DiPE: Deeper into Photometric Errors for Unsupervised Learning of Depth and Ego-motion from Monocular Videos

Hualie Jiang, Laiyan Ding, Zhenglong Sun, and Rui Huang*

Abstract— Unsupervised learning of depth and ego-motion from unlabelled monocular videos has recently drawn great attention, which avoids the use of expensive ground truth in the supervised one. It achieves this by using the photometric errors between the target view and the synthesized views from its adjacent source views as the loss. Despite significant progress, the learning still suffers from occlusion and scene dynamics. This paper shows that carefully manipulating photometric errors can tackle these difficulties better. The primary improvement is achieved by a statistical technique that can mask out the invisible or nonstationary pixels in the photometric error map and thus prevents misleading the networks. With this outlier masking approach, the depth of objects moving in the opposite direction to the camera can be estimated more accurately. To the best of our knowledge, such scenarios have not been seriously considered in the previous works, even though they pose a higher risk in applications like autonomous driving. We also propose an efficient weighted multi-scale scheme to reduce the artifacts in the predicted depth maps. Extensive experiments on the KITTI dataset show the effectiveness of the proposed approaches. The overall system achieves state-of-the-art performance on both depth and ego-motion estimation.

I. INTRODUCTION

The depth and ego-motion estimation is the core problem in Simultaneous Localization And Mapping (SLAM). Recently, Monocular Depth Estimation (MDE) attracts much attention, as it can be flexibly used in many applications, such as autonomous mobile robotics and AR/VR. Tracking the 6-DoF motion for a moving camera is also critical for these applications. Traditional supervised methods require expensively-collected ground truth, resulting in limited ability in generalization. By contrast, unsupervised learning from monocular videos [1] is a much more generalizable solution.

The unsupervised learning models usually contain two networks for predicting the depth map of the target view, and the motion between the target view and its temporally adjacent views. With the network output, the target view can be reconstructed by the adjacent source views with image warping, and the resulted photometric loss can be used as the supervisory signal for learning. However, the image reconstruction is usually destroyed by between-view occlusion and scene dynamics, as illustrated in Fig. 1, and the resulting incorrect supervision harms the network learning.
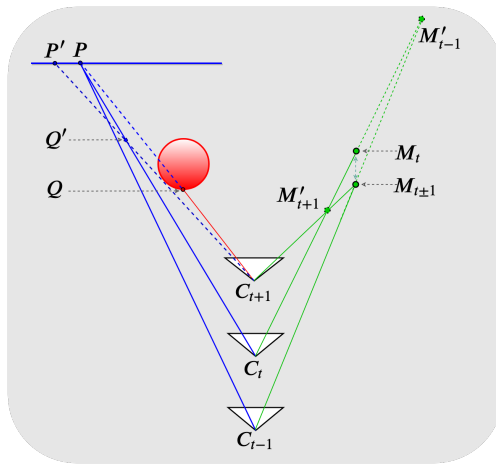
Fig. 1: **The detriment of occlusion and dynamics. (a) Occlusion:** point P is visible in C_t (camera view in time t) but occluded in C_{t+1}, and to achieve photometric consistency (matching P' instead of Q), P is estimated with shorter depth (Q'), making the foreground object blur. **(b) Co-directional motion:** if point M moves forward like camera C (from $t-1$ to t), it is likely estimated with farther depth (M'_{t-1}), producing 'dark holes'. **(c) Contra-directional motion:** when point M moves backward opposite to camera C (from t to $t+1$), it is estimated with shorter depth (M'_{t+1}).

The theory of how minimizing between-view reconstruction errors affects the depth estimation of occluded regions and the common forward and backward moving objects is illustrated in Fig. 1. Many methods have been proposed to cope with the occlusion and dynamics, and considerable improvement has been made. For example, the effect of 'dark holes' by the co-directionally moving objects has been tackled in the latest work [2], [3], [4]. However, as shown in Fig. 4 the latest models make significant underestimation of the depth for the contra-directionally moving objects. To the best of our knowledge, the inaccuracy of such objects has not been reported in the literature, which may cause trouble in practical applications. For instance, in autonomous driving, if the distance of oncoming cars is underrated, unnecessary braking or avoiding may be executed.

