Self-supervised Monocular Depth and Visual Odometry Learning with Scale-consistent Geometric Constraints

Mingkang Xiong, **Zhenghong Zhang**, **Weilin Zhong**, **Jinsheng Ji**, **Jiyuan Liu** and **Huilin Xiong***
Shanghai Key Laboratory of Intelligent Sensing and Recognition, Shanghai Jiao Tong University, Shanghai, China
{mkxiong, art_zzh, zhongweilin, jinshengji, liujiyuan, hlxiong}@sjtu.edu.cn

Abstract

The self-supervised learning-based depth and visual odometry (VO) estimators trained on monocular videos without ground truth have drawn significant attention recently. Prior works use photometric consistency as supervision, which is fragile under complex realistic environments due to illumination variations. More importantly, it suffers from scale inconsistency in the depth and pose estimation results. In this paper, robust geometric losses are proposed to deal with this problem. Specifically, we first align the scales of two reconstructed depth maps estimated from the adjacent image frames, and then enforce forward-backward relative pose consistency to formulate scale-consistent geometric constraints. Finally, a novel training framework is constructed to implement the proposed losses. Extensive evaluations on KITTI and Make3D datasets demonstrate that, i) by incorporating the proposed constraints as supervision, the depth estimation model can achieve state-of-the-art (SOTA) performance among the self-supervised methods, and ii) it is effective to use the proposed training framework to obtain a uniform global scale VO model.

1 Introduction

The depth and pose estimation from images is essential for many applications, such as augmented reality and self-driving cars in robotics and computer vision. Traditional methods are mainly hand-crafted features systems. With the progress of deep learning, depth can be predicted from a single image by Convolutional Neural Network (CNN) in a supervised paradigm [Eigen *et al.*, 2014], [Laina *et al.*, 2016], [Li *et al.*, 2018a]. However, these methods are limited by requiring large amounts of labeled data and it is challenging to collect ground truth depth especially in outdoor environments. Without the need for annotated depth, self-supervised monocular depth estimation methods from stereo images have been proposed in [Garg *et al.*, 2016] and [Godard *et al.*, 2017]. [Zhou *et al.*, 2017] shows a promising direction that depth and pose estimation models can be jointly trained from single view video sequences without ground truth. These self-supervised methods mainly use view synthesis [Zhou *et al.*, 2016] from adjacent frames as supervision. Previous works [Zhou *et al.*, 2017], [Wang *et al.*, 2018] utilize photometric error or gradient-based losses to penalize the differences among these synthetic views. However, these appearance losses are fragile in illumination variations environments and lack relevant constraints to get global scale consistent visual odometry results as proposed in [Bian *et al.*, 2019]. In this paper, to tackle these issues, we leverage more geometric information to formulate supervision beyond using appearance losses. The proposed method is also based on a self-supervised learning framework of depth and pose estimation from monocular videos. Different from the recent work [Bian *et al.*, 2019] directly using the predicted depth, we first derive the reconstructed depth maps estimated from consecutive frames and then use a simple but efficient approach to enforce the depth scale consistency. Besides depth reconstruction loss, we simultaneously consider the forward-backward relative poses error in our geometric loss function to get more accurate results. Depth and pose networks are jointly trained by the proposed loss functions and tested separately in the respective task. Qualitative evaluation on the KITTI Eigen test split is shown in Fig. 1 and the depth estimation results of the proposed method have significant improvement. Extensive qualitative and quantitative evaluation experiments are conducted in Sec. 4.

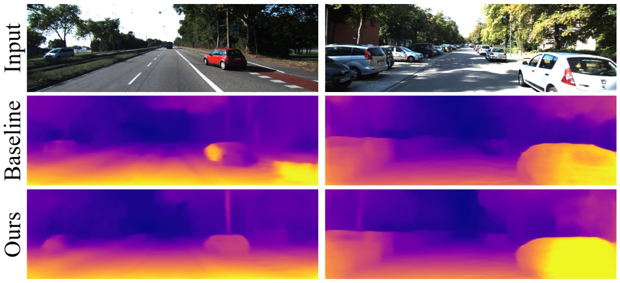


Figure 1: Qualitative comparison samples between the baseline method and our approach on the KITTI Eigen test split. By incorporating the proposed scale-consistent geometric loss, the depth estimation model achieves better results.

