Can Scale-Consistent Monocular Depth Be Learned in a Self-Supervised Scale-Invariant Manner?

Lijun Wang[1], Yifan Wang[1,*], Linzhao Wang,[2] Yunlong Zhan,[2] Ying Wang,[2] and Huchuan Lu[1,3,*]
[1]Dalian University of Technology, [2]Huawei Technologies Co., Ltd., [3]Peng Cheng Lab
{ljwang,wyfan,lhchuan}@dlut.edu.cn, {wanglinzhao,zhanyunlong,wangying110}@huawei.com

Abstract

Geometric constraints are shown to enforce scale consistency and remedy the scale ambiguity issue in self-supervised monocular depth estimation. Meanwhile, scale-invariant losses focus on learning relative depth, leading to accurate relative depth prediction. To combine the best of both worlds, we learn scale-consistent self-supervised depth in a scale-invariant manner. Towards this goal, we present a scale-aware geometric (SAG) loss, which enforces scale consistency through point cloud alignment. Compared to prior arts, SAG loss takes relative scale into consideration during relative motion estimation, enabling more precise alignment and explicit supervision for scale inference. In addition, a novel two-stream architecture for depth estimation is designed, which disentangles scale from depth estimation and allows depth to be learned in a scale-invariant manner. The integration of SAG loss and two-stream network enables more consistent scale inference and more accurate relative depth estimation. Our method achieves state-of-the-art performance under both scale-invariant and scale-dependent evaluation settings.

1. Introduction

To alleviate the need of high-quality ground truth depth data, there is a recent surge of interest in self-supervised monocular depth estimation [36, 10]. The basic idea is to jointly learn depth estimation and ego-motion prediction supervised by a photometric reconstruction loss. Although, these approaches have achieved remarkable success in popular benchmarks, they are known to suffer from the per-frame scale ambiguity issue [27, 2]. For one thing, the estimated depths are not guaranteed to be scale consistent and the ego-motion network also fails to predict globally consistent trajectories for long videos. For another, without proper constraints, the depth network has to adapt its scales according to the ego-motion prediction and vice versa, which confuses network training and results in performance degeneration or even training divergence [27]. Meanwhile, recent evidence [7] also indicates that the global scale is a fundamental source of uncertainty in supervised depth estimation, and that scale-invariant losses focus on relative depth learning and largely benefit depth estimation in terms of accuracy as well as generalization abilities [19].

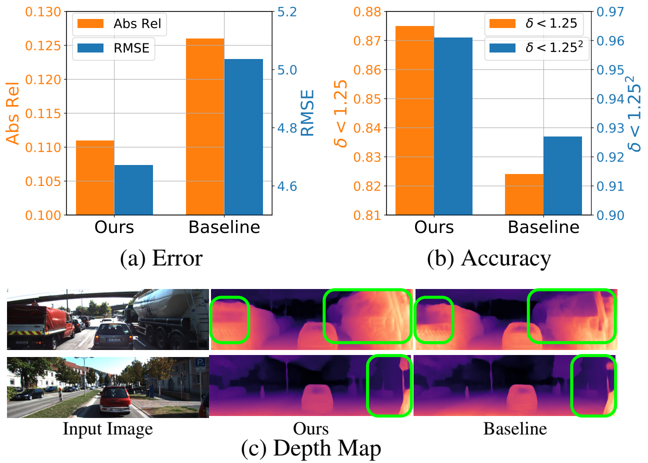

Figure 1. Strength of our scale-consistent depth estimation learned in a scale-invariant manner. (a) Scale-dependent evaluation on KITTI [9], where predictions are aligned to ground truth using one global scale on a per-sequence basis rather than conventional per-frame scale alignments (See Sec. 4.2.2). (b) Scale-invariant learning allows our method to produce more accurate relative depth.

In light of the above analysis, an interesting question to ask is *whether we can achieve scale-consistent depth estimation but also enjoy the advantages of scale-invariant training under the self-supervised framework?* We make the first attempt to answer this question by proposing a new paradigm for self-supervised monocular depth estimation.

To this end, we present a scale-aware geometric loss (dubbed SAG loss) that operates in the 3D space. The estimated depth of adjacent frames are first projected to 3D point clouds and then transformed into a common view using the predicted ego-motion. Instead of directly penal-

[*]Corresponding author

izing the coordinate differences between point clouds, we estimate their relative motion parameters in a least-square sense, which include not only the rotation and translation, but also the relative scale factor. The SAG loss are computed by incorporating the three motion parameters to enforce scale consistent depth estimation of adjacent frames. With iterative training, depth consistency can finally be propagated through entire sequences (See Fig. 1 (a)(b)).

In order to enjoy the benefits of scale invariant training, we propose to decompose the depth estimation task into two sub-tasks: normalized depth prediction and scale inference. Through careful designing, we ensure that the SAG loss is invariant to the scale of the normalized depth prediction and that scale inference can be learned in an explicit manner to guarantee depth scale consistency. We present a concrete implementation of the above idea through a new depth network with a two-stream architecture.