This issue can be largely avoided by our proposed outlier masking technique, which helps to exclude the occluded and moving regions, especially the oncoming objects. The technique is driven by our observation that the photometric errors of occluded and dynamic regions are much larger. In theory, the visible background usually dominates the scenes
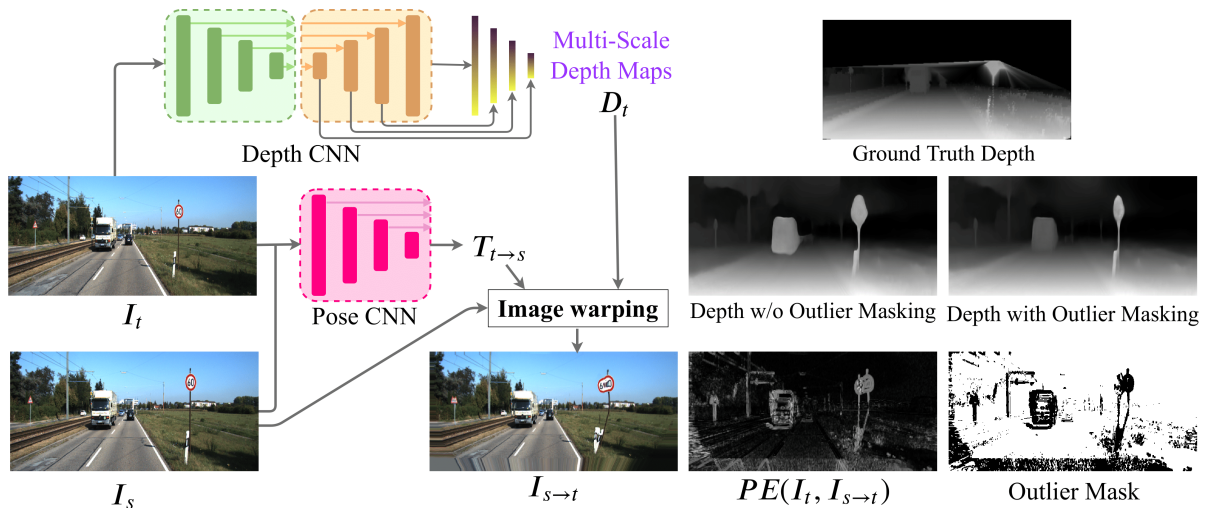
Fig. 2: **The Unsupervised Learning Flow and Effect of Outlier Masking. (a)** Depth CNN: A standard fully convolutional U-net that predicts the multi-scale depth maps for the target image. Pose CNN: A standard CNN that inputs the target view and one source view and predicts their relative motion. With D_t and $T_{t \to s}$ by the networks, the synthesized image $I_{s \to t}$ from the source view I_s to the target view I_t be differentiablly warped. The photometric errors between I_t and $I_{s \to t}$ can work as the training objective for both the Depth CNN and Pose CNN. **(b)** The outlier masking can exclude many invisible and nonstatic pixels, particularly those belonging to contra-moving objects, thus predicting a more accurate depth map. Without outlier masking, the oncoming vehicle is predicted to be very close, and the foreground object boundary significantly dilates.

and the invisible or moving pixels are inconsistent with the background, thus making their errors difficult to optimize. Besides, we also propose an efficient weighted multi-scale scheme to reduce artifacts and work with the outlier masking to produce better depth maps.

The effectiveness of our two main contributions, as mentioned above, is experimentally proven on the driving KITTI dataset. Together with a simple baseline model and some other masking practices, we build an overall state-of-the-art unsupervised monocular depth and ego-motion estimation system, called DiPE.

II. RELATED WORK

In recent years, deep Convolutional Neural Networks (CNN) have boosted the performance of MDE. One typical approach is using a deep CNN to densely regress the ground truth depth obtained with physical sensors [5], [6], [7], [8]. Other approaches can be categorized as combining deep learning with graphical models [9], [10], [11] or casting MDE as a dense classification problem [12], [13], [14]. However, models trained on publicly available datasets with ground truth depth, like the NYUDepthV2 [15] or KITTI [16], usually do not generalize well to real scenarios.

Instead of depending on ground truth, unsupervised learning schemes adopt more available resources, the stereo images [17], [18] or adjacent monocular video frames [1] to construct the supervisory signal. Specifically, the loss is the photometric difference between a view and its synthesis. The synthesis can be computed from the additional view by its estimated depth and the known or estimated pose between the two views. To take advantage of both spatial and temporal cues, stereo videos are exploited for training in [19], [20],

[2], [4]. Compared with stereo images, the monocular videos are more generalized and available, thus this paper focuses on the latter one.