Our contributions are summarized as follows: i) we pro-

[*] Corresponding Author

pose novel scale-consistent geometric constraints by simultaneously considering depth reconstruction error and forward-backward relative poses consistency as supervision; ii) a novel self-supervised learning framework with the proposed loss functions is presented to get accurate depth estimation and scale-consistent visual odometry results; iii) the proposed models can achieve not only SOTA depth estimation performance on KITTI and Make3D datasets but also SOTA pose estimation results on the KITTI Odometry dataset comparing with self-supervised methods.

2 Related Work

Traditionally, depth and pose estimation are mostly solved by hand-crafted features methods, which are the basics of many SLAM and SfM algorithms. ORB-SLAM [Mur-Artal *et al.*, 2015] is typically a visual SLAM system that is based on ORB [Rublee *et al.*, 2011] features. [Schonberger and Frahm, 2016] revisits SfM algorithms and develops the COLMAP system. Recently, with the development of deep learning, CNNs have been successfully applied to estimate the monocular depth and ego-motion.

2.1 Supervised Methods

[Eigen *et al.*, 2014] firstly introduces deep learning to estimate single-view depth in a supervised fashion. Their network can refine the coarse global prediction locally. [Laina *et al.*, 2016] proposes a fully convolutional residual network to model the mapping between RGB images and depth maps. They use the berHu loss and propose a new up-sampling method. [Li *et al.*, 2018a] tackles monocular depth estimation as a multi-category formulation. They fuse the outputs from their dilated CNN in a hierarchical way and utilize the soft-weighted-sum inference to get continuous depth results.

For VO, [Konda and Memisevic, 2015] firstly develops a deep-learning-based VO system. The approach predicts changes in velocity and direction by using the softmax layer. [Wang *et al.*, 2017] formulates VO as a sequence learning problem. Historical information is introduced to infer current relative motions through a recurrent convolutional network. [Xue *et al.*, 2019] presents a VO framework composed of Memory and Refining components. To distill features from previous results in the Memory unit, they adopt a spatial-temporal attention mechanism to model the Refining component.

These methods are supervised by ground truth and therefore have the limitations of demanding enormous labeled data for training.

2.2 Self-supervised Methods

By leveraging view synthesis [Zhou *et al.*, 2016] as supervision, the depth and pose estimators can be jointly learned in a self-supervised paradigm. [Zhou *et al.*, 2017] proposes a novel framework for estimating depth and ego-motion using monocular video sequences. The framework consists of two separate deep networks that use the photometric error as a supervisory signal without depth or pose ground truth. Following this work, by considering the consistency of consecutive 3D point clouds, [Mahjourian *et al.*, 2018] introduces an ICP loss for aligning 3D structures to further improve the depth and pose estimation with combining 2D photometry-based losses. UnDeepVO [Li *et al.*, 2018b] can infer absolute scale results from monocular image sequences by utilizing stereo image pairs as training datasets. DDVO [Wang *et al.*, 2018] combines the traditional direct visual odometry with the learning framework to refine the depth estimation results. However, these methods still suffer from the scale inconsistent problem as proposed in recent work [Bian *et al.*, 2019]. Similarly, [Bian *et al.*, 2019] also considers geometric constraints to cope with this issue. Different from their work, we explicitly align the depth scale and enforce pose consistency in the proposed constraints.

These self-supervised learning methods have shown a promising way for single view depth and pose estimation without ground truth. The details of our loss functions are described in Sec. 3 and comparisons with SOTA approaches are shown in Sec. 4.

3 Method

In this section, we first describe the main idea behind self-supervised monocular depth and pose estimation, and then introduce our framework and loss functions. Finally, we present the baseline and final learning methods.

3.1 Problem Formulation

To learn depth and pose from monocular videos without ground truth, following previous methods [Zhou *et al.*, 2017], [Mahjourian *et al.*, 2018], [Bian *et al.*, 2019], we use photometric reprojection error as supervision. Given a pixel coordinate point $\mathbf{x} = (x, y)^{\mathrm{T}}$ and its estimated depth $d = d(\mathbf{x})$ predicted from the depth network, we can reconstruct its 3D point $\mathbf{X} = (X, Y, Z, 1)^{\mathrm{T}}$ based on the pinhole camera model by the inverse projection function

$$\pi^{-1}(\mathbf{x}, d(\mathbf{x})) = \left(\frac{(x - c_x)d}{f_x}, \frac{(y - c_y)d}{f_y}, d, 1\right)^{\mathrm{T}} \tag{1}$$

where f_x, f_y represent focal lengths and c_x, c_y stand for optical centers. The projection function is performed as

$$\pi(\mathbf{X}) = \left(\frac{Xf_x}{d} + c_x, \frac{Yf_y}{d} + c_y\right)^{\mathrm{T}}. \tag{2}$$

