns 2D coordinates of the projected depths D_t in $I_{t'}$. K is camera intrinsics.

TABLE I

DEPTH EVALUATION RESULTS ON THE KITTI EIGEN SPLIT [10]. PREVIOUS SELF-SUPERVISED MONOCULAR DEPTH ESTIMATION METHODS ARE COMPARED WITH OURS. WE ALSO LIST SOME SUPERVISED LEARNING METHODS. **D**: DEPTH SUPERVISION. **S**: SELF-SUPERVISED FROM STEREO VIDEO. **M**: SELF-SUPERVISED FROM MONOCULAR VIDEO. **AT**: AUXILIARY TASK. THE BEST RESULTS IN EACH SUBSECTION ARE IN BOLD AND THE SECOND BEST IS UNDERLINED

Methods	with AT	Train	AbsRel	SqRel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
			Lower is better				Higher is better		
Eigen et al. [10]		D	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu at al. [51]		D	0.202	1.614	6.523	0.275	0.678	0.895	0.965
Garg et al. [52]		S	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Godard et al. [12]		S	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Zhan et al. [14]		S	0.144	1.391	5.869	0.241	0.803	0.928	0.969
Kuznietsov et al. [53]		DS	0.113	0.741	4.621	0.189	0.862	0.960	0.986
Guo et al. [54]		DS	0.096	0.641	4.095	0.168	0.892	0.967	0.986
P3Depth [55]		D	**0.071**	**0.270**	**2.842**	**0.103**	**0.953**	**0.993**	**0.998**
Geonet-VGG [56]	✓	M	0.164	1.303	6.090	0.247	0.765	0.919	0.968
Geonet-Resnet [56]	✓	M	0.155	1.296	5.857	0.233	0.793	0.931	0.973
Shen et al. [57]		M	0.156	1.309	5.37	0.236	0.797	0.929	0.969
DPSNet [58]	✓	M	0.159	1.355	5.679	0.232	0.785	0.935	0.973
Wang et al. [59]		M	0.151	1.257	5.583	0.228	0.810	0.936	0.974
DF-Net [60]	✓	M	0.150	1.124	5.507	0.223	0.806	0.933	0.973
GANVO [61]		M	0.150	1.141	5.448	0.216	0.808	0.939	0.975
CC [62]	✓	M	0.140	1.070	5.326	0.217	0.826	0.941	0.975
EPC++ [63]	✓	M	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2depth [36]		M	0.141	1.026	5.291	0.215	0.816	0.945	0.979
DOP [64]	✓	M	0.140	1.068	5.255	0.217	0.827	0.943	0.977
SC-SFM [65]		M	0.137	1.089	5.439	0.217	0.830	0.942	0.975
Monodepth2 [30]		M	0.115	0.882	4.701	0.190	0.879	0.961	0.982
Li et al. [66]		M	0.130	0.950	5.138	0.209	0.843	0.948	0.978
ΩNet [67]	✓	M	0.125	0.805	4.795	0.195	0.849	0.955	**0.983**
Chen et al. [68]		M	0.129	0.976	4.958	0.203	8.848	0.951	0.979
Jia et al. [69]		M	0.130	0.957	4.907	0.203	0.851	0.954	0.980
Packent-SFM [70]		M	0.111	0.785	4.601	0.189	0.878	0.960	0.982
Wang et al. [71]		M	0.106	0.799	4.662	0.187	0.889	0.961	0.982
RM-Depth [72]		M	0.108	**0.710**	4.513	0.183	0.884	0.964	**0.983**
ManyDepth [43]		M	0.098	0.770	4.459	**0.176**	**0.900**	**0.965**	**0.983**
PLPE-Depth (Ours)		M	**0.096**	0.716	**4.414**	**0.176**	**0.900**	0.964	**0.983**

The reconstruction loss pe is the weighted sum of L1 loss and SSIM.

$$pe(I_a, I_b) = \alpha SSIM(I_a, I_b) + (1-\alpha)\|I_a - I_b\|_1, \quad (3)$$

where α is a weighting factor.

In the training process, depth estimation is performed twice for a set of images.