The first method training with monocular videos, SfM-Learner [1] adopts an additional Pose CNN to estimate the relative motion between sequential views to make view synthesis attainable. However, the photometric consistency between nearby views is usually unsatisfied due to occlusion and moving objects. To improve this advantageous framework, many methods have been proposed, which can be mainly classified as: masking photometric errors [1], [4], [21], [22], joint learning with optical flow [2], [23], [24], modelling object motion [3], [25], [26]. The masking strategies also do not necessarily guarantee flexibility. Some masking techniques, such as the explainability mask [1] and the uncertainty map [21] also requires an extra network to learn. Joint learning with the optical flow has to construct a new network for learning optical flow to explain or compensate for the photometric inconsistency caused by occlusion or scene dynamics. Similarly, modelling object motion also requires additional modules to estimate the segmentation and motion of objects.

Different from the above methods, the overlap and blank masks geometrically derived from the image warping process [22] is a light-weight design for occlusion. A simpler method for occlusion is the minimum reprojection in Monodepth2 [4], which takes the minimum photometric errors from all source views, thus is also a masking technique. Monodepth2 also adopts a auto-masking technique for moving objects in a close speed with the camera. This simple and efficient masking strategy has been proved effective by Monodepth2, compared with other state-of-the-art methods. However, the

oncoming moving objects, have not been noticed and solved. The outlier masking method is proposed in this paper for such objects. Further, our outlier masking technique can help the minimum reprojection to recover a more accurate boundary for the foreground objects in predicted depth maps.

III. METHODOLOGY

A. Preliminaries

The monocular unsupervised learning scheme is shown in Fig. 2. A training sample contains the target frame I_t at time t and some source frames I_s at nearby times, $s \in \mathcal{S}$. Conventionally, $\mathcal{S} = \{t-1, t+1\}$ or $\{t-2, t-1, t+1, t+2\}$. Suppose that K is the shared intrinsic matrix of these frames. With the predicted depth D_t and transformation $T_{t,s}$, the synthesis from the source view s to the target view t can be expressed as,

$$I_{s \to t} = I_s \langle proj(D_t, T_{t \to s}, K) \rangle, \quad (1)$$

where $\langle \rangle$ is the differentiable bilinear sampling operator [27] and $proj()$ is the operation projecting the pixel p_t in the target image to the point p_s in the source image,

$$p_s \simeq K T_{t \to s} D(p_t) K^{-1} p_t, \quad (2)$$

where p_t and p_s are expressed in homogeneous coordinates.

In this paper, we adopt the popular combination of L1 and SSIM by [18] to compute the photometric errors,

$$\mathcal{PE}(I_a, I_b) = 0.85 \frac{1 - \text{SSIM}(I_a, I_b)}{2} + 0.15 \|I_a - I_b\|_1, \quad (3)$$

In addition, an edge-aware smoothness term is usually also applied in unsupervised training. We use the one by [4],

$$L_{es} = mean \left(|\partial_x d_t^*| e^{-|\partial_x I_t|} + |\partial_y d_t^*| e^{-|\partial_y I_t|} \right), \quad (4)$$

where $d_t^* = d_t / \overline{d_t}$ is the mean-normalized inverse depth from [28] to discourage shrinking of the estimated depth. Both losses are applied in 4 scales to avoid gradient locality.

B. Outlier Masking

As has been discussed, the image reconstruction can be damaged by some adverse factors, such as occlusion and scene dynamics. Therefore a portion of pixels in the photometric error map is invalid, and the incorporation of them in training can be misleading. We have the observation that most pixels are visible and stationary, and other occluded and moving pixels always produce more significant photometric errors. The outlier masking technique is based on this observation, which is simple but effective. The outlier mask is automatically determined by the statistical information of photometric errors. Specifically,, we first compute the mean and standard deviation of pixel photometric errors from all source images for every training sample,

$$\mu = mean\{\mathcal{PE}(I_t, I_{s \to t}) | s \in \mathcal{S}\}, \quad (5)$$
$$\sigma = std\{\mathcal{PE}(I_t, I_{s \to t}) | s \in \mathcal{S}\}. \quad (6)$$

Then, we compute an outlier mask for the photometric error map $\mathcal{PE}(I_t, I_{s \to t})$,