Considering two consecutive frames $\{I_{t-1}, I_t\}$, we can use the relative pose $\mathbf{T} = \mathbf{T}_{t\rightarrow t-1}$ predicted by pose network to get warping transform function

$$\omega(\mathbf{x}, \mathbf{T}) = \pi(\mathbf{T}\pi^{-1}(\mathbf{x}, d(\mathbf{x}))). \tag{3}$$

With this warping transformation, the synthesized images $I_{t-1}(\omega(\mathbf{x}^{i,j}, \mathbf{T}))$ are generated by using a differentiable bilinear sampling mechanism [Jaderberg *et al.*, 2015] and photometric consistency loss is formulated as

$$L_{\mathrm{ph}} = \frac{1}{N}\sum_{i,j}\left|I_{t-1}(\omega(\mathbf{x}^{i,j}, \mathbf{T})) - I_t(\mathbf{x}^{i,j})\right| \tag{4}$$

where N is the number of valid points.

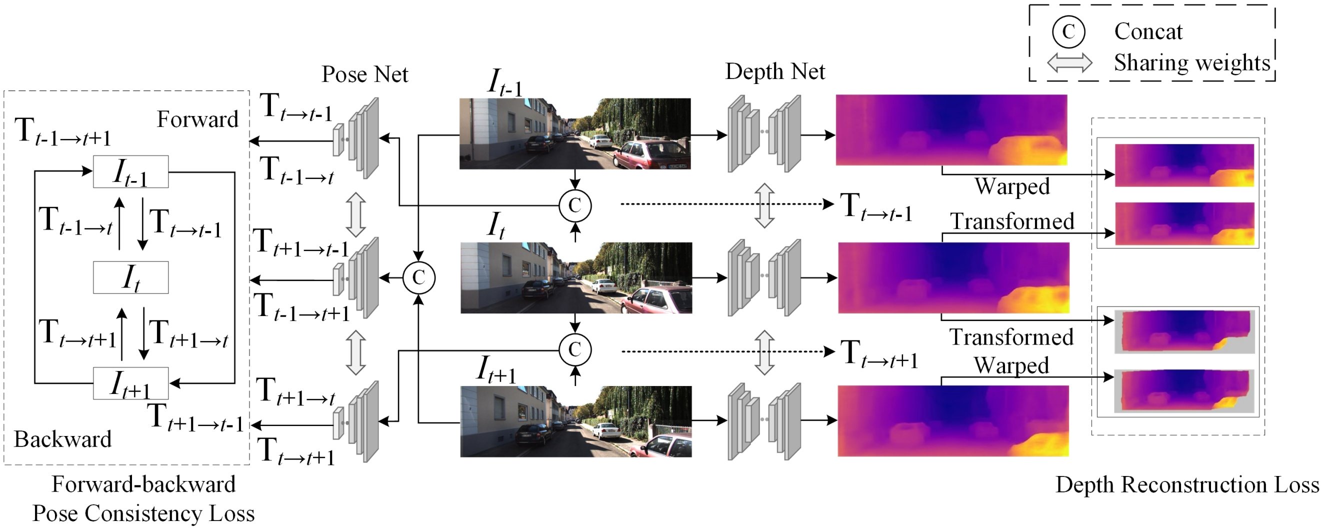


Figure 2: The main idea of our method. Given three sequential frames $\{I_{t-1}, I_t, I_{t+1}\}$, each depth map is predicted by depth net with sharing weights. The concat of image pairs is the input of pose net and its output relative poses $\mathbf{T}_{t\to t-1}, \mathbf{T}_{t\to t+1}$ are used to obtain the synthesized and reconstructed depth map. After aligning the depth scale, the scale-consistent depth reconstruction loss can be formulated as supervision to train neural networks. By applying the reversed image sequence, we can get the constraint of backward relative poses. With the combination of our scale-consistent geometric constraint and appearance losses proposed in the previous works, depth and pose networks can be trained by these self-supervised signals.

3.2 Scale-consistent Geometric Constraints

Besides minimizing photometric inconsistency, in this section, we consider decreasing depth maps and pose differences among consecutive frames as geometric supervision. The details of our loss functions will be introduced respectively.

