2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

# Self-Supervised Learning of Monocular Visual Odometry and Depth with Uncertainty-Aware Scale Consistency

Changhao Wang, Guanwen Zhang, and Wei Zhou

***Abstract*— The inherent scale ambiguity issue greatly limits the performance of monocular visual odometry. In recent years, a variety of methods have been proposed for self-supervised learning of ego-motion and depth estimation, incorporating specifically designed scale-consistency constraints that utilize estimated depth as a reference. However, these existing methods neglect the influence of the depth uncertainty introduced by the dominant photometric loss, which leads to unreliable depth estimation in difficult regions and detrimentally affects scale alignment. To solve these problems, we introduces a feature-based visual odometry learning system with an effective scale recovery strategy in this paper. Additionally, we propose a learning method to estimate the photometric-sensitive depth uncertainty for guiding the scale recovery. The proposed method is evaluated on KITTI odometry, and the experimental results demonstrate that our system can predict scale-consistent trajectories from monocular videos and achieves state-of-the-art performance. Moreover, the proposed method achieves competitive performance on KITTI depth estimation.**

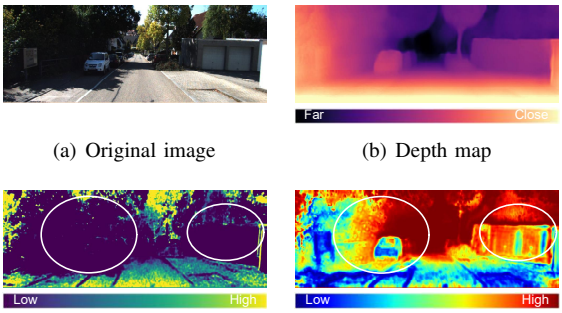
## I. Introduction

Ego-motion and depth estimation are crucial for autonomous systems to comprehend the scene structure around them and to recognize their location. These are fundamental challenges of visual Simultaneous Localization and Mapping (V-SLAM) and visual odometry (VO) systems, which play significant roles in many applications, such as AR/VR, intelligent robots, and autonomous vehicles.

For years, the monocular-based method has drawn attention for its low requirements of cost and easy of deployment. However, the inherent scale ambiguity issue existing in the monocular-based method can lead to predicting scale-inconsistent motion and structure. To address this problem, conventional V-SLAM and VO systems typically employ bundle adjustment in local windows or closing loops [1], [2]. In parallel, with the prosperity of deep learning, a lot of learning-based ego-motion and depth estimation methods that incorporate scale alignment strategies have been proposed to mitigate the scale ambiguity issue [3]–[6].

Recently, the self-supervised learning framework that jointly learns ego-motion and depth in an end-to-end manner is attractive [3], [4], [7]–[9]. These methods typically use two separate networks to simultaneously regress the pose and depth from image sequences, under the constraint of photometric error. Although the pose and depth will align with a consistent scale as well as possible during the regression process, achieving competitive performance with such solutions remains challenging. To leverage the learning-based methods with the two-view geometry, some works have integrated the geometry constraint into the self-supervised learning frameworks and achieve a substantial improvement [5], [6]. Specifically, these methods replace the pose regression network with a flow net and use the optical flow correspondences to solve the pose using the fundamental matrix. To acquire scale-consistent predictions, they align the estimated depth with the depth triangulated from the dense optical flow correspondences to recover the scaling factor between the predicted motion and structure. However, these works neglect that learning depth under the supervision of photometric error easily causes inaccurate estimation in difficult regions, and directly aligning the estimated depth with the triangulated depth limits the reliability of the recovered scale. Moreover, as the entangled depth estimation can provide sufficient structural information, predicting dense optical flow only for solving relative pose involves information redundancy and lacks efficiency.

From the above discussion, we present a feature-based monocular visual odometry system with an effective uncertainty-aware scale recovery strategy in this paper. We introduce a point network to predict sparse interest points, each of which contains a coordinate, a descriptor, and a score. According to the predictions of the point network, we perform feature matching with top-K points to generate




(a) Original image (b) Depth map

(c) Photometric uncertainty (d) Depth uncertainty

Fig. 1. A sample from the KITTI dataset. Using the proposed photometric-sensitive learning method, the hard regions that can converge to very small photometric error (c) may still have high depth uncertainty (d).

This work was supported in part by the National Key R&D Program of China (2018AAA0102801 and 2018AAA0102803), the Natural Science Basic Research Program of Shaanxi Province under Grant (2018ZE53052 and 2021JM-074), and the National Natural Science Foundation of China (61772424, 61702418, and 61602383). *(Corresponding author: Guanwen Zhang.)*

Changhao Wang, Guanwen Zhang, and Wei Zhou are with the Visual Intelligent Processing Laboratory, School of Electronics and Information, Northwestern Polytechnical University, Xi'an Shannxi, P.R.China. (e-mail: wangch@mail.nwpu.edu.cn; guanwen.zh@nwpu.edu.cn; zhouwei@nwpu.edu.cn)

Code is available at https://github.com/wangch-g/KPDepth-VO.