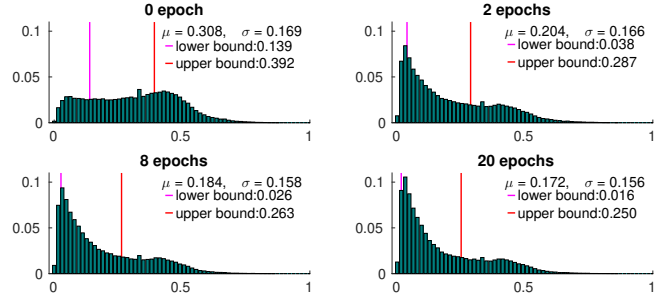$$M_s^{ol} = \mu - l\sigma < \mathcal{PE}(I_t, I_{s \to t}) < \mu + u\sigma, \quad (7)$$

Fig. 3: **The variation of the photometric error distribution.** The photometric error distribution of a validation sample changes during training. Before training, even with a little long tail, the errors are distributed somewhat evenly. After 2 epochs, the majority of errors converge to the lower bound, and a notable long tail forms. Next, the errors under the upper bound continue to decrease and converge, but the errors in the long tail do not change much.

where l and u are the lower and upper thresholds.

We can use the computed mask to exclude the possible occluded or moving regions, as shown in Fig. 2. By visualizing the resulted masks during training, we find that it is good to set u as 0.5 because a higher value cannot sufficiently mask the moving objects, and a lower value can mask out many stationary objects. Besides, l is set as 1 to mask some pixels with very small photometric errors because these pixels usually belong to homogeneous regions and not very valuable for network training. This selection for u and l can retain the principal photometric errors for optimizing, as shown in Fig. 3.

C. Weighted Multi-Scale Scheme

To avoid getting stuck in local minima due to the gradient locality of the bilinear sampler [27], the unsupervised learning models usually predict 4 scale depth maps (Fig. 2) and compute multi-scale photometric losses for training. However, it has been pointed out that this scheme tends to produce 'holes' in large low-texture regions in the intermediate lower resolution depth maps, as well as texture-copy artifacts [4]. To alleviate this phenomenon, Monodepth2 [4] adopts a full resolution multi-scale scheme, i.e., to upsample the multi-scale depth maps to the full resolution, perform the image warping using the full-resolution images, and compute photometric losses at the full resolution.

However, we find that this full-resolution scheme considerably increases the computation and GPU memory during training. To suppress the phenomenon without raising training overhead, we propose a weighted multi-scale scheme to devalue the low-resolution photometric losses and lighten the disadvantage they bring. Explicitly, we define a scale factor $f < 1$ to compute the weight for the scale r,

$$w_r = f^r, \quad (8)$$

where $r \in \{0, 1, 2, 3\}$.

D. Integrated Objective Function

Although the proposed outlier masking technique can exclude most irregular pixels, it has some failure cases. For example, the outlier masking cannot eliminate the pixels that move out of the image boundary, as illustrated by the bottom of the outlier mask in Fig. 2. In fact, it is easy to mask the out-of-box pixels by the principled masking technique [29], which only retains the pixels that are reprojected inside the image box of the source images. Besides, the outlier masking cannot mask out the objects with a very close speed to the camera, as these objects are usually estimated to the maximum depth, and the corresponding photometric errors can exactly lie in the statistical inlier region. As illustrated in Fig. 2, the car in the same lane is not well masked in the outlier mask. Fortunately, the auto-masking technique in Monodepth2 [4] can handle such cases and after including this masking method, the car in the front is not estimated to be very far away. Moreover, we also find that our outlier masking can collaborate with the minimum reprojection in [4] well to produce more accurate foreground object boundaries in the predicted depth maps. Therefore, we build a baseline with these three techniques.

The auto-masking excludes the pixels that hold larger photometric errors by reconstruction than the direct photometric error between the target view and the source view. The minimum reprojection is also a masking technique, and it only retains the pixels with the minimum photometric error among all source views. We express the masks of these three masking methods for the photometric error map $\mathcal{PE}(I_t, I_{s \to t})$ as M_s^p, M_s^a and M_s^{mr},

$$M_s^p = [p_s \text{ within image box}], \quad (9)$$

$$M_s^a = \mathcal{PE}(I_t, I_{s \to t}) < \mathcal{PE}(I_t, I_s), \quad (10)$$

$$M_s^{mr} = \mathcal{PE}(I_t, I_{s \to t}) \leq \min_s \mathcal{PE}(I_t, I_{s \to t}), \quad (11)$$