Scale-consistent Depth Reconstruction Loss

We use $\{d_{t-1}, d_t\}$ to represent the estimated depth of images $\{I_{t-1}, I_t\}$. Similarly to photometric error, warping function Eq. 3 is used to get synthesized depth map $d^{'}_{t-1}$. Its equation is shown as follows:

$$d^{'}_{t-1} = d_{t-1}(\omega(\mathbf{x}, \mathbf{T})). \qquad (5)$$

Rather than directly utilizing d_t predicted from depth net, the transformation matrix $\mathbf{T}$ is applied to get the reconstructed depth map $d^{'}_{t}$ [Kerl *et al.*, 2013]. The equation is given as

$$d^{'}_{t} = [\mathbf{T}\pi^{-1}(\mathbf{x}, d_t(\mathbf{x}))]_d \qquad (6)$$

where $[\cdot]_d$ represents d component of a 3D point.

Learning depth only from monocular sequences suffers from scale inconsistent problem [Bian *et al.*, 2019], that is to say, the estimated depth maps from different images could have a different scale. Therefore, to penalize the inconsistent between synthesized and reconstructed depth maps, the scale inconsistent issue is reconsidered here. In details, we firstly align the depth scale by using their mean $\{\bar{d'}_{t-1}, \bar{d'}_t\}$ to normalize depth maps, and then apply a term to penalize the error between $\bar{d'}_{t-1}$ and $\bar{d'}_t$ to keep scale consistent. Finally, our scale-consistent depth reconstruction loss is written as

$$L_{\mathrm{d}} = \frac{1}{N}\sum_{i,j}\left|\frac{d^{'i,j}_{t-1}}{\bar{d'}_{t-1}} - \frac{d^{'i,j}_{t}}{\bar{d'}_{t}}\right| + \left|\bar{d'}_{t-1} - \bar{d'}_{t}\right| \qquad (7)$$

where N represents the number of valid points in the depth map.

Forward-backward Pose Consistency Loss

We use the vectors $\mathbf{r}$ and $\mathbf{t}$ to represent the rotational and translational items of the relative pose $\mathbf{T}$. Inspired by [Godard *et al.*, 2017] which uses mirror consistency between left and right views, we propose forward-backward pose constraint to get more accurate results. As Fig. 2 shows, two concat images are the input of pose network. The relative poses $\{\mathbf{T}_{t\to t-1}, \mathbf{T}_{t+1\to t}, \mathbf{T}_{t+1\to t-1}\}$ can be predicted from the images pairs $\{\{I_{t-1}, I_t\}, \{I_t, I_{t+1}\}, \{I_{t-1}, I_{t+1}\}\}$. If the estimated relative poses are accurate, the product of $\mathbf{T}_{t\to t-1}$ and $\mathbf{T}_{t+1\to t}$ should be equal to $\mathbf{T}_{t+1\to t-1}$ and vice versa. For simplicity, we use the rotational and translation vectors to represent the estimated relative poses. Our forward-backward pose consistency loss is formulated as

$$\begin{aligned} L_{\mathrm{po}} = \sum_{m=1,-1} & \|\mathbf{r}_{t\to t-m} + \mathbf{r}_{t+m\to t} - \mathbf{r}_{t+m\to t-m}\|_1 \\ & + \|\mathbf{t}_{t\to t-m} + \mathbf{t}_{t+m\to t} - \mathbf{t}_{t+m\to t-m}\|_1 \qquad (8) \end{aligned}$$

where m represents the forward-backward operation.

Note that the relative poses have been used to get the synthesized and reconstructed depth maps in Eq. 5, 6. So the scale inconsistent problem does not need to be considered in pose estimation. The forward-backward relative poses have been predicted by pose network, therefore the forward-backward idea can be easily extended to depth reconstruction loss function. For simplicity, we do not show this operation in our depth reconstruction loss and our scale-consistent geometric constraint is given as

$$L_{\mathrm{SC}} = \lambda_{\mathrm{d}} L_{\mathrm{d}} + \lambda_{\mathrm{po}} L_{\mathrm{po}}. \qquad (9)$$

where empirical parameters λ_{d} and λ_{po} have been used in our formulation and more details can be found in Sec. 4.2.

3.3 Total Training Loss and Framework

Similar to existing work [Godard *et al.*, 2017], [Mahjourian *et al.*, 2018], [Bian *et al.*, 2019], structural similarity (SSIM) [Wang *et al.*, 2004] loss has been used in our work to solve the complex illumination variations problem. Considering two images patches p_1 and p_2, $\mathrm{SSIM}(p_1, p_2)$ is defined as $\frac{(2\mu_1\mu_2+c_1)(2\sigma_{12}+c_2)}{(\mu_1^2+\mu_2^2+c_1)(\sigma_1^2+\sigma_2^2+c_2)}$. The loss function is given as

$$L_{\mathrm{SSIM}} = \frac{1 - \mathrm{SSIM}(I_{t-1}(\omega, \mathbf{T}), I_t)}{2}. \tag{10}$$