Learning scale-consistent depth in a scale-invariant manner seems to be contradictory at the first glance and becomes even more intractable under self-supervised frameworks, due to the scale ambiguity and lack of scale supervision. Thanks to the proposed SAG loss, scale supervision can be explicitly calculated during self-supervised learning to simultaneously ensure scale consistency and allow the disentanglement of depth and scale. As a matter of fact, the idea of taking scale into explicit consideration itself is shown to be beneficial to motion estimation when scale inconsistency indeed exists. In addition, we also explore a new strategy for finding corresponding points between point clouds, which further facilitates motion estimation, leading to a more effective SAG loss. By combining the SAG loss with our two-stream depth network, our method is able to take advantages of scale-invariant depth learning, giving rise to geometrically more consistent and quantitatively more accurate depth estimation (See Fig. 1 (c)).

The contribution of this work can be summarized into three folds.

- A new self-supervised depth estimation framework that enjoys the strengths of scale-invariant learning and delivers scale-consistent depth.
- A scale-aware geometric loss to enforce depth consistency and to provide supervisions for explicit scale inference during self-supervised learning.
- A two-stream depth network to disentangle depth and scale prediction, allowing the normalized depth to be learned irrespective of the global scale.

Experiments on KITTI datasets demonstrate that our method can not only improve depth accuracy but also benefit long-term ego-motion estimation. Extensive ablation studies have also been conducted, which further confirm the effectiveness of our contribution.

2. Related Work

In the deep learning era, fully-supervised CNN models have shown record-breaking performance [7, 18, 8, 29]. Self-supervised learning has also been highlighted by recent studies [12, 30, 16] to alleviate the needs of ground truth depth annotation. In the seminal work of Zhou *et al*. [36], self-supervised depth estimation is achieved in a purely monocular setting. Following this line of work, rapid progress [23, 34, 22, 17] has been made by exploring new architectures and training strategies to further improve the accuracy and robustness. For instance, some works [37, 32, 4, 34] propose to incorporate optical flow to handle moving objects, while others [14, 5] leverage semantic labels to guide self-supervised learning. Later on, the photometric loss is replaced by the deep feature reconstruction loss, where the deep features are either pre-trained [33] or jointly learned with depth networks [24]. In [11], Godard *et al*. propose a new appearance matching loss with automask techniques, which further closes the performance gap between stereo and monocular self-supervised depth estimation. Recently, a self-supervised depth estimation network with symmetrical 3D packing and unpacking blocks is designed in [13], which is shown to even outperform supervised counterparts.

Scale-Consistent Depth Learning. To ensure depth scale consistency in monocular self-supervised learning, a geometric consistency loss is proposed in [2], which directly minimizes the differences between depth predictions of consecutive frames. In a similar spirit, [4] enforces depth scale and structure consistency by projecting multi-view depth into the 3D space and penalizing the coordinate differences of corresponding points. In [13], the camera velocity is leveraged as an additional supervision to solve the scale ambiguity issue. In comparision, [25] uses bundle-adjusted scene structures and poses as supervision to learn more consistent depth estimation. Compared to the above methods, our unique contribution is to combine the benefits of scale-consistent depth estimation with scale-invariant learning under the self-supervised learning framework. The most related work to ours is [20] which proposes a 3D constraint to align point clouds through an approximate back-propagation algorithm. Our method differs from [20] mainly in three aspects. First, point clouds alignment in [20] is performed by estimating a 6-DOF transformation including rotation and translation, where we take scale into explicit consideration, which is shown to result in more accurate alignment and can be used to provide direct supervision for depth scale inference. Second, in [20] correspondences between 3D points are determined by using a closest point heuristic [1], which is solved in an iterative manner and can only provide local optimal. In comparison, we leverage the correspondences learned from view synthesis, which is more accurate and enable closed-form solutions for point cloud alignment. Finally, by combining the proposed two-stream network and SAG loss, our method can be trained in a scale-invariant manner, which can not only ensure scale-consistency, but also deliver more accurate relative depth structures.

Scale-Invariant Depth Learning. The advantage of scale-invariant training is first explored in supervised methods [7, 3, 31]. Eigen *et al.* [7] observe that the global depth scale is ambiguous in monocular images, and propose a scale-invariant error to learn relative depth irrespective of scales. The idea is further extended in [19] to improve the generalization ability across datasets with different scales. Later on, Wang *et al.* [28] propose a new architecture to disentangle depth and scale estimation for fully-supervised learning. In the self-supervised domain, Wang *et al.* [27] achieve scale-invariant learning through depth normalization. However, it is still an open question to leverage scale-invariant training while ensuring scale-consistency.

3. Self-Supervised Scale-Consistent Depth

This section will elaborate on our major contributions, *i.e.*, the two-stream depth network with disentangled scale inference and the scale-aware geometric loss to enforce scale consistency. We first revisit the principles of self-supervised depth estimation in Sec. 3.1 to introduce our motivation and notations. In Sec. 3.2 and 3.3, we present our network architectures and loss functions, respectively. Finally, Sec. 3.4 provides implementation details.

3.1. A Revisit to Self-Supervised Training

The primary idea behind self-supervised depth estimation from monocular videos is to cast the joint learning of depth and ego-motion network to a novel view synthesis problem. More formally, given a target frame I_t and a source frame I_s that are adjacent to each other, the depth D_t of I_t and the camera motion $M = [R, T]$ (with rotation matrix R and translation T) from the source to the target frame can be estimated using the depth and ego-motion networks, respectively. The target frame depth can be projected into a point cloud P_t as follows,

$$P_t^{ij} = K^{-1} D_t^{ij} \; [i, j, 1]^\intercal, \tag{1}$$

where K denotes the camera intrinsics; $[i, j, 1]$ indicates the homogeneous coordinate of a pixel at location $[i, j]$ of the image plane; while P_t^{ij} and D_t^{ij} represent the corresponding 3D point and depth of that pixel.