correspondences for solving relative pose and triangulating sparse depth. Meanwhile, we use a depth network to output the depth map with uncertainty estimation. Then, the scaling factor is recovered according to the predicted depth and the triangulated depth. With the recovered scaling factor, the predicted motion can be scaled to align with the structure during both the training and inference stages. Instead of recovering scale by comparing the two types of depth directly, we propose to take the depth uncertainty into account for scale recovery, which greatly alleviates the negative effect of the unreliable depth estimation in hard regions and ensures the accuracy and robustness of the system. In self-supervised learning of ego-motion and depth, the depth predictions are usually more uncertain in hard regions where the brightness constancy assumption may be invalid and the photometric loss may converge to a vary small value despite inaccurate depth estimation, such as non-Lambertian surfaces and large textureless areas [10]. To predict the depth uncertainty introduced by the inherent problem of photometric loss, we propose a novel photometric-sensitive learning method that will effectively assign high uncertainty to hard regions. As shown in Fig. 1(c) and 1(d), the regions that can converge to a very small photometric error under the supervision of photometric loss may still have high depth uncertainty. We evaluate the proposed method on KITTI odometry, the experimental results demonstrate that our system can predict scale-consistent trajectories from monocular videos and achieves state-of-the-art performance. Furthermore, as for depth estimation, our method achieves competitive performance on KITTI depth estimation.

In summary, our major contributions are as follows:

- We propose a feature-based learning system with an effective uncertainty-aware scale recovery strategy for monocular visual odometry.
- We propose a photometric-sensitive method to learn depth uncertainty introduced by the deficiency of photometric constraint for guiding scale recovery.
- The proposed method effectively solves the scale ambiguity problem and significantly improves the accuracy of monocular visual odometry. Experimental results demonstrate that our system achieves state-of-the-art performance on KITTI odometry and performs competitively on KITTI depth estimation.

## II. Related work

**Self-supervised visual odometry**. Ego-motion estimation from monocular video has been studied for decades, and various remarkable VO/SLAM systems are proposed on the basis of multiple view geometry [1], [2], [11]–[15]. Recent years, camera pose estimation based on the self-supervised learning method became prevalent. Zhou *et al*. [7] proposed the first full self-supervised pose and depth learning framework that jointly optimizes the predictions by minimizing the photometric error. Their work enlightened a lot of following researchers [3], [4], [9], [16], [17]. Bian *et al*. [3], [4] proposed a geometry consistency constraint and a self-discovered mask for predicting scale-consistent trajectory and improving the accuracy of depth estimation. However, the monocular ego-motion estimation generally suffers from the scale ambiguity issue severely. Although some works, such as [3], [4], [18], proposed implicit constraints to enforce the networks predicting scale-consistent pose and depth, the improvement is still limited. To address this problem, Zhan *et al*. [5] proposed a hybrid visual odometry system that leverages the advantages of two-view geometry constraint with deep learning. They solve the relative pose and triangulate the depth map from the optical flow correspondences. By comparing the triangulated depth with the predicted depth, they can estimate a scaling factor to align the motion and structure, which boosts the accuracy of trajectory estimation. Similar as [5], Zhao *et al*. [6] also proposed a flow-based pose and depth learning framework to solve the scale-inconsistency problem. Differently, they align the depth to the pose by supervising the depth estimation with the triangulated depth during the training phase while aligning the pose to the depth during inference.

**Depth and uncertainty estimation**. Learning depth in a supervised manner requires expensive ground truth information [19]–[21]. Therefore, self-supervised depth estimation has attracted numerous attention. With the known displacement, the depth can be learned from the left and right images with view synthesis [22], [23]. Without the stereo information, the self-supervised monocular depth estimation is always tightly coupled with the ego-motion estimation [7], [8], [24]–[27]. Godard *et al*. [8] proposed a per-pixel minimum photometric loss and an auto-masking loss to overcome the occlusion and dynamic problems respectively, which effectively improves the performance of self-supervised depth learning. Zhou *et al*. [25] proposed an architecture that benefits the accuracy of depth estimation with the especially designed multi-scale feature modulation module and the iterative inverse depth update strategy. Besides, some works showed that modeling the uncertainty of prediction is helpful for learning depth [24], [28], [29]. Klodt and Vedaldi [29] introduced the photometric uncertainty into loss function for improving the depth estimation. Despite the photometric uncertainty, Dikov and Vugt [10] proposed a Variational Depth Networks (VDN) to learn the depth uncertainty that reflects the reliability of predicted depth in different regions.

## III. Method

In this section, we introduce the proposed feature-based pose and depth learning system at first. Secondly, we present the method of photometric-sensitive depth uncertainty estimation. Finally, we detail the optimization and inference of the proposed system. The overview of our method is shown in Fig. 2.

### A. Feature-based pose and depth leaning

**Self-supervised learning**. On the basis of the brightness constancy assumption, the pose and depth can be learned simultaneously in a self-supervised learning manner by minimizing the photometric error between the target image $I_t$ and the synthesized image $I_{t'\to t}$. The synthesized image $I_{t'\to t}$ is