Following previous methods [Godard *et al.*, 2017], [Mahjourian *et al.*, 2018], [Bian *et al.*, 2019], it is assumed that the estimated depth maps should be locally smooth. An edge-aware term is used to weight the cost by using depth maps gradients. The depth smoothness loss is shown as follow:

$$L_{\mathrm{sm}} = \frac{1}{N}\sum_{i,j}\left|\nabla d_t^{ij}\right| e^{-\left|\nabla I_t^{ij}\right|}. \tag{11}$$

To demonstrate that the proposed constraint can contribute to the depth and pose estimation, we use photometric consistency loss with additional appearance losses as the baseline method which is represented by

$$L_{\mathrm{baseline}} = \lambda_{\mathrm{ph}} L_{\mathrm{ph}} + \lambda_{\mathrm{S}} L_{\mathrm{SSIM}} + \lambda_{\mathrm{sm}} L_{\mathrm{sm}}. \tag{12}$$

Via adding our proposed scale-consistent geometric constraint to baseline formulation, the total loss is written as

$$L_{\mathrm{total}} = L_{\mathrm{baseline}} + L_{\mathrm{SC}}. \tag{13}$$

As shown in Fig. 2, our depth network is based on the DispNet [Zhou *et al.*, 2017] which is an encoder-decoder architecture. We replace the original encoder-decoder convolutional blocks with residual blocks [He *et al.*, 2016]. The input of this network is an RGB image and the output is a disparity map (the inverse of a depth map). For pose network, we adopt the architecture PoseNet proposed in [Bian *et al.*, 2019]. The network uses a concat of two images as input then output rotational and translational vectors of the relative pose. In Fig. 2, we present the main idea of our proposed scale-consistent geometric constraint. Three sequential images are formulated as a training sample and their depth maps are predicted respectively. The relative poses from image pairs are utilized in Eq. 5 and 6. To get backward pose results, the image sequences need to be reversed.

4 Experiments

4.1 Datasets

Our models are mainly trained on KITTI datasets [Geiger *et al.*, 2012]. For monocular depth estimation, we use Eigen split [Eigen *et al.*, 2014] of Raw data for a fair comparison with previous methods. The split selects 697 images as test datasets for monocular depth estimation and the others are applied for training. The ground truth depth maps of test datasets are obtained from lidar sensors. The original image size is 1242×375 and we resize it as 416×128 or 832×256 to formulate training datasets. For pose estimation, we train our networks on the KITTI Odometry dataset [Geiger *et al.*, 2012], which contains 11 sequences with public ground truth poses. The sequences 00-08 are utilized for training and 09-10 are test sets. The ground truth poses are not used in our training framework. Note that for the depth or pose evaluation, we separately train our networks on two datasets (Raw data and Odometry data). Because there are overlapping scenes in these datasets.

As presented in previous works [Zhou *et al.*, 2017], [Bian *et al.*, 2019], depth and pose estimation results can be improved by pretraining on Cityscapes [Cordts *et al.*, 2016] datasets. To compare with SOTA methods, we pre-train networks on Cityscapes and finetune on KITTI with the parameters in Sec. 4.2.

The Make3D datasets [Saxena *et al.*, 2009] are used to evaluate the generalization ability of the depth estimation model. We directly test our models on this set without training or fine-tuning. The datasets contain 534 single-view images with ground truth depth maps and 134 images are used for evaluation.

4.2 Implementation Details

We use PyTorch [Paszke *et al.*, 2019] to implement our framework and train it with a TITAN XP GPU. Adam optimizer is adopted and parameters are set as $\beta_1 = 0.9$ and $\beta_2 = 0.999$. In Eq. 12 and 13, the combination $[\lambda_{\mathrm{ph}}, \lambda_{\mathrm{S}}, \lambda_{\mathrm{sm}}, \lambda_{\mathrm{d}}, \lambda_{\mathrm{po}}] = [0.15, 0.85, 0.1, 0.001, 0.1]$ is used. We utilize three sequential frames as training samples and batch size is set to 4. The learning rate is set as 10^{-4}. Our models are trained in 100 epochs and we randomly select 1000 samples in every epoch. We pre-train the model on Cityscapes and then finetune KITTI datasets. The data is augmented with random brightness, contrast, and saturation.