With the predicted camera motion, we can transform the point cloud P_t to that of the source frame $\hat{P}_s = RP_t + T$, and then project the point cloud back to the source image plane. The whole process can be represented as follows[1].

$$[\hat{i}, \hat{j}, 1]^\intercal \sim KRD_t^{ij} K^{-1} [i, j, 1]^\intercal + KT, \tag{2}$$

where $[\hat{i}, \hat{j}]$ and $[i, j]$ are coordinates of corresponding pixels in the source and target frames, respectively. According to this pixel-level mapping, we can reconstruct the target frame using the source frame through bilinear warping.

The above view synthesis process is entirely differentiable, and the predicted depth and ego-motion are involved as intermediate variables. Therefore, the depth and ego-motion networks can be jointly trained by minimizing the photometric reconstruction error. Most existing approaches implement photometric loss with the combination of the L1 and SSIM loss:

$$L_P = \frac{\alpha}{2}(1 - \text{SSIM}(\hat{I}_t, I_t)) + (1 - \alpha)\|\hat{I}_t - I_t\|_1, \tag{3}$$

where $\hat{I}_t$ are the reconstructed target frame and $\alpha = 0.85$.

An edge-aware gradient smoothness constraint has also been introduced in [12] to regularize the predicted depth:

$$L_S = \sum_{i,j} |\partial_x D_t^{ij}| e^{-|\partial_x I_t^{ij}|} + |\partial_y D_t^{ij}| e^{-|\partial_y I_t^{ij}|}. \tag{4}$$

It can be easily shown that the photometric loss in (3) is scale ambiguous to the joint prediction of depth and ego-motion. To demonstrate this, one can consider another set of prediction $D'_t = aD_t$ and $M' = [R, T']$ with $T' = aT$, which have different scales from the original prediction D_t and M. By substituting D_t and M with D'_t and M' in (2), the same pixel mapping is established, leading to the same photometric loss. As a consequence of the scale ambiguity issue, the learned depth and ego-motion are not scale consistent across one video sequence. In addition, the depth and ego-motion networks have to learn to co-adapt their scales, which potentially confuses network learning and even leads to training divergence.

3.2. Disentangled Depth and Scale Estimation

Our main goal is to alleviate the above scale ambiguity issue. Meanwhile, as scale-invariant learning is shown to benefit fully-supervised depth estimation, we also expect that this advantage can be transferred to the self-supervised domain. Our first step towards this goal is to disentangle scale inference from depth estimation in the network architecture level. As such, one part of our network can focus on learning to predict accurate relative depth irrespective of scales, while the other part is able to explicitly learn depth scale inference to ensure scale consistency as described.

We implement the above idea through the design of a two-stream depth network as shown in Fig. 2. It consists

[1] For notation conciseness, we omit the detailed conversion steps to the homogeneous coordinates

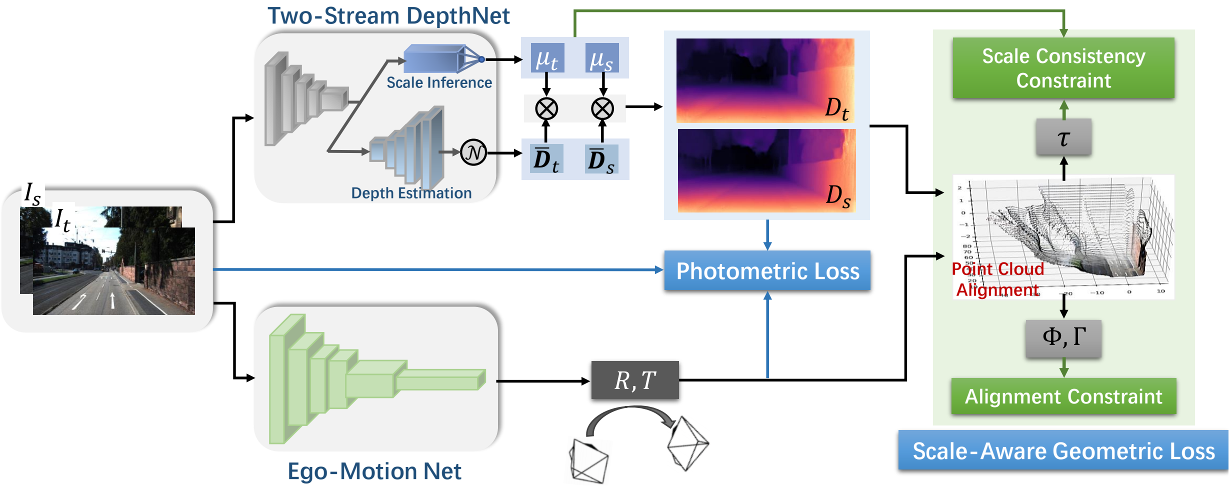

Figure 2. Pipeline of our proposed method.

of the depth estimation and scale inference stream building on top of a shared backbone network. Following prior arts, we adopt the ResNet18 network [15] as our backbone, which can already achieve satisfactory results but is more lightweight and efficient. We believe other more sophisticated networks can also meet our purpose. Given an input frame, the backbone network generates a multi-scale feature pyramid (*i.e.*, the output feature maps produced by Res2-Res5 stages of ResNet18), which serves as the input to the depth estimation stream.