where p_s is calculated by Eqn. 2. Then we can compute the final mask for the photometric error map $\mathcal{PE}(I_t, I_{s \to t})$ by combining three type masks,

$$M = M_s^{ol} \bullet M_s^p \bullet M_s^a \bullet M_s^{mr}, \quad (12)$$

where $\bullet$ represents the element-wise logical conjunction. Finally, the overall objective function is computed by,

$$L = \eta \sum_r f^r \sum_s \frac{M_s P_s}{\#\{M_s = 1\}} + \lambda \sum_r e^r L_{es}^r, \quad (13)$$

where we denote $\mathcal{PE}(I_t, I_{s \to t})$ as P_s, η and λ are weights to balance the two types of losses, and e is a weighting factor for the edge-aware smoothness loss from different scales.

IV. EXPERIMENTS

A. Implementation Details

We implement the proposed approaches based on Monodepth2 [4] and maintain the most basic experimental settings. The depth CNN is a fully convolutional encoder-decoder network with an input/output resolution of 640×192. The Pose CNN is a stand CNN with fully connected layer to regress the 6-Dof relative camera motion. Both networks use a ResNet18 [32] pretrained on ImageNet [33] as backbone for all of the experiments. In depth estimation experiments, as Monodetph2 [4], we only use the nearby 2 frames ($\mathcal{S} = \{t-1, t+1\}$) and the pair-input Pose (Fig. 2). In ego-motion estimation, however, we also experiment with the all-input Pose CNN with the nearby 4 frames ($\mathcal{S} = \{t-2, t-1, t+1, t+2\}$) and 2 frames ($\mathcal{S} = \{t-1, t+1\}$) for [1], [29].

The hyper-parameter η, λ, and e in the final loss function are empirically set to 1, 0.001, and 0.5. The factor f of the weighted multi-scale scheme is chosen as 0.25 by examining several values in the validation set. DiPE is also trained for 20 epochs using Adam [34]. As our weighted multi-scale scheme consumes less memory, DiPE is trained with a bigger batch size of 16 than 12 in Monodepth2 and the training spends only 9 hours on a single Titan Xp while Monodepth2 uses 12 hours. DiPE also uses an initial learning rate of 10^{-4} but divides it by 5 after 15 and 18 epochs, whereas Monodepth2 divides it by 10 only after 15 epochs. As the outlier masking further reduces the errors for training and decreasing the learning rate can help DiPE converges better. Monodepth2 uses the same intrinsic parameters for all training samples by approximating the principal point of the camera to the image center and averaging the focal length on the whole dataset. More precisely, we use the calibrated intrinsic parameters for every training samples, and when performing horizontal flips in data augmentation, the horizontal coordinate of the principal point is also flipped.

B. KITTI Eigen Split

We adopt the standard Eigen split [5] of the KITTI dataset [16] in the monocular depth estimation experiments. Following Zhou *et al.* [1], we use a subset of the training set that contains no static frames for training. There are 39,810, 4,424, and 697 samples for training, validation, and test. We also only use about one-tenth (432) of the validation set for validation, because we evaluate all the validation samples after every epoch rather than evaluate a batch of validation samples for certain steps, and this is better for monitoring the training process without spending too much time on validation. In evaluation, every predicted depth map is aligned to the ground truth depth map by multiplying the median value ratio [1] as other unsupervised monocular methods, and we also adopt the conventional metrics and cropping region in [5], and the standard depth cap $80m$ [18]. There are 4 error metrics, namely, absolute relative error (Abs Rel), square relative error (Sq Rel), root mean square error (RMSE) and the root mean square error in log space (RMSE log). Other 3 accuracy metrics are the percentages of pixels where the ratio (δ) between the estimated depth and ground truth depth smaller than 1.25, 1.25^2 and 1.25^3.

1) Performance Comparison: We quantitatively and qualitatively compare the results of our model and other state-of-the-art methods. The quantitative results are shown in Table I and the results of other methods are taken from the corresponding papers. The comparison is mainly among the unsupervised monocular training methods, but some