4.3 Depth Estimation Results

Our method is evaluated on standard Eigen split [Eigen *et al.*, 2014] by using the metric proposed in [Eigen *et al.*, 2014]. As shown in Tab. 1, we compare the depth estimation results with SOTA self-supervised methods. These models are trained on KITTI or Cityscapes datasets. The evaluation results of previous works are taken from their public papers. Two maximum depth range 80m and 50m have been used in the evaluation and most models are tested at 80m depth range. Without the utilization of traditional direct visual odometry, the proposed method outperforms DDVO [Wang *et al.*, 2018]. Different from the methods using complex auxiliary tasks [Yin and Shi, 2018] [Ranjan *et al.*, 2019] or complicated networks [Vankadari *et al.*, 2019], our approach mainly focus on leveraging geometric information. It is important to note that although the models vid2depth [Mahjourian *et al.*, 2018] and SC-SfMLearner [Bian *et al.*, 2019] have considered the depth consistency in geometry, our methods not only mitigate depth scale inconsistency but also take into account relative pose consistency during the training phase and thus our model can achieve more accurate depth estimation results.

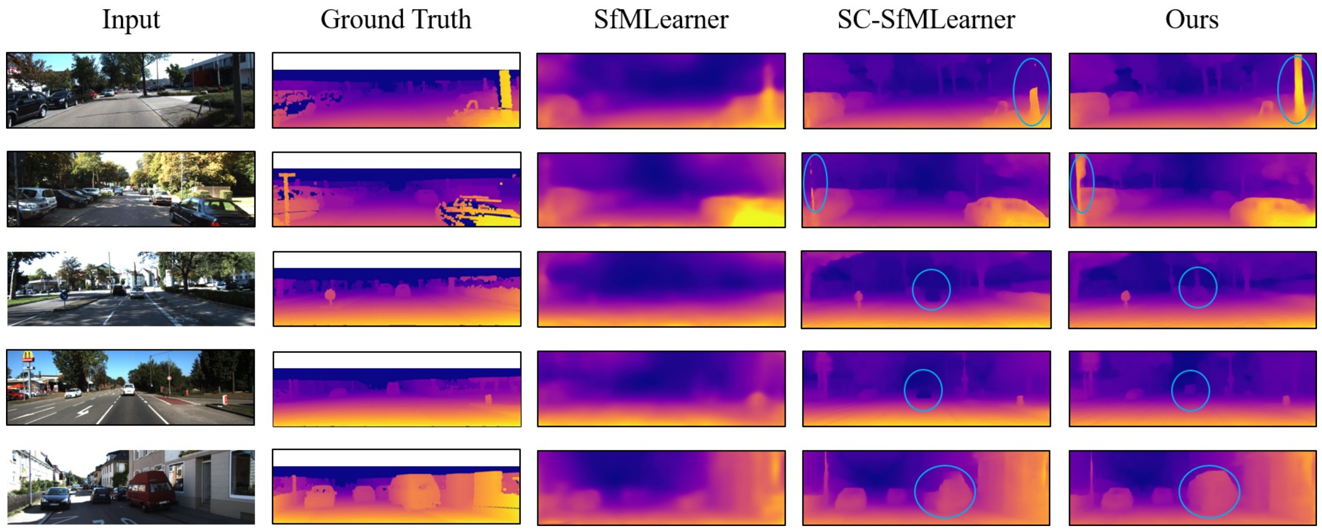


Figure 3: Qualitative comparison samples on the KITTI Eigen test split. The ground truth maps are got from sparse laser data by maximum filtering only for visualization. Comparing with other state-of-art methods, our model can predict more robust depth estimation results (black holes in previous work) and more details in the scene.

Methods	Cap	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
		Error (lower is better)				Accuracy (higher is better)		
SfMLearner [Zhou *et al.*, 2017]	80m	0.198	1.836	6.565	0.275	0.718	0.901	0.960
vid2depth [Mahjourian *et al.*, 2018]		0.159	1.231	5.912	0.243	0.784	0.923	0.970
GeoNet [Yin and Shi, 2018]		0.153	1.328	5.737	0.232	0.802	0.934	0.972
DDVO [Wang *et al.*, 2018]		0.148	1.187	5.496	0.226	0.812	0.938	0.975
UnDepth [Vankadari *et al.*, 2019]		0.127	0.998	5.309	0.226	0.827	0.934	0.971
CC [Ranjan *et al.*, 2019]		0.139	1.032	5.199	0.213	0.827	0.943	0.977
SC-SfMLearner [Bian *et al.*, 2019]		0.128	1.047	5.234	0.208	0.846	0.947	0.976
Ours		**0.126**	**0.902**	**5.052**	**0.205**	**0.851**	**0.950**	**0.979**
SfMLearner [Zhou *et al.*, 2017]	50m	0.190	1.436	4.975	0.258	0.735	0.915	0.968
vid2depth [Mahjourian *et al.*, 2018]		0.151	0.949	4.383	0.227	0.802	0.935	0.974
UnDepth [Vankadari *et al.*, 2019]		0.121	0.749	4.051	0.214	0.840	0.941	0.975
Ours		**0.119**	**0.681**	**3.815**	**0.192**	**0.866**	**0.957**	**0.981**

Table 1: Monocular depth estimation results on the KITTI Eigen test split. Two parts of this table show the depth results capped by 80m and 50m. δ represents the ratio of estimated depth and ground truth.