In each feature level, the depth estimation stream first process the input feature map with a standard 3×3 convolution layer. The processed feature map is then combined with that from the last level through concatenation followed by another 3×3 convolution layer. Finally, the combined feature is upsampled with nearest neighbor interpolation and fed to the next level. The above procedure is progressively conducted from the coarsest to the finest feature level, producing the output feature map. An additional convolution layer takes the produced feature map as input to generate the one-channel depth output. We further explore different strategies to normalize the output depth as follows,

$$\bar{D} = \hat{D}/m, \tag{5}$$

where $\hat{D}$ denotes the output depth of our depth estimation stream; m indicates either the mean or median value of the output depth; and $\bar{D}$ denotes the normalized depth. As a result, the normalized depth is independent to the global scale of the depth stream output, and therefore the depth stream can be learned in a scale-invariant manner.

In parallel to the depth estimation stream, the scale inference stream consumes the coarsest output of backbone (*i.e.*, output feature of Res5 in ResNet18), and consists of four 3×3 convolution layers followed by a global average pooling layer. Finally, a Sigmoid unit acts as the output layer to produce a depth scale μ for the input frame. Given the normalized depth and scale, we compute the final depth by multiplying them $D = \mu\bar{D}$. Further training with our SAG loss allows that the final depth is scale consistent.

For ego-motion estimation, we adopt the architecture proposed in [11], which modifies a ResNet18 backbone to accept a pair of RGB frames as input and predict the 6-DoF relative motion, including the rotation with an axis-angle representation and the translation. Please refer to [11] for more architecture details.

3.3. Scale Aware Geometric Loss

A straightforward idea to enforce scale consistency is to directly penalize the disagreement between the estimated depth of adjacent frames. Our scale-aware geometric (SAG) loss also adheres to this principle but operates on the point clouds. As shown in the following, our SAG loss together with the two-stream depth network can not only ensure scale consistency, but also retain the benefits of scale-invariant depth learning.

Recall that in Sec 3.1, we project the predicted depth of target and source frames into point clouds P_t and P_s, respectively, and convert P_t to the source view $\hat{P}_s$ using the predicted ego-motion. To measure the disagreement between the predicted target and source depth, we first align the point cloud $\hat{P}_s$ with P_s through a least-square estimation of transformation parameters between them. Since $\hat{P}_s$ and P_s are scale inconsistent under most circumstances, the transformation parameters consist of not only rotation Φ and translation Γ, but also a scale factor τ, to ensure more precise point alignment. Let us assume that the correspondences between two point clouds are given, the least-square estimation can be formally described as follows,

$$\Phi, \Gamma, \tau = \arg\min_{\tilde{\Phi},\tilde{\Gamma},\tilde{\tau}} \sum_i \left\| \tilde{\tau}\tilde{\Phi}\hat{P}_s^i + \tilde{\Gamma} - P_s^{N(i)} \right\|_2^2, \tag{6}$$

where i and $N(i)$ indicates the indices of two corresponding points in point cloud $\hat{P}_s$ and P_s. The above least-square estimation can be solved in a closed-form as shown in [26], and the solutions can be further represented as differentiable functions of the input point clouds, allowing gradient backward-propagation. Our SAG loss is defined using the estimated transformation parameters and comprises an alignment constraint and a scale consistency constraint.

The alignment constraint is used to enforce the accuracy of relative depth and ego-motion prediction irrespective of the scale. If the predicted relative depth and ego-motion are accurate, the scaled point cloud $\tau\hat{P}_s$ should already be perfectly aligned with P_s. Otherwise, the estimated rotation Φ and transformation Γ in (6) will imply their misalignment. Therefore, our alignment constraint penalizes the inaccurate predictions by forcing the estimated transformation Φ and Γ to approximate an identity mapping.

Our scale consistency constraint pursues a more direct supervision for depth scale inference. Let us denote μ_t and μ_s as the depth scales of target and source frame predicted by our scale inference stream. According to the point cloud alignment (6), the scaled point cloud $\tau\hat{P}_s$ (with depth scale $\tau\mu_t$) is already scale-consistent to P_s (with depth scale μ_s). Therefore, for scale-consistent predictions, the ideal depth scales of the target and source frames should be $\tau k\mu_t$ and $k\mu_s$, respectively, up to an unknown factor k, where the estimated scale factor τ embodies the inconsistency between the predicted scales, and servers as an amendment to the target scale. By treating the ideal scales as our objectives, we eliminate the unknown factor through division and define our scale consistency constraint as the differences between the predicted and objective scale ratios of the two frames:

$$\begin{aligned} L_C &= \left\| \frac{\mu_t}{\mu_s} - \frac{\dot{\tau} k \dot{\mu}_t}{k \dot{\mu}_s} \right\|_1 \\ &= \left\| \frac{\mu_t}{\mu_s} - \frac{\dot{\tau} \dot{\mu}_t}{\dot{\mu}_s} \right\|_1, \end{aligned} \tag{7}$$

where the notation $\dot{x}$ indicates that x is used as a constant to compute the ground truth and its gradient backward-propagation is disabled. We provide more detailed derivation and explanation to interpret the scale consistency constraint in the supplementary material.