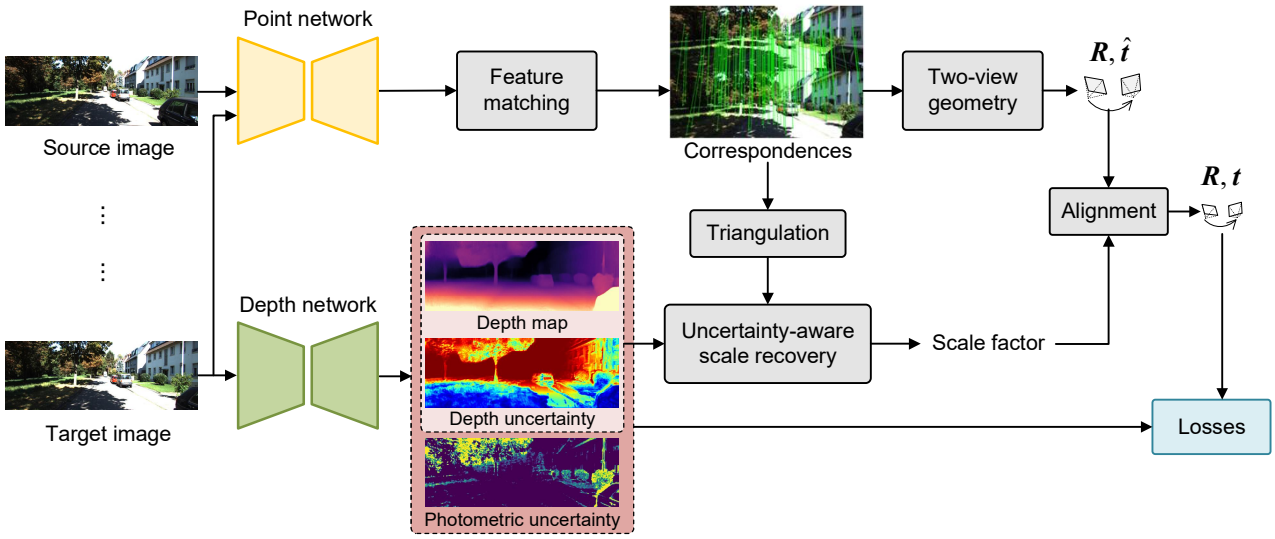



Fig. 2. The overview of the proposed system. The proposed system contains a point network and a depth network. The point network takes the target image and the source image as inputs to predict interest points. According to the descriptor similarity of interest points, we can acquire the correspondences between the target and source images. Meanwhile, the depth network takes the target image as input and outputs a depth map, a depth uncertainty map, and a photometric uncertainty map. Next, we solve the motion from the target viewpoint to the source viewpoint and triangulate the depth of sparse points based on the correspondences. And, the depth uncertainty is introduced to recover the scale factor by aligning the triangulated depth with the predicted depth. The scaling factor ensures the consistency between the motion and the structure, and it is crucial for both training and inference.

reconstructed by sampling the pixel values that are projected into the target viewpoint from the source image $I_{t'}$ with the target depth map $D_t$, the relative pose $[\boldsymbol{R}_{t\to t'}, \boldsymbol{t}_{t\to t'}]$, and the camera intrinsics $\boldsymbol{K}$. According to [23], the photometric error at pixel point $\boldsymbol{p}$ is formulated with L1 loss and SSIM [30] as:

$$pe(I_t(\boldsymbol{p}), I_{t'\to t}(\boldsymbol{p})) = \frac{\alpha}{2}(1 - \text{SSIM}(I_t(\boldsymbol{p}), I_{t'\to t}(\boldsymbol{p}))) + (1-\alpha)||I_t(\boldsymbol{p}) - I_{t'\to t}(\boldsymbol{p})||_1, \quad (1)$$

where $\alpha$ is commonly set to $0.85$. Further, we use the per-pixel minimum operation proposed in Monodepth2 [8] to formulate the photometric loss as:

$$L_p = \frac{1}{|N|}\sum_{\boldsymbol{p}\in N} \min_{t'} pe(I_t(\boldsymbol{p}), I_{t'\to t}(\boldsymbol{p})), \quad (2)$$

where $N$ is the set of all pixel points on an image and $t' \in [t-1, t+1]$ indicates that the previous and the next images with respective to the target image are used to compute the photometric loss [8].

Moreover, considering the probabilistic formulation, the pixel values on the synthesized images $I_{t'\to t}(\boldsymbol{p})$ can be seen as the observation of that on the target images $I_t(\boldsymbol{p})$. Following [29], with the predicted photometric uncertainty $\Sigma_t^p(\boldsymbol{p})$, $I_t(\boldsymbol{p})$ can be modeled by a posterior probability distribution $p(I_t(\boldsymbol{p})|I_{t'\to t}(\boldsymbol{p}), \Sigma_t(\boldsymbol{p}))$. As assuming the distribution is Laplacian, the learning objective is to minimize the negative log-likelihood, and the Eq. 2 can be further rewritten as:

$$L_p = \frac{1}{|N|}\sum_{\boldsymbol{p}\in N} \frac{\min_{t'} pe(I_t(\boldsymbol{p}), I_{t'\to t}(\boldsymbol{p}))}{\Sigma_t^p(\boldsymbol{p})} + \log \Sigma_t^p(\boldsymbol{p}). \quad (3)$$

**Solving pose with feature matching**. With the photometric loss, most existing methods regress the pose and depth using two separate deep neural networks in self-supervised learning. Generally, this kind of method reports promising depth estimation results, but the accuracy of the regressive pose is inferior compared with that of the conventional methods. In this paper, we use a point network to predict interest points and solve the pose with the point correspondences while jointly learning a depth network. Currently, some works achieve promising performance on learning interest points without the supervision of ground truth 3D information [31]–[33]. We follow the previous works and pre-train a point network to predict interest points from video sequences. Each of the predicted interest points has a coordinate, a descriptor, and a score. In the process of pose estimation, we select top-K points on both $I_t$ and $I_{t'}$ with high point scores and then perform feature matching to acquire correspondences. Notably, different from the flow-based methods that have to filter the occluded regions to generate reliable correspondences, we do not need to consider the occlusion problem since the correspondences are matched according to the similarity of the point descriptor. We set the K to be $30\%$ both in optimization and inference. With the pixel-wise correspondences, we can solve the fundamental matrix for recovering the relative pose $[\boldsymbol{R}_{t\to t'}, \hat{\boldsymbol{t}}_{t\to t'}]$ [34], [35].