1) First Depth Estimation: For the first depth estimation, the pose estimation results of the image-input pose estimation network θ_{image_pose} are used. The depth maps of $I_{t'}$ and I_t are as follows:

$$T_{t\_image} = \theta_{image\_pose}(I_{t'}, I_t). \quad (4)$$

$$T_{t'\_image} = \theta_{image\_pose}(I_{t'-1}, I_{t'}). \quad (5)$$

$$D_{t\_raw} = \theta_{depth}(I_{t'}, I_t, T_{t\_image}). \quad (6)$$

$$D_{t'} = \theta_{depth}(I_{t'-1}, I_{t'}, T_{t'\_image}). \quad (7)$$

2) Pseudo-LiDAR Pose Estimation: The pseudo-LiDAR is back projected from the depth map using the camera intrinsics matrix K. After preprocessing, the adjacent pseudo-LiDAR $\{PC_{t'}, PC_t\}$ is input into the pseudo-LiDAR pose estimation network θ_{pc_pose}.

$$PC_t = [recons(D_t, K)], \; PC_{t'} = [recons(D_{t'}, K)], \quad (8)$$

where $recons$ means reconstruction from two dimensions to three dimensions, and $[\ ]$ represents the preprocessing operation in Sec. III-C.2.

$$T_{t\_pc} = \theta_{pc\_pose}(PC_{t'}, PC_t). \quad (9)$$

3) Depth Re-estimation: We calculate the reconstruction loss using T_{t_image} and T_{t_pc}, respectively. The one with minimum reconstruction loss is selected as the optimal pose:

$$l_{pe} = pe(I_t, I_{t'} \langle proj(D_{t\_raw}, T, K)\rangle). \quad (10)$$

$$T_{t\_optimal} = \underset{T \in \{T_{t\_image}, T_{t\_pc}\}}{\operatorname{argmin}} l_{pe}. \quad (11)$$

The depth map of I_t is re-estimated using the optimal pose.

$$D_{t\_new} = \theta_{depth}(I_{t'}, I_t, T_{t\_optimal}). \quad (12)$$

4) Pseudo-LiDAR Pose Enhancement Loss: Following [30], [43], we use the minimum of reconstruction loss pe over all source images as per-pixel reconstruction loss L_p.

$$L_p = \min_{t'} pe(I_t, I_{t'} \langle proj(D_{t\_raw}, T_{t\_optimal}, K)\rangle). \quad (13)$$

The raw loss is:

$$L_{raw} = (1-M)L_p + L_{consistency} + L_{smooth}, \quad (14)$$

where M is a binary mask that is one in unreliable pixels and zero in other pixels and $L_{consistency}$ is the consistency