By combining the alignment and scale consistency constraints, our SAG loss can be described as follows,

$$L_{SAG} = \|\Phi - E\|_1 + \|\Gamma\|_1 + \left\| \frac{\mu_t}{\mu_s} - \frac{\dot{\tau}\dot{\mu}_t}{\dot{\mu}_s} \right\|_1, \tag{8}$$

where E denotes the identity matrix.

Until now, one remaining problem is how to obtain the correspondences $\big(i, N(i)\big)$ between point clouds $\hat{P}_s$ and P_s in order to estimate the transformation in (6). In [20], this is approached by the iterative closet points (ICP) method [1] that alternates between finding correspondences and the least-square estimation of transformation parameters. However, the correspondences are found merely based on a closest point heuristic and the photometric appearance information is left unused. In our preliminary experiments, we find that the ICP algorithm is computational inefficient and can only deliver local optimums, especially when the two point clouds are scale inconsistent. Therefore, we propose to use the pixel-level correspondences established by (2), which eliminates the need of iterative estimation. Since these correspondences are directly learned through view synthesis, they are more accurate than those only relying on closet point heuristics. As training proceeds, the depth and ego-motion network becomes stronger, giving rise to more accurate correspondences. However, one may still concern that the correspondences established by (2) may still be inaccurate, particularly during the initial training stage. As detailed in Sec. 3.4, we remedy this issue by exploring a selection mechanism to perform least-square estimation using only reliable correspondences.

Discussion. Compared to prior methods, our SAG loss requires explicit estimation of depth scales, and thus the name "scale-aware". Our experiments show that taking scale into account can facilitate more accurate translation estimation (6), and thus more effective 3D constraints. Besides, we further leverage the estimated scale factor to construct the scale consistency constraint (7) which provides direct and explicit supervision for depth scale inference. By training our disentangled depth and scale prediction network using our SAG loss, we can finally achieve scale consistency while retain the advantages of scale-invariant training.

3.4. Implementation

We initialize the backbones of depth and ego-motion network using ResNet18 pre-trained on ImageNet [6]. To balance efficiency with precision, our networks operate in an input resolution of 640 × 192 pixels, though higher resolutions are shown to further boost accuracy. Our final loss function combines the photometric loss, the smoothness constraint, and the proposed SAG loss:

$$L = L_P + \lambda_1 L_{SAG} + \lambda_2 L_S, \tag{9}$$

where the loss weights are empirically set to $\lambda_1 = 0.05$ and $\lambda_2 = 0.001$. During training, we predict depth and compute the loss values at multiple scales following prior works. To address issues caused by occlusion, out of view, and static camera, we adopt the strategies proposed in [11]. Specifically, instead of computing the averaged photometric losses over all pixels, we compare, at each pixel, the losses a) computed by warping the target frame to all source views and b) computed by using original target frame. A per-pixel mask can then be obtained indicating potentially valid pixels. We use this mask to weight the photometric loss and to select

Method	Year	Error ↓				Accuracy ↑		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SfMLearner [36]	CVPR 2017	0.198	1.836	6.565	0.275	0.718	0.901	0.960
DDVO *et al.* [27]	CVPR 2018	0.151	1.257	5.583	0.228	0.810	0.936	0.974
Mahjourian *et al.* [20]	CVPR 2018	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Zhan *et al.* [33]	CVPR 2018	0.135	1.132	5.585	0.229	0.820	0.933	0.971
DF-Net [37]	ECCV 2018	0.146	1.182	5.215	0.213	0.818	0.943	0.978
Bian *et al.* [2]	NeurIPS 2019	0.137	1.089	5.439	0.217	0.830	0.942	0.975
CC [23]	CVPR 2019	0.140	1.070	5.326	0.217	0.826	0.941	0.975
Zhou *et al.* [35]	ICCV 2019	0.121	0.837	4.945	0.197	0.853	0.955	0.982
Monodepth2 [10]	ICCV 2019	0.115	0.903	4.863	0.193	0.877	0.959	0.981
SGDepth [17]	ECCV 2020	0.113	0.835	4.693	0.191	0.879	0.961	0.981
pRGBD-Refined [25]	ECCV 2020	0.113	0.793	4.655	0.188	0.874	0.960	**0.983**
DeaFet [24]	CVPR 2020	0.126	0.925	5.035	0.200	0.862	0.954	0.980
Johonston *et al.* [16]*	CVPR 2020	0.111	0.941	4.817	0.189	**0.885**	0.961	0.981
PackNet-SfM [13]	CVPR 2020	0.111	0.785	**4.601**	0.189	0.878	0.960	0.982
Ours	–	**0.109**	**0.779**	4.641	**0.186**	0.883	**0.962**	0.982

Table 1. Comparison on KITTI benchmark. The best and second best methods are in **bold** and underlined, respectively. * denotes results achieved with the ResNet18 backbone for fair comparison.

correspondences for point cloud alignment (6). We adopt a sequence length of three frames for training, with the central frame as the target view and the rest as the source views. The Adam optimizer is used to learn network parameters with an initial learning rate of $1e-4$ for the first 20 epochs and $1e-5$ for another 15 epochs. Data augmentation strategies including random color jittering and horizontal flipping has also been adopted to improve generalization abilities.