**Uncertainty-aware scale recovery**. Since the translation solved from the fundamental matrix is up-to-scale, aligning the scale between the predicted pose and depth is necessary. Previous works either align depth to pose by supervising the depth estimation with the depth triangulated from the correspondences [6], or align pose to depth by

scaling the translation with a scaling factor that is recovered by comparing the predicted depth and the triangulated depth [5]. However, existing works have not yet taken the depth uncertainty introduced by the photometric training constraint into account during the process of scale recovery. Intuitively, aligning the scale on the pixel points with low depth uncertainty is more convincing than that on the pixel points with high depth uncertainty. Therefore, we propose an uncertainty-aware strategy that weights the scale computed from each pixel point by the certainty of depth to recover the scale. The uncertainty-aware scale recovery is formulated as:

$$s = \sum_{\boldsymbol{p} \in M} \left( \frac{1 - \Sigma_t^d(\boldsymbol{p})}{\sqrt{\sum_{\boldsymbol{p} \in M} (1 - \Sigma_t^d(\boldsymbol{p}))^2}} \right)^2 \frac{D_t(\boldsymbol{p})}{D_t^{tri}(\boldsymbol{p})}, \quad (4)$$

where $D_t$ and $D_t^{tri}$ are the predicted depth map and the triangulated depth map of the target image respectively, $\Sigma_t^d$ is the target depth uncertainty estimated by the depth network (Sec.III-B), and $M$ is the set of interest points that are filtered with cheirality constraint. With the scaling factor $s$, the aligned relative pose $[\boldsymbol{R}_{t \to t'}, \boldsymbol{t}_{t \to t'}]$ is computed as:

$$[\boldsymbol{R}_{t \to t'}, \boldsymbol{t}_{t \to t'}] = [\boldsymbol{R}_{t \to t'}, s\hat{\boldsymbol{t}}_{t \to t'}]. \quad (5)$$

### B. Photometric-sensitive depth uncertainty

As we mentioned before, estimating the depth uncertainty introduced by the photometric constraint is favourable for the scale alignment. Photometric constraint works upon the brightness constancy assumption. However, although the brightness consistency is satisfied, the photometric constraint may still lead to inaccurate prediction in some regions, in which the photometric error can be small by computing with a wrong depth estimation, such as textureless regions [10]. To this end, we propose a loss function for learning the photometric-sensitive depth uncertainty. During the training process, the depth network predicts the target depth map $D_t$ with its uncertainty map $\Sigma_t^d$. We create a noisy depth map $\hat{D}_t$ by sampling from the normal distribution that the location and the scale are defined by $D_t$ and $\Sigma_t^d$ respectively:

$$\hat{D}_t = \mathrm{NormalSample}(D_t, \Sigma_t^d). \quad (6)$$

Using the noisy depth map $\hat{D}_t$, we can reconstruct the synthesized image $\hat{I}_{t' \to t}$ and compute the photometric error $\hat{pe}$. By comparing with the photometric error $pe$, the proposed photometric-sensitive depth uncertainty loss is formulated as:

$$L_{ps} = \frac{1}{|V|} \sum_{\boldsymbol{p} \in V} (|pe(\boldsymbol{p}) - \hat{pe}(\boldsymbol{p})| - \gamma pe(\boldsymbol{p})) \Sigma_t^d(\boldsymbol{p}), \quad (7)$$

where $V$ is the set of pixel points that satisfy the minimum condition [8], $\gamma$ is a hyperparameter. From Eq.7, the gradient of $\Sigma_t^d$ will be positive when the difference between the $pe$ and the $\hat{pe}$ is larger than the tolerance threshold $\gamma pe$ or will be negative. In other words, by minimizing the $L_{ps}$, the regions that have the lower tolerance to varying the depth for unchanging the photometric error will be assigned with the lower depth uncertainty or will be assigned with the higher. By substituting the photometric-sensitive depth uncertainty $\Sigma_t^d$ to Eq.4, we can recover the scaling factor to align the relative pose with the predicted depth both.

### C. Optimization and inference

**Optimization**. We optimize the proposed system in two stages. At first, we follow the approach of [33] and use the same loss function to optimize the point network in a self-supervised learning manner. Next, we freeze the parameters of the point network and optimize the depth network with the proposed learning framework. In the second stage of optimization, apart from the $L_p$ and the $L_{ps}$, we additionally compute the depth consistency loss $L_{dc}$ between the target depth map $D_t$ and the reconstructed depth map $D_{t' \to t}$ [36].

$$L_{dc} = \frac{1}{|N|} \sum_{\boldsymbol{p} \in N} \min_{t'} \left| \frac{1}{D_t(\boldsymbol{p})} - \frac{1}{D_{t' \to t}(\boldsymbol{p})} \right|. \quad (8)$$

Besides, we also use the edge-aware smoothness loss $L_s$ to smooth the depth prediction [23].

$$L_s = |\partial_x D_t^*| e^{-|\partial_x I_t|} + |\partial_y D_t^*| e^{-|\partial_y I_t|}, \quad (9)$$

where $D_t^*$ is the mean-normalized inverse depth.