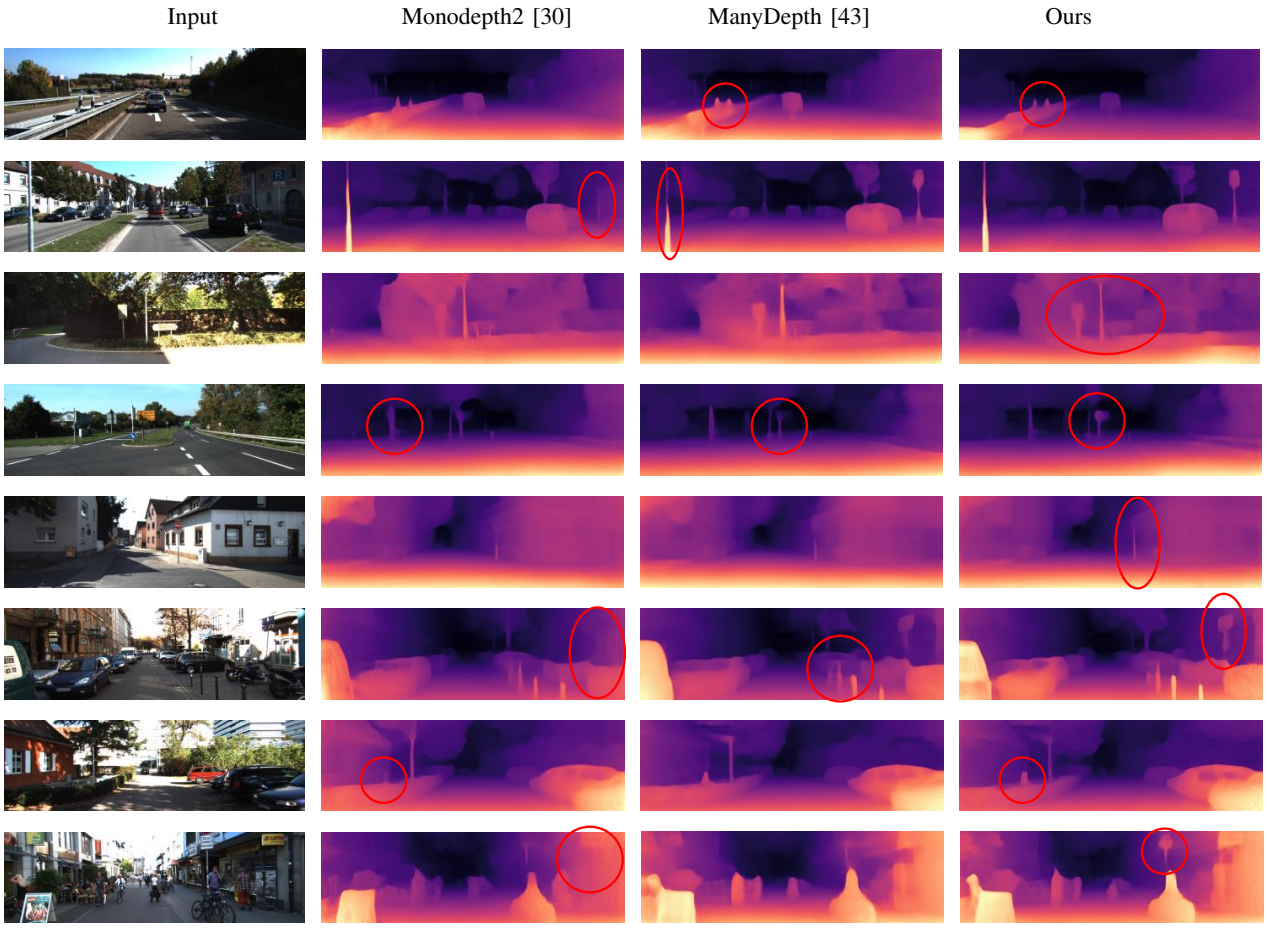

Fig. 2. Depth visualization. Monodepth2 [30] uses I_t, while ManyDepth [43] and ours use $\{I_{t-1}, I_t\}$. Our algorithm can provide more refined results on posts and road signs.

loss. They are both from [43]. L_{smooth} is the smoothness loss from [30].

The new depth map and pseudo-LiDAR pose are used to calculate the pseudo-LiDAR pose enhancement loss.

$$L_{p\_aug} = \min_{t'} pe(I_t, I_{t'} \langle proj(D_{t\_new}, T_{t\_pc}, K)\rangle)). \quad (15)$$

$$L_{aug} = (1 - M')L_{p\_aug} + L'_{consistency} + L'_{smooth}. \quad (16)$$

Our final loss is $L = L_{raw} + L_{aug}$.

It is worth noting that although we use the future frame of the target image in the training process, we only use the past frame of the target image in the test because the future frame is unpredictable in reality.

IV. EXPERIMENTS

A. Implementation Details

a) KITTI Raw Dataset: The KITTI raw dataset [73] is a benchmark dataset for depth estimation and pose estimation. It contains image sequences with depth ground truth in a variety of scenes. We use the data split of Eigen et al. [10] to train and test models. Following previous works, the range of the depth estimate is set to $[10^{-3}, 80m]$, because few pixels are deeper than $80m$.

b) Network Architecture: The architectures of image-input pose estimation network θ_{image_pose} and depth estimation network θ_{depth} are the same as in [43]. Following [43], there is also a teacher monocular network $\theta_{consistency}$, which uses the standard architecture from [30]. For pseudo-LiDAR pose estimation network θ_{pc_pose}, we use the standard architecture with 8192 points input from [22].