4. Experiments

4.1. Evaluation of Monocular Depth

We evaluate our method on KITTI benchmark [9] by following the training protocol established by Eigen *et al.* [7]. We also adopt Zhou *et al.*'s [36] pre-processing strategy to remove static frames, giving rise to 3910 monocular triplets for training, 4424 for validation and 697 for testing. Except the ImageNet pre-trained backbones, we do not perform any additional pre-training on depth datasets. Source code and pre-trained models will be released at https://bit.ly/3m8GFON.

The comparison results between our method and state-of-the-art approaches are shown in Tab. 1. Unless otherwise specified, all compared methods are trained using the same protocol with same input resolutions. Since self-supervised learning cannot restore accurate scales, we compute all the metrics after scale alignment to ground truth on a per-frame basis. It can be shown that our method can consistently outperform other compared methods in terms of all metrics. Among others, [20] also uses 3D constraints for consistent depth estimation. Unlike ours, their approaches do not take scale into explicit account. Our two-stream depth network adopts the same backbone architecture with Godard *et al.* [10]. The additional overhead brought by our scale inference stream is very limited. However, the improvement of our method over [10] is substantial, which verifies the strength of scale-invariant training enabled by our two-stream architecture as well as the SAG loss. Fig. 3 further visualizes the predicted depth and point cloud reconstruction using our method.

4.2. Ablation Analysis on Monocular Depth

To understand the impact of learning scale-consistent depth in scale-invariant manner, we conduct ablation studies on KITTI dataset. We have also evaluated the performance of our method with stronger backbones and higher input resolutions. Since these are not our main contributions, they are included in the supplementary material.

4.2.1 Scale-Invariant Training

We compare 5 variants of our method to analyze the strength of scale-invariant training in self-supervised depth estimation. Among them, Baseline only contains the depth estimation stream, while Baseline+MN further normalizes each predicted depth map with its mean value. Two-stream adopts the same architecture as ours with disentangled depth and scale estimation. Meanwhile, Two-stream+MN and Two+stream+MdN normalize the predicted depth with mean and median depth values, respectively. All the above methods are trained using the conventional photometric loss with smoothness constraints. Comparison results are shown in Tab. 2 (a). Compared with Baseline, Baseline+MN learns depth estimation in a scale-invariant manner. However, its improvement over Baseline is marginal, which is not consistent to our knowledge obtained from the self-supervised

	Method	Scale-Disen.	Scale-Inv.	Scale-Consis.	Error ↓		Accuracy ↑
					Abs Rel	RMSE	$\delta < 1.25$
(a)	Baseline				0.118	4.956	0.862
	Baseline+MN		✓		0.117	4.956	0.864
	Two-stream	✓			0.116	4.954	0.863
	Two-stream+MN	✓	✓		0.112	4.689	0.878
	Two-stream+MdN	✓	✓		0.115	4.864	0.862
	Two-stream+MN+SAG-w/o-scale	✓	✓	✓	0.111	4.689	0.880
	Two-stream+MN+SAG-ICP	✓	✓	✓	0.115	4.854	0.865
	Two-stream+MN+SAG	✓	✓	✓	**0.109**	**4.641**	**0.883**
(b)	Baseline				0.126	5.035	0.824
	Two-stream+MN	✓			0.124	4.956	0.826
	Two-stream+MN+SAG-w/o-scale	✓	✓	✓	0.114	4.723	0.870
	Two-stream+MN+SAG	✓	✓	✓	**0.111**	**4.672**	**0.875**

Table 2. Ablation study on KITTI benchmark. (a) Scale-invariant evaluation, where scales of predicted depths are aligned to ground-truth for each frames. (b) Scale-dependent setting, where predictions are aligned to ground-truth with the same scale factor per-sequence. Scale-Disen., Scale-Inv., and Scale-Consis. indicate that the method is trained in scale-disentangled, scale-invariant, and scale-consistent manner, respectively. MN and MdN denote mean-value and median-value normalization, respectively. The best results are in **bold**.

methods. One possible reason may be that depth and ego-motion networks have to co-adapt their scales in self-supervised learning. Directly normalizing depth will confuse ego-motion network which further affects the training of depth estimation. Since the model size of Two-stream is comparable to Baseline, their performances are also similar when learned with the same training strategy. Through output mean normalization, Two-stream+MN not only disentangles scale inference from depth estimation, but also ensures depth estimation to be learned in a scale-invariant manner, which significantly improves depth accuracy under self-supervised learning. By comparing Two-stream+MN and Two-stream+MdN, it is clear that mean-value normalization is more superior than median-value normalization.