Therefore, the total loss function used in the second stage of optimization is defined as:

$$L = \lambda_p L_p + \lambda_{ps} L_{ps} + \lambda_{dc} L_{dc} + \lambda_s L_s, \quad (10)$$

where $[\lambda_{pe}, \lambda_{ps}, \lambda_{dc}, \lambda_s]$ are the loss weights.

**Inference**. During inference, our system predicts the relative pose and depth as the same as the training process. Differently, for stable estimation, we follow the approach of [5], [6] that uses perspective-n-point (PnP) method instead of epipolar geometry to solve the relative pose when the parallax between the consecutive frames is small.

## IV. Experiments

### A. Implementation details

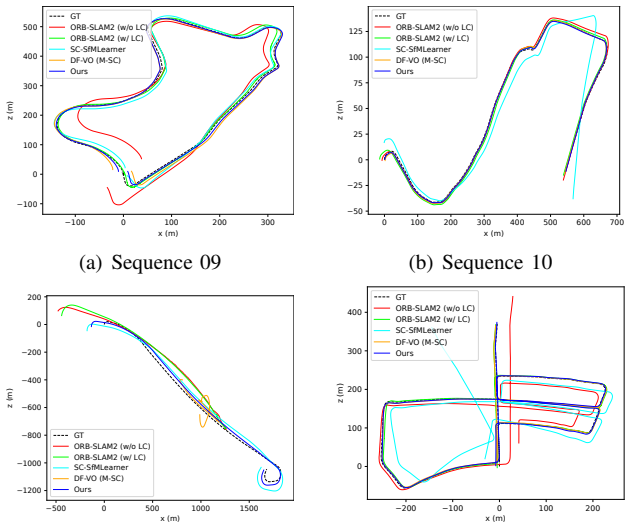
**Network Architectures**. We use the LANet [33] as our point network since it is lightweight and easy to train. For the depth network, we use the same architecture as the Monodepth2 [8] with the backbone of ResNet18 [37] pretrained on ImageNet [38]. We add two output channels to the last layer of the decoder in the depth network to output the depth uncertainty and the photometric uncertainty respectively.

**Training**. We use COCO 2017 dataset [39] and the same settings as [33] to train the point network at first. Next, the depth network is trained on KITTI dataset [40]. For KITTI odometry, we use sequences 00-08 for training and 09-10 for testing [7]. For depth estimation, we train and test the depth network using the Eigen split protocol [19]. The depth network is trained by Adam optimizer for 12 epochs. The learning rate is set to $10^{-4}$ initially and reduced to $10^{-5}$ at the last 5 epochs. The input images are resized to $320 \times 1024$ with a batch size of 8. The loss weights are set as $[\lambda_{pe}, \lambda_{ps}, \lambda_{dc}, \lambda_s] = [1.0, 1.0, 0.1, 0.001]$ and the hyperparameter in $L_{ps}$ is set as $\gamma = 0.25$. The proposed system is implemented with PyTorch [41] on a single NVIDIA GeForce RTX 3090 GPU.

TABLE I

COMPARISONS ON KITTI ODOMETRY. THE BEST AND THE SECOND BEST RESULTS ARE IN BOLD AND UNDERLINED RESPECTIVELY EXPECT FOR THE METHODS WITH ONLINE ADAPTATION. THE RESULTS OF ONLINE ADAPTATION METHODS THAT ACHIEVE THE BEST ARE SUPERSCRIBED WITH $^*$.

