Generalizing to the Open World: Deep Visual Odometry with Online Adaptation

Shunkai Li  Xin Wu  Yingdian Cao  Hongbin Zha
Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
PKU-SenseTime Machine Vision Joint Lab
{lishunkai, wuxin1998, yingdianc}@pku.edu.cn  zha@cis.pku.edu.cn

Abstract

Despite learning-based visual odometry (VO) has shown impressive results in recent years, the pretrained networks may easily collapse in unseen environments. The large domain gap between training and testing data makes them difficult to generalize to new scenes. In this paper, we propose an online adaptation framework for deep VO with the assistance of scene-agnostic geometric computations and Bayesian inference. In contrast to learning-based pose estimation, our method solves pose from optical flow and depth while the single-view depth estimation is continuously improved with new observations by online learned uncertainties. Meanwhile, an online learned photometric uncertainty is used for further depth and pose optimization by a differentiable Gauss-Newton layer. Our method enables fast adaptation of deep VO networks to unseen environments in a self-supervised manner. Extensive experiments including Cityscapes to KITTI and outdoor KITTI to indoor TUM demonstrate that our method achieves state-of-the-art generalization ability among self-supervised VO methods.

1. Introduction

Estimating camera motion from monocular videos plays an essential role in many real-world applications, such as autonomous driving and robotics. This problem is usually solved by visual odometry (VO) or simultaneous localization and mapping (SLAM). Classic SLAM/VO methods [7, 8, 10, 26] perform well in favorable conditions but often fail in challenging situations (e.g. textureless region, dynamic object) due to the reliance on low-level features and hand-crafted pipeline. Since deep neural networks are able to extract high-level features and infer end-to-end by learning from data, many learning-based VO methods [21, 22, 39, 47] have been proposed to break through the limitations of classic SLAM/VO. Among them, self-supervised VO methods are able to jointly learn camera pose, depth and optical flow by minimizing photometric error [39], which have shown promising results in recent years.

However, learning-based VO often fails during inference when the scenes are different from the training data. The inability of pretrained VO to generalize to unseen environments limits its wide applications [21, 44]. To this end, the pretrained networks are required to achieve real-time online adaptation in a self-supervised manner.

As a result, several previous works [3, 21, 44] have been proposed to mitigate the domain generalization problem of stereo matching and VO. However, the performance is still much inferior to classic methods in terms of accuracy and the pretrained networks suffer from slow convergence. These methods treat VO as a black-box by learning all components (pose, depth, optical flow, etc.) but ignore well-defined geometric computations and optimization methods, which leads to slow convergence during online adaptation.

Figure 1. In this paper, we focus on the generalization ability to unseen environments of deep VO. When the test data are different from the training data, previous methods fail to generalize while our method still performs well with very small trajectory error.
Existing deep VO methods predict depth by single-view estimation, which is an ill-posed problem [22]. The learned depth has a strong reliance on the training dataset. During inference, the camera intrinsics, scene layouts and distances are usually different. Meanwhile, the camera pose is learned rather than calculated analytically, which requires favorable camera motion with sufficient disparity (e.g. KITTI dataset). Therefore, these methods tend to fail when faced with unseen or more complicated motion patterns. In addition, existing learning-based methods do not explicitly ensure multi-view geometric consistency during inference, which leads to large scale drift in trajectories.

In order to improve the online adaptation of VO to unseen environments, we propose a self-supervised framework that combines the advantage of deep learning and geometric computations. The proposed framework utilizes scene-agnostic 3D geometry constraints and Bayesian inference formulations to speed up online adaptation. During inference, the single-view depth estimation is used as a prior of the current scene geometry and is continuously improved with incoming observations by a probabilistic Bayesian updating framework. The refined depth is used as Maximum A Posteriori (MAP) to train DepthNet for better estimation at the next timestep. Instead of predicting pose by PoseNet, our framework solves pose analytically from optical flow and refined depth. Meanwhile, in order to deal with observation noise, the proposed method online learns depth and photometric uncertainties which are used in the depth refinement process and differentiable Gauss-Newton optimization, respectively. Finally, the optimized pose, depth and flow are used for online self-supervision. Our framework ensures scale consistency by exploiting multi-view geometric constraints. The well-defined scene-agnostic computation helps our VO framework achieve good generalization ability across different scene conditions. Our contributions can be summarized as follows:

- We propose a generalizable deep VO that uses scene-agnostic geometric formulation and Bayesian inference to speed up self-supervised online adaptation.
- The predicted depth is continuously refined by a Bayesian fusion framework, which is further used to train depth and optical flow during online learning.
- We introduce online learned depth and photometric uncertainties for better depth refinement and differentiable Gauss-Newton optimization.

Our method achieves much better generalization than state-of-the-art baselines when tested across different domains, including Cityscapes [4] to KITTI [13] and outdoor KITTI to indoor TUM [31] datasets. Meanwhile, we also achieve state-of-the-art depth estimation results on KITTI and NYUv2 [30] datasets.

2. Related works

Learning-based VO has been widely studied in recent years and shown impressive results [35, 37, 45]. Deep-TAM [45] mimics the framework of parallel tracking and mapping in classic SLAM/VO by using two networks for depth and pose estimation simultaneously. Xue et al [37] extends the VO pipeline to tracking, selecting memory and refining modules, which shows superior performance under challenging conditions. However, these methods require ground truth which is often impractical to obtain. In order to alleviate the need of ground truth data, self-supervised VO has been proposed. SfMLearner [47] learns depth and pose simultaneously by minimizing photometric loss between warped and input image. Zhao et al. [43] and Ranjan et al. [28] extend this idea to joint estimation of pose, depth and optical flow. Monodepth2 [15] explicitly handles non-rigid and occluded cases which are against static-scene assumption. SAVO [22] exploits spatial-temporal correlations over long sequence and utilizes RNN to reduce scale drift. In this paper, we use the depth network of Monodepth2 [15] for single-view depth estimation.

Online adaptation Most machine learning algorithms assume that the training and testing data are sampled from the same feature distribution. However, when the test data are different from the training set, most pretrained models suffer from a significant reduce in performance. In this situation, online learning [23, 33] is an effective method to solve the domain shift problem. Previous methods use online gradient update [5] and probabilistic filtering [2] to accelerate domain adaptation. In the computer vision field, Zhong et al. [44] proposes a self-supervised framework for stereo matching in the open world. Li et al. [21] proposes an online meta-learning algorithm for VO to continuously adapt to unseen environments. However, these methods learn all components by deep networks, leading to slow convergence and inferior performance. In contrast, our method combines the advantage of deep learning and well-defined geometric computations to achieve better generalization.

3D Geometric computations In classic 3D computer vision, the relative pose between two images and scene depth can be solved analytically by multi-view geometric constraints. Given a set of correspondences, the pose can be solved by epipolar geometry [16, 17] with 2D-2D matching or Perspective-n-Point (PnP) [19] with 3D-2D matching. The depth of each correspondence can be recovered by mid-point triangulation [26]. On the other hand, the depth and pose can also be solved by minimizing photometric error [7, 8] via classic optimizations. If more observations are available, the 3D map can be further refined by Bundle Adjustment (BA) [26] or filtering [10]. In this paper, we adopt a Bayesian depth fusion method to refine single-view depth estimation and propose a differentiable Gauss-Newton layer to minimize weighted photometric residuals.
3. Method