4.2.2 Scale-Consistent Training

On top of the Two-stream+MN variant, we investigate different constraints to understand the impact of scale-consistent training. As shown in Tab. 2 (a), SAG-w/o-scale denotes a simplified version of our SAG loss, which does not consider scale during point cloud alignment and only consists of the alignment constraint. SAG-ICP performs point cloud alignment with the iterative ICP algorithm. The performances of Two-stream with and without SAG-w/o-scale are similar, which is reasonable due to the scale-invariance of our evaluation process, *i.e.*, the per-frame scale alignment to ground-truth depth is conducted before computing metrics. Nonetheless, the improvement of SAG loss over SAG-w/o-scale is still considerable, suggesting that taking scale into explicit consideration can deliver more precise point cloud alignment, thus more superior depth estimation performance. Meanwhile, the performance of SAG-ICP is not satisfactory, indicating point correspondences learned from view-synthesis is more accurate than its counterparts based on the closest point heuristic.

Method	Sequence 9	Sequence 10	Frames
ORB-Slam [21]	0.014±0.008	0.012±0.011	–
SfMLearner [36]	0.021±0.017	0.020±0.015	5
DF-Net [37]	0.017±0.007	0.015±0.009	5
CC [23]	**0.012±0.007**	**0.012±0.008**	5
DDVO [27]	0.045±0.108	0.033±0.074	3
Mahjourian [20]	0.013±0.010	0.012±0.011	3
Monodepth2 [11]	0.017±0.008	0.015±0.010	2
Baseline	0.020±0.010	0.016±0.011	2
Ours	0.014±0.008	0.014±0.010	2

Table 3. Comparison on KITTI Odometry benchmark. The best results are in **bold** font.

The above evaluation process is scale-invariant and is thus not in favor of our SAG loss. To further demonstrate the power of our SAG loss, we replace the per-frame depth scale alignment with the per-sequence alignment before computing all metrics, *i.e.*, we align the depth scale for each frame of one video sequence with the same scale factor computed using the ground-truth depth. As shown in Tab. 2 (b), the performance gain brought by our SAG loss becomes even more significant, which confirms the effectiveness of our SAG loss in maintaining scale consistency.

4.3. Evaluation of Ego-Motion

Since the depth network is jointly trained with the ego-motion network, their performances are reliant to each other. To further confirm the effectiveness of our method, evaluation results on KITTI Odometry benchmark are reported in Tab. 3. Following [11], our method is trained on