Method	Train	Error metric ↓				Accuracy metric ↑		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen *et al.* [5]	D	0.203	1.548	6.307	0.282	0.702	0.890	0.890
Liu *et al.* [10]	D	0.201	1.584	6.471	0.273	0.680	0.898	0.967
Kuznietsov *et al.* [30]	DS	0.113	0.741	4.621	0.189	0.862	0.960	0.986
DORN [14]	D	**0.072**	**0.307**	**2.727**	**0.120**	**0.932**	**0.984**	**0.994**
Garg [17]	S	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Monodepth R50 [18]†	S	0.133	1.142	5.533	0.230	0.830	0.936	0.970
SuperDepth [31]	S	0.112	0.875	**4.958**	**0.207**	0.852	0.947	**0.977**
Monodepth2 [4]	S	**0.109**	**0.873**	4.960	0.209	**0.864**	**0.948**	0.975
SfMLearner [1]†	M	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Vid2Depth [29]	M	0.163	1.240	6.220	0.250	0.762	0.916	0.968
DF-Net [24]	M	0.150	1.124	5.507	0.223	0.806	0.933	0.973
GeoNet [23]†	M	0.149	1.060	5.567	0.226	0.796	0.935	0.975
DDVO [28]	M	0.151	1.257	5.583	0.228	0.810	0.936	0.974
EPC++ [2]	M	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2depth '(M)' [3]	M	0.141	1.026	5.291	0.215	0.816	0.945	0.979
Gordon *et al.*[26]	M	0.128	0959	5.230	0.212	0.845	0.947	0.976
Monodepth2 [4]	M	0.115	0.903	4.863	0.193	0.877	0.959	**0.981**
DiPE (Ours)	M	**0.112**	**0.875**	**4.795**	**0.190**	**0.880**	**0.960**	**0.981**

TABLE I: **Quantitative Results.** All the methods are trained and evaluated on the Eigen split [5] of the KITTI dataset [16]. Three categories of methods, which perform training with the depth, stereo images. and monocular video frames, respectively, are compared. In each category, the best results are in **bold**. **Legend:** D – depth supervision; S – unsupervised stereo supervision; M – unsupervised mono supervision; †– newer results from the respective online implementations.
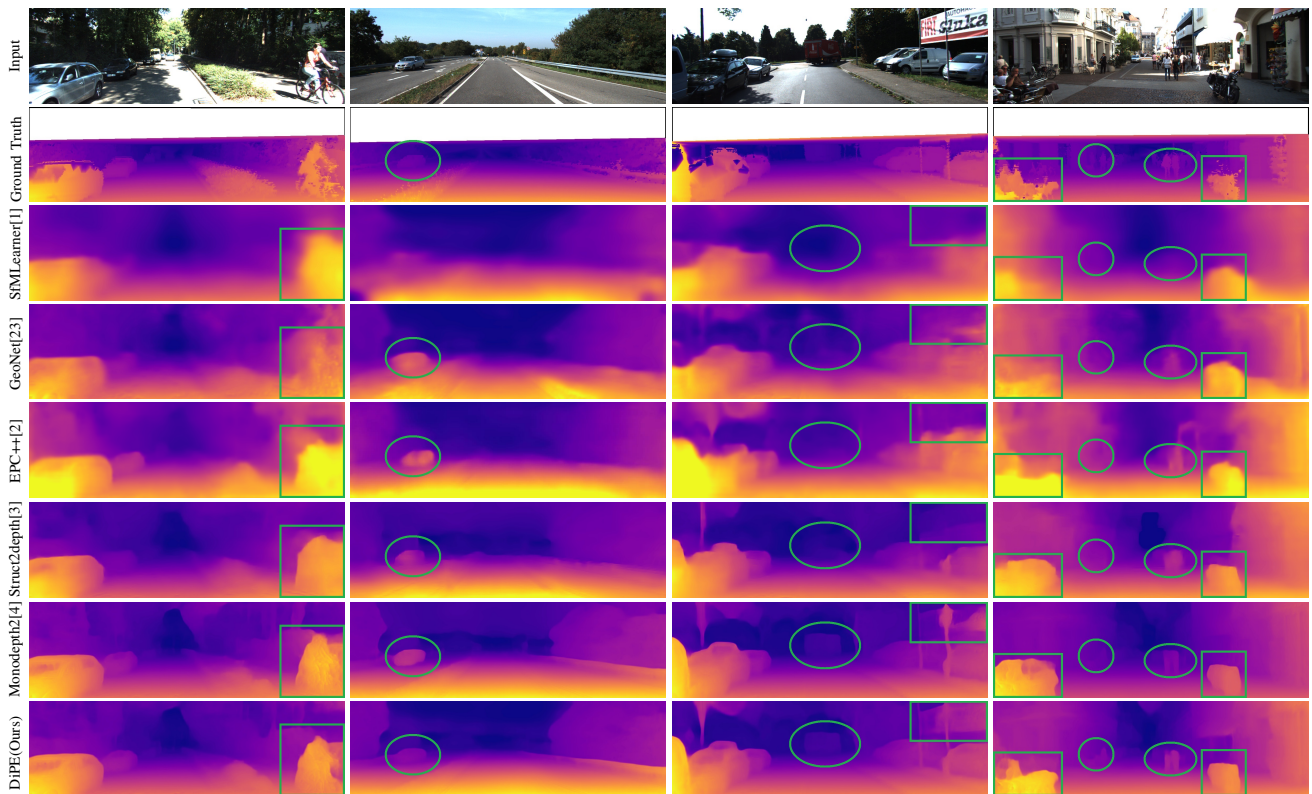