c) Pseudo-LiDAR Pose Estimation Network Pre-training: We use the pretrained weights of θ_{image_pose}, θ_{depth}, and $\theta_{consistency}$ provided by [43] to pre-train our pseudo-LiDAR pose estimation network θ_{pc_pose}. We fix the weights of θ_{image_pose}, θ_{depth}, and $\theta_{consistency}$, and use the pose prediction results of θ_{image_pose} to supervised train the pseudo-LiDAR-input pose estimation network θ_{pc_pose}. The loss function adopted the total loss in [22]. The pseudo-LiDAR preprocessing parameters d_h is $2.2m$, d_l is $1.1m$, and d_f is $30m$. The number of sampling points N is 8192. The learning rate is 10^{-3}, and the batch size is 6. We use the Adam optimizer to train for 20 epochs. The training results are used as the pre-trained weights of θ_{pc_pose}.

We initialize our models with the pre-trained weights. Adjacent monocular images $\{I_{t-2}, I_{t-1}, I_t, I_{t+1}\}$ are used for training and $\{I_{t-1}, I_t\}$ for test. The resolution of the image is 640×192. We use the Adam parameter optimizer

TABLE II
ABLATION RESULTS ON THE KITTI EIGEN SPLIT [10].

Methods	AbsRel	SqRel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
	Lower is better				Higher is better		
w/o Pseudo-LiDAR Pose Estimation	0.100	0.753	4.480	0.178	0.895	0.963	0.982
w/o Depth Re-estimation	0.101	0.759	4.505	0.180	0.892	0.963	**0.983**
w/o Pseudo-LiDAR Pose Enhancement Loss	0.098	0.728	4.428	0.177	0.897	**0.964**	**0.983**
Full model	**0.096**	**0.716**	**4.414**	**0.176**	**0.900**	**0.964**	**0.983**

to train for 20 epochs. The weighting factor α is 0.85. The learning rate is 10^{-4}. After training 15 epochs, the learning rate decays to 10^{-4}. The batch size is 6. All the experiments are performed on an NVIDIA RTX2080 Ti GPU.

B. Evaluation Results

The results of the depth estimation we trained on the KITTI dataset are shown in Table I. As a comparison, we present the results of other self-supervised monocular depth estimators. Similar to us, Wang et al. [71] and ManyDepth [43] use adjacent frames $\{I_{t-1}, I_t\}$ in the test-time. ManyDepth [43] is able to achieve better performance with higher resolution 1024×320 and test-time refinement. To be fair, we list the results of [43] with low resolution 640×192 and without test-time refinement.

As it can be seen, we have achieved the current optimum in AbsRel, RMSE, RMSE log, $\delta < 1.25$, $\delta < 1.25^3$. In SqRel and $\delta < 1.25^2$, we achieve the second best. Compared with ManyDepth2 [43], we are significantly better in AbsRel, SqRel, and RMSM, and equal in RMSE log, $\delta < 1.25$, and $\delta < 1.25^3$. It is only slightly worse on $\delta < 1.25^2$. We approach in AbsRel and outperform advanced supervised method in $\delta < 1.25$. Fig. 2 shows the depth maps of Monodepth [30], ManyDepth [43] and ours. The red circles mark the parts with obvious contrast. Our algorithm can provide more refined results on posts and road signs. This is thanks to our use of pseudo-LiDAR.

C. Ablation Study

Our ablation experiments are divided into three groups.

1) Without Pseudo-LiDAR Pose Estimation: When Pseudo-LiDAR Pose Estimation is removed, the network framework degrades into to ManyDepth [43]. The loss is:

$$L_p = \min_{t'} pe(I_t, \langle proj(D_{t\_raw}, T_{t\_image}, K)\rangle). \quad (17)$$

$$L_{ablation\_1} = (1 - M)L_p + L_{consistency} + L_{smooth}. \quad (18)$$

2) Without Depth Re-estimation: Only the first depth estimation is performed. The image-input pose T_{t_image} is estimated by θ_{pc_pose}. The depth map is:

$$D_t = \theta_{depth}(I_{t'}, I_t, T_{t\_image}). \quad (19)$$

The pseudo-LiDAR pose T_{t_pc} is estimated with the first depth estimation results. Instead of the optimal pose, T_{t_pc} is used to calculate the reconstruction loss.

$$L_p = \min_{t'} pe(I_t, \langle proj(D_{t\_raw}, T_{t\_pc}, K)\rangle). \quad (20)$$

The final loss function is:

$$L_{ablation\_2} = (1 - M)L_p + L_{consistency} + L_{smooth}. \quad (21)$$

3) Without Pseudo-LiDAR Pose Enhancement Loss: Only the raw loss is used as $L_{ablation\_3} = L_{raw}$.