In this section, we will introduce our framework in detail. The system overview is illustrated in Fig. 2. Firstly, the FlowNet predicts dense optical flow between the keyframe $I_k$ and current frame $I_t$ (Section 3.1), and predicts photometric uncertainty map $\Gamma_t$ (Section 3.4) as a side output. Meanwhile, the DepthNet estimates depth mean $D_0^k$ and uncertainty $\Sigma_k$ of keyframe, providing a prior of the current scene geometry (Section 3.2). The relative pose $T_k^t$ is solved by essential matrix or PnP from selected flow correspondences. During online adaptation, we firstly reconstruct the sparse depth of $I_k$ by a differentiable triangulation module. Then, the prior keyframe depth $D_0^k$ is continuously improved by subsequent depth estimations in a Bayesian updating framework (Section 3.3). Next, the differentiable Gauss-Newton layer minimizes the photometric loss of $I_t$ and warped image $\hat{I}_t$ weighted by predicted $\Gamma_t$ (Section 3.5). Finally, the optimized depth $D_k$ and flow $F_{k,t}$ are used as pseudo ground truth to supervise the online learning of DepthNet and FlowNet (Section 3.6).

3.1. Pose recovery from optical flow

We use RAFT [32] to learn dense optical flow $F_{k,t}$ between keyframe $I_k$ and current frame $I_t$. The optical flow between $I_k$ and $I_{t-1}$ is used as a prior to initialize current flow prediction. However, the predicted flow is not accurate for all pixels and the pose estimation error will increase if the displacement becomes small. Thus we select robust correspondences $(p_k, p_t)$ with good forward-backward flow consistency and moderate flow magnitude [42]:

$$\|F_{k,t}(p_k) + F_{t,k}(p_t)\| < \delta_1, \quad \|F_{k,t}(p_k)\| > \delta_2,$$

where we set $\delta_1 = 0.1, \delta_2 = 3$. We select $I_t$ as a new keyframe if the mean flow of robust correspondences is larger than 30. Benefiting from this keyframe-based scheme, the motion disparity between two frames are increased, enabling more accurate pose and depth estimation.

Given 2D correspondences between $p_k, p_t$, the relative pose $T_k^t = [R|t]$ is computed by solving essential matrix $E$ with RANSAC [9] algorithm:

$$p_t^T K^{-T} E K^{-1} p_k = 0, \quad E = [t] \times R,$$

where $K$ denotes camera intrinsics. The scale of up-to-scale pose $T_k^t$ is recovered by aligning triangulated sparse depth (detailed in Section 3.3) with keyframe depth. However, when confronted with small translation or pure rotation, the 2D-2D estimation fails. In these cases, we recover pose with PnP [19] by minimizing reprojection error:
\[ e_r = \sum ||KT_k^oD_kK^{-1}p_k - p_t||_2, \]

where the 2D correspondences in \( I_k \) are lifted to 3D with depth \( D_k \) (detailed in Section 3.2-3.3) and intrinsics \( K \).

### 3.2. Depth modeling

In this paper, we model the depth estimation and updating in a unified Bayesian framework. The inverse depth \( z_i = \frac{1}{\sigma_i} \) of every pixel \( i \) is used since it obeys Gaussian-like distribution and is more robust to distant objects. For inverse depth measurement \( z_i^t \) at time \( t \), we model the good measurement as Gaussian distribution around the ground truth \( z_i \) while the bad one is regarded as observation noise which is uniformly distributed within the interval \([z_i^{\min}, z_i^{\max}]\). For every new observation \( z_i^t \), the probability of being a inlier is \( \rho_i^t \). Thus \( z_i^t \) is modeled as [10]:

\[ p(z_i^t | z_i, \rho_i^t) := \rho_i^t \mathcal{N}(z_i^t | z_i^t, \tau_i^2) + (1 - \rho_i^t) \mathcal{U}(z_i^t | z_i^{\min}, z_i^{\max}), \]

where \( \tau_i^2 \) denotes the variance of a good measurement. We follow [10] to set inverse depth variance \( \tau_i^2 \) as the photometric disparity error of one pixel.