Methods	Datasets	Resolutions	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
			Error (lower is better)				Accuracy (higher is better)		
Baseline	K	416×128	0.157	1.235	5.700	0.237	0.789	0.930	0.972
Total w/o L_{po}			0.152	1.149	5.519	0.231	0.800	0.930	0.972
Total w/o L_d			0.148	1.116	5.603	0.228	0.803	0.932	0.973
Total			**0.148**	**1.077**	**5.506**	**0.228**	**0.806**	**0.934**	**0.973**
Total	K	832×256	0.140	1.061	5.309	0.219	0.823	0.940	0.976
Total	CS+K		**0.126**	**0.902**	**5.052**	**0.205**	**0.851**	**0.950**	**0.979**

Table 2: Ablation studies on monocular depth estimation. The results are evaluated on KITTI Eigen split and are capped at 80m. Two types of image resolutions have been conducted in our experiments. K denotes that our models are only trained on KITTI and CS+K means fine-tuning networks on KITTI with pre-trained parameters on Cityscapes.

Methods	Abs Rel	Sq Rel	RMSE	RMSE log
SfMLearner [Zhou *et al.*, 2017]	0.383	5.321	10.47	0.478
DDVO [Wang *et al.*, 2018]	0.387	4.720	8.09	0.204
SfMLearner (updated)	0.361	3.680	7.749	0.181
SC-SfMLearner [Bian *et al.*, 2019]	0.337	3.302	7.162	0.171
Ours	**0.320**	**3.170**	**7.062**	**0.163**

Table 3: Monocular depth estimation results on the Make3D dataset. The model trained on Cityscapes and KITTI datasets is directly used to predict the test set. Except for the first two row results are attained from the paper, the others are tested with online provided models by the same metrics.

Methods	Metric	Seq. 09	Seq. 10
SfMLearner [Zhou *et al.*, 2017]	$t_{err}(\%)$	17.84	37.91
	$r_{err}(^\circ/100\text{m})$	6.78	17.78
Depth-VO-Feat [Zhan *et al.*, 2018]	$t_{err}(\%)$	11.93	12.45
	$r_{err}(^\circ/100\text{m})$	3.91	3.46
SC-SfMLearner [Bian *et al.*, 2019]	$t_{err}(\%)$	8.24	10.7
	$r_{err}(^\circ/100\text{m})$	2.19	4.58
Ours	$t_{err}(\%)$	**5.85**	**10.11**
	$r_{err}(^\circ/100\text{m})$	**1.73**	**3.89**

Table 4: Visual odometry estimation results on the KITTI Odometry dataset. The models are tested on sequence 09 and 10. t_{err} and r_{err} respectively represents average translation and rotation error.

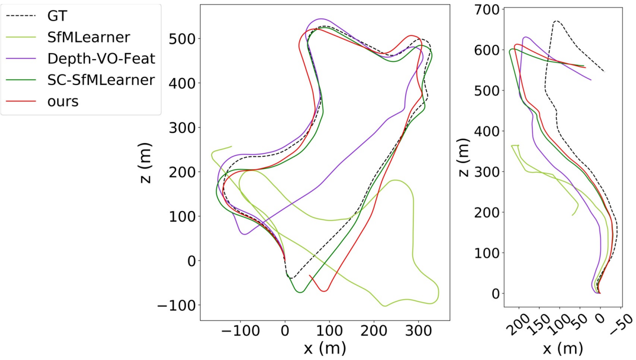
Our model is directly tested on Make3D datasets without finetuning to verify the generalization ability. In Tab. 3, we compare our models with the online public SOTA models. In this unseen environment, the proposed method still has significant improvement compared with the other methods.