| Methods | Metric | Seq. 00 | Seq. 01 | Seq. 02 | Seq. 03 | Seq. 04 | Seq. 05 | Seq. 06 | Seq. 07 | Seq. 08 | Seq. 09 | Seq. 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ORB-SLAM2 [2] (w/o loop closure) | $t_{err}$ | 11.43 | 107.57 | 10.34 | 0.97 | 1.30 | 9.04 | 14.56 | 9.77 | 11.46 | 9.30 | 2.57 |
| | $r_{err}$ | 0.58 | 0.89 | **0.26** | **0.19** | **0.27** | 0.26 | 0.26 | 0.36 | **0.28** | 0.26 | 0.32 |
| | ATE | 40.65 | 502.20 | 47.82 | 0.94 | 1.30 | 29.95 | 40.82 | 16.04 | 43.09 | 38.77 | 5.42 |
| ORB-SLAM2 [2] (w/ loop closure) | $t_{err}$ | 2.35 | 109.10 | 3.32 | **0.91** | 1.56 | 1.84 | 4.99 | 1.91 | 9.41 | 2.88 | 3.30 |
| | $r_{err}$ | **0.35** | **0.45** | 0.31 | **0.19** | **0.27** | **0.20** | **0.23** | **0.28** | 0.30 | **0.25** | **0.30** |
| | ATE | **6.03** | 508.34 | **14.76** | 1.02 | 1.57 | 4.04 | 11.16 | 2.19 | 38.85 | 8.39 | 6.63 |
| SfM-Learner [7] | $t_{err}$ | 21.32 | 22.41 | 24.10 | 12.56 | 4.32 | 12.99 | 15.55 | 12.61 | 10.66 | 11.32 | 15.25 |
| | $r_{err}$ | 6.19 | 2.79 | 4.18 | 4.52 | 3.28 | 4.66 | 5.58 | 6.31 | 3.75 | 4.07 | 4.06 |
| | ATE | 104.87 | 109.61 | 185.43 | 8.42 | 3.10 | 60.89 | 52.19 | 20.12 | 30.97 | 26.93 | 24.09 |
| Depth-VO-Feat [42] | $t_{err}$ | 6.23 | 23.78 | 6.59 | 15.76 | 3.14 | 4.94 | 5.80 | 6.49 | 5.45 | 11.89 | 12.82 |
| | $r_{err}$ | 2.44 | 1.75 | 2.26 | 10.62 | 2.02 | 2.34 | 2.06 | 3.56 | 2.39 | 3.60 | 3.41 |
| | ATE | 64.45 | 203.44 | 85.13 | 21.34 | 3.12 | 22.15 | 14.31 | 15.35 | 29.53 | 52.12 | 24.70 |
| SC-SfMLearner [3] | $t_{err}$ | 11.01 | 27.09 | 6.74 | 9.22 | 4.22 | 6.70 | 5.36 | 8.29 | 8.11 | 7.64 | 10.74 |
| | $r_{err}$ | 3.39 | 1.31 | 1.96 | 4.93 | 2.01 | 2.38 | 1.65 | 4.53 | 2.61 | 2.19 | 4.58 |
| | ATE | 93.04 | 85.90 | 70.37 | 10.21 | 2.97 | 40.56 | 12.56 | 21.01 | 56.15 | 15.02 | 20.19 |
| Li *et al.* [43] | $t_{err}$ | 1.32$^*$ | 2.83$^*$ | 1.42$^*$ | 1.77 | 1.22 | 1.07 | 1.02 | 2.06 | 1.50$^*$ | 1.87 | 1.93 |
| | $r_{err}$ | 0.45 | 0.65 | 0.45 | 0.39 | 0.27$^*$ | 0.44 | 0.41 | 1.18 | 0.42 | 0.46 | 0.30$^*$ |
| DOC+ [44] | $t_{err}$ | - | - | - | - | - | - | - | - | - | 2.02 | 2.29 |
| | $r_{err}$ | - | - | - | - | - | - | - | - | - | 0.61 | 1.10 |
| | ATE | - | - | - | - | - | - | - | - | - | 4.76$^*$ | 3.38 |
| DF-VO (M-SC) [5] | $t_{err}$ | 2.25 | 66.98 | 3.60 | 2.67 | 1.43 | 1.15 | 1.03 | **0.93** | 2.23 | 2.47 | 1.96 |
| | $r_{err}$ | 0.58 | 17.04 | 0.52 | 0.50 | 0.29 | 0.30 | 0.26 | 0.29 | 0.30 | 0.30 | 0.31 |
| | ATE | 12.64 | 695.75 | 23.11 | 1.23 | 1.36 | 3.75 | 2.63 | **1.74** | 7.87 | 11.02 | 3.37 |
| TrainFlow [6] | $t_{err}$ | - | - | - | - | - | - | - | - | - | 6.93 | 4.66 |
| | $r_{err}$ | - | - | - | - | - | - | - | - | - | 0.44 | 0.62 |
| **Ours** | $t_{err}$ | **1.96** | **15.03** | **2.76** | 2.66 | **0.78** | **0.95** | **0.81** | 2.93 | **1.62** | **1.86** | **1.55** |
| | $r_{err}$ | 0.63 | 0.52 | 0.54 | 0.53 | 0.44 | 0.34 | 0.29 | 1.11 | 0.38 | 0.29 | 0.31 |
| | ATE | 12.95 | **54.08** | 18.39 | **0.93** | **0.65** | **3.21** | **2.10** | 6.27 | **6.14** | **6.06** | **2.64** |




(a) Sequence 09 (b) Sequence 10

(c) Sequence 01 (d) Sequence 05

Fig. 3. Qualitative tracking trajectories on KITTI odometry.

## B. Visual odometry evaluation

We adopt the commonly used evaluation metrics of the relative translation error $t_{err}$ (%), relative rotation error $r_{err}$ (°/100m), and absolute trajectory error ATE (m) for testing on KITTI odometry. We report the comparison results in Tab. I. The compared methods are categorized into four sets with geometry-based methods [2], pure learning-based methods [3], [7], [42], online adaptation methods [43], [44], and hybrid methods [5], [6]. Since the real-world scale is unknown for monocular system, the evaluation results are aligned to the ground truth for comparison. From Tab. I, expect for comparing the methods with online adaptation, our method can achieve the best of $t_{err}$ and ATE on both sequence 09 and sequence 10. By comparing the hybrid methods, our system outperforms the TrainFlow [6] by a large margin and surpasses the DF-VO (M-SC) [5] on overall performance. The tracking trajectories of sequences 09 and 10 are shown in Fig. 3(a) and 3(b). Furthermore, we also report the results on sequences 00-08. As shown in Tab. I, our method achieves the best $t_{err}$ and ATE on the majority of sequences and outperforms the classic ORB-SLAM2 with loop closure without any backend optimization. The results indicate that our system can predict the scale-consistent trajectory and proves the effectiveness of the proposed scale recovery strategy. Two qualitative results are demonstrated in Fig. 3(c) and 3(d). As for sequence 01, most of the methods fail to track trajectory while our system shows remarkable robustness and succeeds in tracking. By comparing with the online adaptation methods, our system performs comparably with the method of Li *et al.* [43] that updates the network parameters with the meta-learning strategy during testing.