The experimental conditions are the same. The results of the ablation experiments are shown in Table II. As one can see, in the absence of pseudo-LiDAR pose estimation, depth re-estimation, and pseudo-LiDAR pose enhancement loss, the depth estimation results are significantly worse. The results prove that each of our innovations promotes depth estimation.

D. Future Improvement

Although our depth estimates have performed quite well, there is still much need to be improved. We notice that the current pose estimation is still not accurate enough, which may limit depth estimation optimization. One possible solution is to use denser point clouds or fusion of image and point clouds for pose estimation. Another way is to sample more critical points in the point cloud instead of sampling randomly. Another problem is that the depth re-estimation increases the computational load. How to simplify the calculation is an optimization direction. In addition, we only used depth re-estimation during training. How to use pseudo-LiDAR to correct the initial depth estimation at test-time needs further exploration.

V. CONCLUSION

We propose a new self-supervised learning framework for multi-frame depth estimation and pose estimation. We introduce pseudo-LiDAR pose estimation and use the result of depth estimation to estimate pose. We select the one with smaller reconstruction loss between the pseudo-LiDAR-input pose and image-input pose for multi-frame depth re-estimation. This strengthens the coupling between depth estimation and pose estimation. We also improve the reconstruction loss and design a Pseudo-LiDAR pose enhancement loss to facilitate the joint learning. As far as we know, this paper is the first attempt to use pseudo-LiDAR to enhance depth estimation and has achieved pretty good performance on the KITTI dataset. Our framework still has room for improvement, we will consider pseudo-LiDAR and image features fusion for pose estimation, and try some other depth re-estimation methods. In addition, we will try to remove more points with large errors in the pseudo-LiDAR to reduce the estimation error in our future work.

REFERENCES

[1] H. Marzbani, H. Khayyam, C. N. TO, V. Quoc, and R. N. Jazar, "Autonomous vehicles: Autodriver algorithm and vehicle dynamics," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3201–3211, 2019.

[2] H. Qiao, S. Zhong, Z. Chen, and H. Wang, "Improving performance of robots using human-inspired approaches: a survey," *Science China Information Sciences*, vol. 65, no. 12, pp. 1–31, 2022.

[3] H. Qiao, Y.-X. Wu, S.-L. Zhong, P.-J. Yin, and J.-H. Chen, "Brain-inspired intelligent robotics: Theoretical analysis and systematic application," *Machine Intelligence Research*, vol. 20, no. 1, pp. 1–18, 2023.

[4] J. Biswas and M. Veloso, "Depth camera based indoor mobile robot localization and navigation," in *IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 1697–1702.

[5] T.-N. Nguyen, H.-H. Huynh, and J. Meunier, "3d reconstruction with time-of-flight depth camera and multiple mirrors," *IEEE Access*, vol. 6, pp. 38 106–38 114, 2018.

[6] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8445–8453.

[7] R. Qian, D. Garg, Y. Wang, Y. You, S. Belongie, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. Chao, "End-to-end pseudo-lidar for image-based 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5881–5890.

[8] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*. Springer, 2008, pp. 44–57.

[9] H. Tao, H. S. Sawhney, and R. Kumar, "A global matching framework for stereo computation," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 1. IEEE, 2001, pp. 532–539.

[10] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[11] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[12] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 270–279.

[13] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2624–2632.

[14] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 340–349.

[15] C. Ling, X. Zhang, and H. Chen, "Unsupervised monocular depth estimation using attention and multi-warp reconstruction," *IEEE Transactions on Multimedia*, vol. 24, pp. 2938–2949, 2021.