4.4 Pose Estimation Results

Although depth and pose nets are jointly learned during training, they are an independent model in the testing phase. Frame-to-frame pose estimation results without post-processing have been integrated over sequence 09 and 10 of the KITTI Odometry dataset. As shown in Tab. 4, we evaluate average translation and rotation error of (100,200,...,800) meters sub-sequences by using the standard evaluation metrics proposed in [Geiger *et al.*, 2012]. Fig. 4 shows the qualitative results of trajectories. Our estimated trajectory scale only needs to be aligned once with ground truth rather than aligning every estimated pose results scale as SfMLearner [Zhou *et al.*, 2017]. Different from Depth-VO-Feat [Zhan *et al.*, 2018] using a stereo sequence for training, only monocular videos are utilized in the proposed method. By simultaneously considering pose and aligned depth consistency, our model achieves better pose estimation results than SC-SfMLearner [Bian *et al.*, 2019].

4.5 Ablation Studies

We present ablation studies to show the importance of our methods in depth estimation. The experiments are conducted on KITTI Eigen split with variations of our models and use the same metrics mentioned in Sec. 4.3.



Sequence 09 **Sequence 10**

Figure 4: Visual odometry trajectories on sequence 09 and 10 of KITTI Odometry split.

Tab. 2 presents the quantitative results and Fig. 1 shows qualitative comparisons. Our baseline method is similar to the basic approach of [Bian *et al.*, 2019] and is based on the loss proposed in Eq. 12. We apply the network architecture mentioned in Sec. 3.3 to ablation studies. It can be found that, without any contributions of our methods, the baseline model performs worst. The combination of depth reconstruction and pose consistency loss proposed in Eq. 9 makes a significant improvement. Excluding one of our loss functions decline the performance of depth estimation. Moreover, we use larger image resolution 832×256 and pre-train our models on the Cityscapes dataset to get better results.

5 Conclusion

In this paper, we present a self-supervised learning framework of depth and pose estimation with scale-consistent geometric constraints. Beyond using photometric consistency as supervision, robust geometric information is utilized and scale alignment operation is also conducted in the proposed method. The scale-consistent depth and visual odometry results can be predicted by our models, which is essential for many applications such as obstacle avoidance. Comparing with prior self-supervised learning frameworks, the proposed monocular depth and pose estimation models obtain SOTA results on KITTI and Make3D datasets. To the best of our knowledge, it is the first work to simultaneously use depth and pose consistency to formulate the self-supervised signal from monocular video sequences.

Moreover, little current work uses higher-level features as supervision. It is interesting to fuse semantic knowledge with the proposed method for future work. Traditional VO systems have been widely used in many applications, we will try to incorporate self-supervised learning ideas into traditional VO or visual SLAM to get more robust and accurate pose estimation results.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 61673274.

References

[Bian *et al.*, 2019] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *NeurIPS*, 2019.

[Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[Eigen *et al.*, 2014] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014.

[Garg *et al.*, 2016] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[Godard *et al.*, 2017] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NeurIPS*, 2015.

[Kerl *et al.*, 2013] Christian Kerl, Jürgen Sturm, and Daniel Cremers. Dense visual slam for rgb-d cameras. In *IROS*, 2013.

[Konda and Memisevic, 2015] Kishore Reddy Konda and Roland Memisevic. Learning visual odometry with a convolutional network. In *VISAPP (1)*, 2015.

[Laina *et al.*, 2016] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016.

[Li *et al.*, 2018a] Bo Li, Yuchao Dai, and Mingyi He. Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference. *Pattern Recognition*, 83:328–339, 2018.

[Li *et al.*, 2018b] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. In *ICRA*, 2018.

[Mahjourian *et al.*, 2018] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 2018.

[Mur-Artal *et al.*, 2015] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[Ranjan *et al.*, 2019] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *CVPR*, 2019.

[Rublee *et al.*, 2011] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, 2011.

[Saxena *et al.*, 2009] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840, 2009.

[Schonberger and Frahm, 2016] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

[Vankadari *et al.*, 2019] Madhu Vankadari, Swagat Kumar, Anima Majumder, and Kaushik Das. Unsupervised learning of monocular depth and ego-motion using conditional patchgans. In *IJCAI*, 2019.

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

[Wang *et al.*, 2017] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *ICRA*, 2017.

[Wang *et al.*, 2018] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, 2018.

[Xue *et al.*, 2019] Fei Xue, Xin Wang, Shunkai Li, Qiuyuan Wang, Junqiu Wang, and Hongbin Zha. Beyond tracking: Selecting memory and refining poses for deep visual odometry. In *CVPR*, 2019.

[Yin and Shi, 2018] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018.

[Zhan *et al.*, 2018] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, 2018.

[Zhou *et al.*, 2016] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *ECCV*, 2016.

[Zhou *et al.*, 2017] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.