TABLE II

DEPTH ESTIMATION ON KITTI. ALL METHODS ARE TRAINED ON KITTI WITH MONOCULAR MODALITY AT THE HIGH-RESOLUTION SETTINGS. "U-A": UNCERTAINTY-AWARE STRATEGY, "P-S DEPTH UNCERTAINTY": PHOTOMETRIC-SENSITIVE DEPTH UNCERTAINTY.

| Methods | Error ↓ | | | | Accuracy ↑ | | |
|---|---|---|---|---|---|---|---|
| | Abs Rel | Sq Rel | RMSE | RMSE log | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
| Monodepth2 [8] | 0.115 | 0.882 | 4.701 | 0.190 | 0.879 | 0.961 | 0.982 |
| R-MSFM3 [25] | 0.112 | 0.773 | 4.581 | 0.189 | 0.879 | 0.960 | 0.982 |
| R-MSFM6 [25] | 0.108 | 0.748 | **4.470** | 0.185 | 0.889 | 0.963 | 0.982 |
| HR-Depth [26] | 0.106 | 0.755 | 4.472 | 0.181 | 0.892 | **0.966** | **0.984** |
| Zhang *et al*. [27] | **0.105** | 0.755 | 4.492 | 0.183 | **0.893** | 0.964 | 0.983 |
| **Ours** w/o U-A (average) | 0.110 | **0.718** | 4.545 | 0.186 | 0.880 | 0.962 | 0.983 |
| **Ours** w/ U-A (photometric uncertainty) | 0.107 | 0.719 | 4.552 | 0.184 | 0.884 | 0.963 | 0.983 |
| **Ours** w/ U-A (P-S depth uncertainty) | **0.105** | 0.727 | 4.491 | **0.180** | 0.888 | 0.964 | **0.984** |

TABLE III

ABLATION STUDY ON SEQUENCES 09-10 OF KITTI ODOMETRY.

| Methods | | Seq. 09 | | | Seq.10 | | |
|---|---|---|---|---|---|---|---|
| | | $t_{err}$ | $r_{err}$ | ATE | $t_{err}$ | $r_{err}$ | ATE |
| Baseline | PointNet only | 8.65 | **0.29** | 28.13 | 10.89 | 0.89 | 26.61 |
| | PointNet + DepthNet + PnP | 3.19 | 0.82 | 9.29 | 2.60 | 1.27 | 4.76 |
| Ours | w/o U-A (average) | 4.20 | 0.36 | 15.06 | 2.52 | 0.40 | 4.67 |
| | w/ U-A (photometric uncertainty) | 4.21 | 0.31 | 14.69 | 2.28 | **0.29** | 4.23 |
| | w/ U-A (P-S depth uncertainty) | **1.86** | **0.29** | **6.06** | **1.55** | 0.31 | **2.64** |

### C. Depth estimation

The depth network of the proposed system is evaluated on KITTI dataset splitting by the Eigen testing protocol [19]. We compare our depth network with the recent models that are trained in monocular modality using the high-resolution inputs of $320 \times 1024$ and report the results using metrics of Abs Rel, Sq Rel, RMSE, RMSElog, $\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$. From Tab. II, our depth network achieves competitive overall performance. Moreover, we explore the effects of the proposed uncertainty-aware strategy. Firstly, we train the system without the uncertainty-aware strategy but using the average of all scale factors. As shown in the third line from the last of Tab. II, expect for the square relative error, the performance drops obviously. Next, compared with using the photometric uncertainty, the proposed photometric-sensitive depth uncertainty improves the performance on all metrics expect for the square relative error. The results indicate that the depth network can benefit from the scale-consistency relative pose estimation of our system.

### D. Ablation study

We perform an ablation study on sequences 09-10 of KITTI odometry, and the results are shown in Tab. III. We additionally report the results of two baseline settings: (1) solving the relative pose by the epipolar geometry with the outputs of the point network only; (2) solving the relative pose by the PnP method with the predictions of the point network and the depth network.

Compared with the results of PointNet only, our system dramatically reduces the $t_{err}$ and ATE. From the results of the second row, it seems that using the PnP method can achieve an acceptable performance. That is because the depth network is optimized with the proposed training scheme, from which solving the relative pose by the PnP method benefits. Further, we train and evaluate a model without the proposed uncertainty-aware but the average scale recovery strategy, the results are reported in the third row of Tab. III. By comparing with that, recovering the scaling factor by our proposed method outperforms the average strategy by a large margin on $t_{err}$ and ATE, and slightly improves the performance on $r_{err}$. Moreover, we also train and evaluate a model using the photometric uncertainty in uncertainty-aware by replacing $\Sigma_t^d(\boldsymbol{p})$ to $\Sigma_t^p(\boldsymbol{p})$ in Eq. 4, the results are shown in the fourth row of Tab. III. Compared with the photometric uncertainty, weighting the scale factors by the proposed photometric-sensitive depth uncertainty can significantly reduce the $t_{err}$ and ATE. The results show that the proposed uncertainty-aware scale recovery and the photometric-sensitive depth uncertainty are effective to predict scale-consistent motion for monocular visual odometry.