[16] J. Wei, S. Pan, W. Gao, and T. Zhao, "Triaxial squeeze attention module and mutual-exclusion loss based unsupervised monocular depth estimation," *Neural Processing Letters*, pp. 1–16, 2022.

[17] G. Wang, J. Zhong, S. Zhao, W. Wu, Z. Liu, and H. Wang, "3d hierarchical refinement and augmentation for unsupervised learning of depth and pose from monocular video," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[18] G. Wang, X. Wu, Z. Liu, and H. Wang, "Hierarchical attention learning of scene flow in 3d point clouds," *IEEE Transactions on Image Processing*, vol. 30, pp. 5168–5181, 2021.

[19] G. Wang, Y. Hu, X. Wu, and H. Wang, "Residual 3-d scene flow learning with context-aware feature extraction," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–9, 2022.

[20] G. Wang, C. Jiang, Z. Shen, Y. Miao, and H. Wang, "Sfgan: Unsupervised generative adversarial learning of 3d scene flow from the 3d scene self," *Advanced Intelligent Systems*, vol. 4, no. 4, p. 2100197, 2022.

[21] G. Wang, Y. Hu, Z. Liu, Y. Zhou, M. Tomizuka, W. Zhan, and H. Wang, "What matters for 3d scene flow network," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*. Springer, 2022, pp. 38–55.

[22] G. Wang, X. Wu, Z. Liu, and H. Wang, "Pwclo-net: Deep lidar odometry in 3d point clouds using hierarchical embedding mask optimization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 910–15 919.

[23] G. Wang, X. Wu, S. Jiang, Z. Liu, and H. Wang, "Efficient 3d deep lidar odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[24] J. Nubert, S. Khattak, and M. Hutter, "Self-supervised learning of lidar odometry for robotic applications," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 9601–9607.

[25] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2650–2658.

[26] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 239–248.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3174–3182, 2017.

[29] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4009–4018.

[30] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3828–3838.

[31] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, "Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance," in *European Conference on Computer Vision*. Springer, 2020, pp. 582–600.

[32] H. Jung, E. Park, and S. Yoo, "Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 642–12 652.

[33] J. Watson, M. Firman, G. J. Brostow, and D. Turmukhambetov, "Self-supervised monocular depth hints," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2162–2171.

[34] S. Zhao, H. Fu, M. Gong, and D. Tao, "Geometry-aware symmetric domain adaptation for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9788–9798.

[35] X. Song, W. Li, D. Zhou, Y. Dai, J. Fang, H. Li, and L. Zhang, "Mlda-net: Multi-level dual attention-based network for self-supervised monocular depth estimation," *IEEE Transactions on Image Processing*, vol. 30, pp. 4691–4705, 2021.

[36] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8001–8008.

[37] Y. Chen, C. Schmid, and C. Sminchisescu, "Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7063–7072.

[38] X. Luo, J.-B. Huang, R. Szeliski, K. Matzen, and J. Kopf, "Consistent video depth estimation," *ACM Transactions on Graphics (ToG)*, vol. 39, no. 4, pp. 71–1, 2020.

[39] C. Shu, K. Yu, Z. Duan, and K. Yang, "Feature-metric loss for self-supervised learning of depth and egomotion," in *European Conference on Computer Vision*. Springer, 2020, pp. 572–588.

[40] Y. Kuznietsov, M. Proesmans, and L. Van Gool, "Comoda: Continuous monocular depth adaptation using past experiences," in *Proceedings of*

the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 2907–2917.

[41] V. Patil, W. Van Gansbeke, D. Dai, and L. Van Gool, "Don't forget the past: Recurrent depth estimation from monocular video," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6813–6820, 2020.

[42] K. Li, Z. Fu, H. Wang, Z. Chen, and Y. Guo, "Adv-depth: self-supervised monocular depth estimation with an adversarial loss," *IEEE Signal Processing Letters*, vol. 28, pp. 638–642, 2021.

[43] J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, and M. Firman, "The temporal opportunist: Self-supervised multi-frame monocular depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1164–1174.

[44] X. Weng and K. Kitani, "Monocular 3d object detection with pseudo-lidar point cloud," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[45] G. Wang, X. Tian, R. Ding, and H. Wang, "Unsupervised learning of 3d scene flow from monocular camera," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4325–4331.