Fig. 4: **Qualitative comparison.** Our model DiPE produces very high-quality depth maps, and it reduces most artifacts due to occlusion and scene dynamics. More importantly, recent models, for instance, joint learning with optical flow [23], [2], *e.g.*oncoming cars, including modelling object motion [3] and auto-masking moving objects [4], underestimate the depth for the objects moving in an opposite direction, *e.g.*oncoming cars, while our DiPE succeeds (the second column). Best viewed in color and zoomed in.

representative depth supervised, and unsupervised stereo training models are also included. DiPE archives state-of-the-

art performance, as it markedly outperforms the current state of the art, Monodepth2 [4]. Also, DiPE has a comparable

	Weighted multi-scale	Outlier masking	Error metric ↓				Accuracy metric ↑		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Baseline			0.120	0.927	4.938	0.198	0.868	0.956	0.980
w/o Weighted multi-scale		✓	0.117	0.921	4.915	0.196	0.871	0.957	**0.981**
w/o Outlier masking	✓		0.115	0.910	4.865	0.193	0.876	0.958	0.980
DiPE	✓	✓	**0.112**	**0.875**	**4.795**	**0.190**	**0.880**	**0.960**	**0.981**

TABLE II: **Ablation Experiments.** There are 4 model variants with monocular training on the Eigen split [5] of the KITTI dataset [16]. The baseline model adopts existing principled masking, auto-masking and minimum reprojection techniques. Other 3 models include either one or both of our two contributions, the outlier making and weighted multi-scale methods.
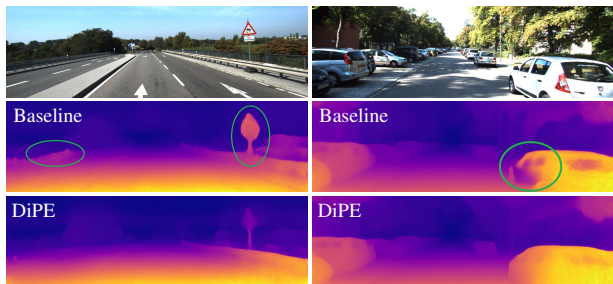

Fig. 5: **Artifacts.** DiPE can solve the artifacts better and success in the two failure cases by Monodepth2 as is reported in the paper of Monodepth2 [4].

or even better performance to the models of the other two categories. Fig 4 demonstrates the qualitative comparison among the predicted depth maps by DiPE and many state-of-the-art unsupervised monocular training methods. The predicted depth maps of other models are either shared by the authors or obtained by running the codes provided by the authors. DiPE handles the the scene dynamics and artifacts better than other methods. More results between DiPE and Monodepth2 [4] about coming vehicles can been seen from the attached video.

2) Ablation Study: We also perform ablation experiments to examine the effectiveness of our contributions. As mentioned in Section III-D, the baseline model uses the three existing masking techniques, i.e., the principled masking, auto-masking, and minimum reprojection. We experiment with four possible combinations of whether including our two contributions, the weighted multi-scale scheme, and the outlier masking technique. The results are shown in Table II.

It can be observed that, our two contributions can obviously improve the performance individually, and the performance gain when they combine together is more than double of their separate performance gain, which indicates that the two techniques can collaborate well. Furthermore, the weighted multi-scale scheme also helps DiPE address the artifacts better than Monodepth2 [4]. DiPE can handle the two failure cases in Monodepth2, as illustrated in Fig. 5.

C. KITTI Odometry

To prove the effectiveness of our DiPE model in the ego-motion estimation, we also experiment on the official odometry split of the KITTI dataset [16]. We use three different input settings for the ego-motion network, with the

number of frames as 2, 3, and 5, respectively. For training the ego-motion network with input as 2 or 3 frames, we use the DiPE based on the baseline model. However, for the 5-frame-input network, we do not adopt the minimum reprojection technique, because it almost masks out all the pixels from the source views with indexes of $t-2$ and $t+2$ and the motion estimation for these two views is inferior.