## V. CONCLUSION

In this paper, we present a feature-based visual odometry and depth learning system with an effective uncertainty-aware scale recovery strategy. To estimate the depth uncertainty arising from the limitation of photometric constraint for scale recovery, we propose a photometric-sensitive depth uncertainty learning method. The experimental results on KITTI odometry indicate that the proposed system can predict scale-consistent trajectories from monocular videos and outperforms the conventional method that has strong backend optimization. Moreover, the depth network of our system achieves comparable performance in depth estimation on KITTI dataset.

## REFERENCES


[1] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans. Robotics*, 31(5):1147–1163, 2015.

[2] Raul Mur-Artal and Juan D. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Trans. Robotics*, 33(5):1255–1262, 2017.

[3] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian D. Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *NeurIPS*, pages 35–45, 2019.

[4] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth learning from video. *Int. J. Comput. Vis.*, 129(9):2548–2564, 2021.

[5] Huangying Zhan, Chamara Saroj Weerasekera, Jia-Wang Bian, and Ian Reid. Visual odometry revisited: What should be learnt? In *ICRA*, pages 4203–4210, 2020.

[6] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. In *CVPR*, pages 9148–9158, 2020.

[7] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, pages 6612–6619, 2017.

[8] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3827–3837, 2019.

[9] Cong Wang, Yu-Ping Wang, and Dinesh Manocha. Motionhint: Self-supervised monocular visual odometry with motion constraints. In *ICRA*, pages 1265–1272, 2022.

[10] Georgi Dikov and Joris van Vugt. Variational depth networks: Uncertainty-aware monocular self-supervised depth estimation. In *ECCV*, volume 13808, pages 43–60, 2022.

[11] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[12] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: large-scale direct monocular SLAM. In *ECCV*, volume 8690, pages 834–849, 2014.

[13] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO: fast semi-direct monocular visual odometry. In *ICRA*, pages 15–22. IEEE, 2014.

[14] Christian Forster, Zichao Zhang, Michael Gassner, Manuel Werlberger, and Davide Scaramuzza. SVO: semidirect visual odometry for monocular and multicamera systems. *IEEE Trans. Robotics*, 33(2):249–265, 2017.

[15] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(3):611–625, 2018.

[16] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, pages 1983–1992, 2018.

[17] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *ICCV*, pages 7062–7071, 2019.

[18] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *ECCV*, volume 11209, pages 38–55, 2018.

[19] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Neural Information Processing Systems*, pages 2366–2374, 2014.

[20] 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.

[21] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *ICCV*, pages 5683–5692, 2019.

[22] Ravi Garg, B. G. Vijay Kumar, Gustavo Carneiro, and Ian D. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *ECCV*, volume 9912, pages 740–756, 2016.

[23] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 6602–6611, 2017.

[24] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3VO: deep depth, deep pose and deep uncertainty for monocular visual odometry. In *CVPR*, pages 1278–1289, 2020.

[25] Zhongkai Zhou, Xinnan Fan, Pengfei Shi, and Yuanxue Xin. R-MSFM: recurrent multi-scale feature modulation for monocular depth estimating. In *ICCV*, pages 12757–12766, 2021.

[26] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. In *AAAI*, pages 2294–2301, 2021.

[27] Tianyu Zhang, Dongchen Zhu, Guanghui Zhang, Wenjun Shi, Yanqing Liu, Xiaolin Zhang, and Jiamao Li. Spatiotemporally enhanced photometric loss for self-supervised monocular depth estimation. In *IROS*, pages 1–8, 2022.

[28] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, pages 5574–5584, 2017.

[29] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: Learning SFM from SFM. In *ECCV*, volume 11214, pages 713–728, 2018.

[30] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

[31] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Workshops*, pages 224–236, 2018.

[32] Jiexiong Tang, Hanme Kim, Vitor Guizilini, Sudeep Pillai, and Rares Ambrus. Neural outlier rejection for self-supervised keypoint learning. In *ICLR*, 2020.

[33] Changhao Wang, Guanwen Zhang, Zhengyun Cheng, and Wei Zhou. Rethinking low-level features for interest point detection and description. In *ACCV*, volume 13842, pages 108–123, 2022.

[34] Richard I. Hartley. In defense of the eight-point algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(6):580–593, 1997.

[35] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[36] Huangying Zhan, Chamara Saroj Weerasekera, Ravi Garg, and Ian D. Reid. Self-supervised learning for single view depth and surface normal estimation. In *ICRA*, pages 4811–4817, 2019.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

[39] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, volume 8693, pages 740–755, 2014.

[40] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, pages 3354–3361, 2012.

[41] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zach DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[42] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian D. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, pages 340–349, 2018.

[43] Shunkai Li, Xin Wu, Yingdian Cao, and Hongbin Zha. Generalizing to the open world: Deep visual odometry with online adaptation. In *CVPR*, pages 13184–13193, 2021.

[44] Jiaxin Zhang, Wei Sui, Xinggang Wang, Wenming Meng, Hongmei Zhu, and Qian Zhang. Deep online correction for monocular visual odometry. In *ICRA*, pages 14396–14402, 2021.