[46] H. Deng, G. Wang, Z. Feng, C. Jiang, X. Wu, Y. Miao, and H. Wang, "Pseudo-lidar for visual odometry," *arXiv preprint arXiv:2209.01567*, 2022.

[47] A. Nicolai, R. Skeele, C. Eriksen, and G. A. Hollinger, "Deep learning for laser based odometry estimation," in *RSS workshop Limits and Potentials of Deep Learning in Robotics*, vol. 184, 2016, p. 1.

[48] M. Velas, M. Spanel, M. Hradis, and A. Herout, "Cnn for imu assisted odometry estimation using velodyne lidar," in *2018 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. IEEE, 2018, pp. 71–77.

[49] C. Zheng, Y. Lyu, M. Li, and Z. Zhang, "Lodonet: A deep neural network with 2d keypoint matching for 3d lidar odometry estimation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2391–2399.

[50] W. Wang, M. R. U. Saputra, P. Zhao, P. Gusmao, B. Yang, C. Chen, A. Markham, and N. Trigoni, "Deeppco: End-to-end point cloud odometry through deep parallel neural network," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 3248–3254.

[51] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2024–2039, 2015.

[52] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European conference on computer vision*. Springer, 2016, pp. 740–756.

[53] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6647–6655.

[54] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang, "Learning monocular depth by distilling cross-domain stereo networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 484–500.

[55] V. Patil, C. Sakaridis, A. Liniger, and L. Van Gool, "P3depth: Monocular depth estimation with a piecewise planarity prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1610–1621.

[56] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1983–1992.

[57] T. Shen, Z. Luo, L. Zhou, H. Deng, R. Zhang, T. Fang, and L. Quan, "Beyond photometric loss for self-supervised ego-motion estimation," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6359–6365.

[58] J. Zhang, Q. Su, B. Tang, C. Wang, and Y. Li, "Dpsnet: Multitask learning using geometry reasoning for scene depth and semantics," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[59] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, "Learning depth from monocular videos using direct methods," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2022–2030.

[60] Y. Zou, Z. Luo, and J.-B. Huang, "Df-net: Unsupervised joint learning of depth and flow using cross-task consistency," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 36–53.

[61] Y. Almalioglu, M. R. U. Saputra, P. P. De Gusmao, A. Markham, and N. Trigoni, "Ganvo: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks," in *International conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 5474–5480.

[62] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black, "Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 240–12 249.

[63] C. Luo, Z. Yang, P. Wang, Y. Wang, W. Xu, R. Nevatia, and A. Yuille, "Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2624–2641, 2019.

[64] G. Wang, C. Zhang, H. Wang, J. Wang, Y. Wang, and X. Wang, "Unsupervised learning of depth, optical flow and pose with occlusion from 3d geometry," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 308–320, 2020.

[65] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," *Advances in neural information processing systems*, vol. 32, 2019.

[66] H. Li, A. Gordon, H. Zhao, V. Casser, and A. Angelova, "Unsupervised monocular depth learning in dynamic scenes," in *Conference on Robot Learning*. PMLR, 2021, pp. 1908–1917.

[67] F. Tosi, F. Aleotti, P. Z. Ramirez, M. Poggi, S. Salti, L. D. Stefano, and S. Mattoccia, "Distilled semantics for comprehensive scene understanding from videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4654–4665.

[68] S. Chen, Z. Pu, X. Fan, and B. Zou, "Fixing defect of photometric loss for self-supervised monocular depth estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1328–1338, 2021.

[69] S. Jia, X. Pei, X. Jing, and D. Yao, "Self-supervised 3d reconstruction and ego-motion estimation via on-board monocular video," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 7557–7569, 2021.

[70] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2485–2494.

[71] J. Wang, G. Zhang, Z. Wu, X. Li, and L. Liu, "Self-supervised joint learning framework of depth estimation via implicit cues," *arXiv preprint arXiv:2006.09876*, 2020.

[72] T.-W. Hui, "Rm-depth: Unsupervised learning of recurrent monocular depth in dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1675–1684.

[73] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.