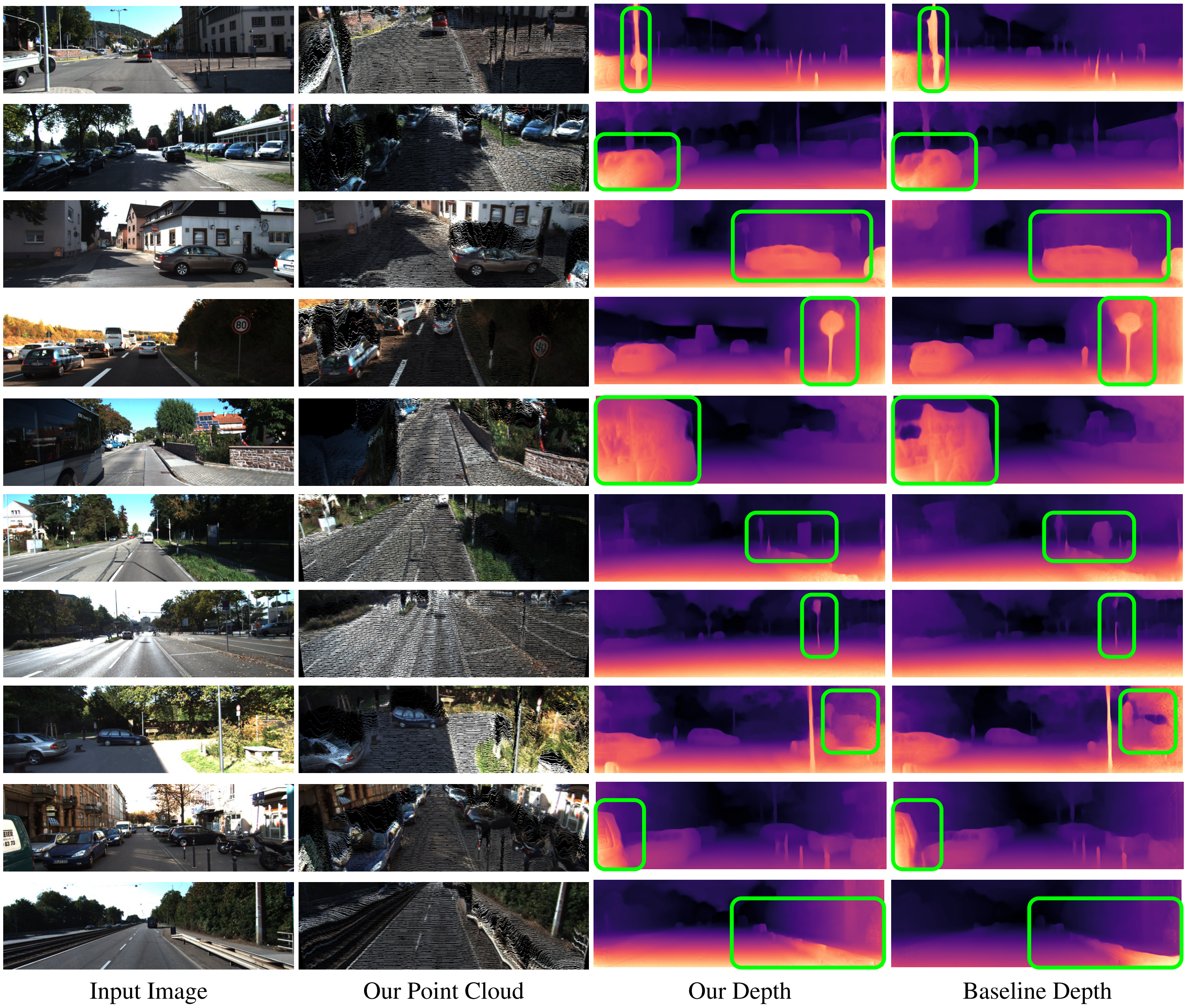

Figure 3. Visual comparison of our method and Baseline. The predicted depth maps of our method are perceptually more accurate with more details. The point clouds reconstructed based on our predictions are also visually plausible. Best viewed in color and zoom in.

sequences 0-8 and evaluated on sequences 9 and 10. The absolute trajectory error is averaged over all overlapping five-frame snippets in the test sequences. Although our ego-motion network accepts only two consecutive frames as input to predict their relative motion, we still compare favorably against existing methods. Besides, our ego-motion network adopts the exactly same architecture as that of [11]. The performance gain of our method is therefore solely brought by our proposed two-stream depth network trained using our SAG loss.

5. Conclusion

We propose a self-supervised depth estimation method, which can ensure scale consistency while enjoying the advantages of scale-invariant learning. The core design of our method is the two-stream depth network and the scale-aware geometric (SAG) loss. On the one hand, the network disentangles scale inference from depth estimation, allowing depth to be learned in a scale-invariant manner. On the other hand, the SAG loss explicitly estimates relative scale factor during 3D geometric alignment, providing direct supervision for consistent scale inference. Experiments on KITTI depth and odometry dataset verify our contribution.

Acknowledgment. This work is supported by National Natural Science Foundation of China (61906031, 62006036, U1903215, 61725202, 61829102), Fundamental Research Funds for Central Universities (DUT21RC(3)025, DUT20RC(3)017), and Dalian Innovation Leaders Support Plan (2018RD07).

References

[1] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606, 1992. 2, 5

[2] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Advances in neural information processing systems*, pages 35–45, 2019. 1, 2, 6

[3] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Adv. Neural Inform. Process. Syst.*, pages 730–738, 2016. 3

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

[5] Bin Cheng, Inderjot Singh Saggu, Raunak Shah, Gaurav Bansal, and Dinesh Bharadia. S3 net: Semantic-aware self-supervised depth estimation with monocular videos and synthetic data. In *Eur. Conf. Comput. Vis.*, pages 52–69, 2020. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009. 5

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Adv. Neural Inform. Process. Syst.*, pages 2366–2374, 2014. 1, 2, 3, 6

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

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research*, 2013. 1, 6

[10] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *Int. Conf. Comput. Vis.*, pages 3827–3837, 2019. 1, 6

[11] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *CVPR*, pages 3828–3838, 2019. 2, 4, 5, 7, 8

[12] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 270–279, 2017. 2, 3

[13] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2485–2494, 2020. 2, 6

[14] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. *arXiv preprint arXiv:2002.12319*, 2020. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 4

[16] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4756–4765, 2020. 2, 6

[17] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *Eur. Conf. Comput. Vis.*, 2020. 2, 6

[18] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision*, pages 239–248, 2016. 2

[19] Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019. 1, 3

[20] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5667–5675, 2018. 2, 5, 6, 7

[21] 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. 7

[22] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3227–3237, 2020. 2

[23] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12240–12249, 2019. 2, 6, 7

[24] Jaime Spencer, Richard Bowden, and Simon Hadfield. Defeat-net: General monocular depth via simultaneous unsupervised representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14402–14413, 2020. 2, 6

[25] Lokender Tiwari, Pan Ji, Quoc-Huy Tran, Bingbing Zhuang, Saket Anand, and Manmohan Chandraker. Pseudo rgb-d for self-improving monocular slam and depth prediction. In *Eur. Conf. Comput. Vis.*, 2020. 2, 6

[26] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, (4):376–380, 1991. 5

[27] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2022–2030, 2018. 1, 3, 6, 7

[28] Lijun Wang, Jianming Zhang, Oliver Wang, Zhe Lin, and Huchuan Lu. SDC-Depth: Semantic divide-and-conquer network for monocular depth estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020. 3

[29] Lijun Wang, Jianming Zhang, Yifan Wang, Huchuan Lu, and Xiang Ruan. Cliffnet for monocular depth estimation with hierarchical embedding loss. In *Eur. Conf. Comput. Vis.*, volume 12350, pages 316–331, 2020. 2

[30] Jamie Watson, Michael Firman, Gabriel J. Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Int. Conf. Comput. Vis.*, pages 2162–2171, 2019. 2

[31] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 611–620, 2020. 3

[32] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1983–1992, 2018. 2

[33] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2018. 2, 6

[34] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9151–9161, 2020. 2

[35] Junsheng Zhou, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. Unsupervised high-resolution depth learning from videos with dual networks. In *Int. Conf. Comput. Vis.*, pages 6872–6881, 2019. 6

[36] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1851–1858, 2017. 1, 2, 6, 7

[37] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Eur. Conf. Comput. Vis.*, pages 36–53, 2018. 2, 6, 7