Method	Sequence 09	Sequence 10	# frames
ORB-SLAM [35]	0.014±0.008	0.012±0.011	-
SfMLearner [1]	0.021±0.017	0.020±0.015	5
DF-Net [24]	0.017±0.007	0.015±0.009	5
GeoNet [23]	0.012±0.007	0.012±0.009	5
DiPE (Ours)	**0.012±0.006**	**0.012±0.008**	5
DDVO [28]	0.045±0.108	0.033±0.074	3
Vid2Depth [29]	0.013±0.010	0.012±0.011	3
EPC++ [2]	0.013±0.007	**0.012±0.008**	3
DiPE (Ours)	**0.012±0.006**	**0.012±0.008**	3
Monodepth2 [4]	0.017±0.008	0.015±0.010	2
DiPE (Ours)	**0.013±0.006**	**0.012±0.008**	2

TABLE III: **Visual odometry results on the odometry split of the KITTI [16] dataset.** Results show the average absolute trajectory error, and standard deviation, in meters.

For evaluation, we adopt the commonly used metric proposed by Zhou *et al.* [1], i.e., the Absolute Trajectory Error (ATE) [35] in 5-frame snippets. The results are shown in Table III and the results of other models are taken from their corresponding papers. Among models with the three different input settings, DiPE achieves the best performance. Notably, in the setting of the pair-input ego-motion network, DiPE significantly outperforms Monodepth2 [4]. Besides, there is no significant performance difference among different motion network settings for DiPE, so DiPE is robust to different motion network input settings.

V. CONCLUSION

In this paper, we have demonstrated that carefully processing the photometric errors for unsupervised learning of depth and ego-motion from monocular videos can successfully solve the intrinsic difficulties, i.e., the occlusion and scene dynamics. We have introduced the outlier masking technique to exclude the irregular photometric errors that may mislead the network learning. This technique is useful in tackling occlusion and scene dynamics, especially for contra-directionally moving objects. Moreover, we have proposed an

efficient and effective weighted multi-scale scheme to avoid the artifacts brought by multi-scale training. Unlike other methods that introducing extra modules, our approaches are simple, as they can be very easily incorporated in the unsupervised geometry learning framework. We have experimentally proven the effectiveness of our two contributions and built a new state-of-the-art model, DiPE, on both monocular depth and ego-motion estimation.

REFERENCES

[1] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1851–1858.

[2] 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*, 2019.

[3] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," in *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

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

[5] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374.

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

[7] Y. Zeng, Y. Hu, S. Liu, Q. Tang, J. Ye, and X. Li, "Geocuedepth: exploiting geometric structure cues to estimate depth from a single image," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 17–22.

[8] H. Jiang and R. Huang, "High quality monocular depth estimation via a multi-scale network and a detail-preserving objective," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1920–1924.

[9] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Towards unified depth and semantic prediction from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2800–2809.

[10] 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.

[11] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," in *Proceedings of CVPR*, 2017.

[12] 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*, 2017.

[13] B. Li, Y. Dai, and M. He, "Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference," *Pattern Recognition*, vol. 83, pp. 328–339, 2018.

[14] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[15] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*. Springer, 2012, pp. 746–760.

[16] 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.

[17] 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.

[18] 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.

[19] H. Zhan, R. Garg, C. Saroj 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.

[20] R. Li, S. Wang, Z. Long, and D. Gu, "Undeepvo: Monocular visual odometry through unsupervised deep learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7286–7291.

[21] M. Klodt and A. Vedaldi, "Supervising the new with the old: learning sfm from sfm," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 698–713.

[22] G. Wang, H. Wang, Y. Liu, and W. Chen, "Unsupervised learning of monocular depth and ego-motion using multiple masks," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4724–4730.

[23] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1983–1992.

[24] 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.

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

[26] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova, "Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[27] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[28] C. Wang, J. Miguel 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.

[29] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5667–5675.

[30] Y. Kuznietsov, J. Stückler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6647–6655.

[31] S. Pillai, R. Ambruş, and A. Gaidon, "Superdepth: Self-supervised, super-resolved monocular depth estimation," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9250–9256.

[32] 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.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

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

[35] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.