During online inference, we seek to find the Maximum A Posteriori (MAP) estimation of \( z_i^t \) at each timestep, which can be approximated [34] by the product of a Gaussian distribution for \( z_i^t \) and a Beta distribution for inlier ratio \( \rho_i^t \):

\[ q(z_i^t, \rho_i^t | a_i^t, b_i^t, \mu_i^t, \sigma_i^t^2) := \text{Beta}(\rho_i^t | a_i^t, b_i^t) \mathcal{N}(z_i^t | \mu_i^t, \sigma_i^t^2), \]

where \( a_i^t, b_i^t \) are the parameters in Beta distribution, and \( \mu_i^t, \sigma_i^t^2 \) the mean and variance of Gaussian depth estimate.

The depth of keyframe is initialized with single-view estimation \( d_0^k \in D_k^0 \) and inverse depth uncertainty \( \sigma_0^t \in \Sigma_k \) from DepthNet as follows:

\[
\begin{align*}
\mu_i^0 &= \frac{1}{d_i^k}, & \sigma_i^0 \in \Sigma_k, & z_i^{\max} = \mu_i^0 + \sigma_i^0, \\
z_i^{\min} &= \begin{cases} 
\mu_i^0 - \sigma_i^0, & \text{if } \mu_i^0 - \sigma_i^0 > 0 \\
1e^{-6}, & \text{else}
\end{cases}
\end{align*}
\]  

(6)

During adaptation, the DepthNet online learns the prior knowledge of the new scene geometry. Besides, the learned uncertainties can also serve to gauge the reliability in probabilistic depth fusion.

### 3.3. Online depth refinement

Given the relative pose \( T_k^t \) and 2D correspondences, the subsequent depth estimation of keyframe can be further calculated by two-view triangulation [26]:

\[ d_i^k = \arg \min_{d_i^k} \{ \text{dis}(L_k, d_i^k)^2 + \text{dis}(L_t, d_i^t)^2 \}, \]

(7)

where \( \text{dis}() \) denotes the distance between \( d_i^k \) and two camera rays \( L_k, L_t \) generated from 2D correspondences. The mid-point triangulation is naturally differentiable, enabling our VO framework to perform end-to-end online learning.

The triangulated depth map \( D_k^t \) is usually very sparse (~2000 points) and we densify each point with a local \( 3 \times 3 \) patch \( D_k^t \). The depth of each patch pixel is assumed the same as the central point. The patch-based representation allows larger region of depth filtering and provides more valid gradients with a wider basin of convergence.

During online adaptation, \( D_k^t \) is used to update the prior depth estimate to get a MAP estimation \( \tilde{z}_i^t \) according to Eq. 5 as illustrated in Fig. 3. Meanwhile, the parameters \( a_i^t, b_i^t, \mu_i^t, \sigma_i^t^2 \) in Eq. 5 are incrementally updated by Bayesian formulation. The updating method can be found in the supplementary materials. We assume the inverse depth \( z_i^t \) have converged to the ground truth \( z_i^t \) once the uncertainty \( \sigma_i^t^2 \) is lower than a threshold.

### 3.4. Photometric residuals with learned uncertainty

Given the estimated pose \( T_k^t \) and refined depth \( D_k \), one can synthesize \( I_t \) by warping \( I_k \) to the target image \( I_t \) [47]:

\[ p_t \sim KT_k^oD_k(p_k)K^{-1}p_k, \]

(8)

However, view synthesis builds on the photometric constancy assumption, which is often violated in practice. In order to alleviate this issue, we regard these corner cases as
observation noise and use deep neural network to predict a posterior probability distribution \( p(I | \mu_t, \gamma) \) for each RGB pixel parametrized by mean \( \mu_t \) and variance \( \gamma \in \Gamma \) over ground truth intensity \( I \). By assuming the observation noise to be Laplacian, the online learning process can be formulated as minimizing the negative log-likelihood, which can be converted to a weighted photometric loss:

\[
\mathcal{L}_{\text{pho}} = \sum \log p(I | \mu_t, \gamma) = \frac{\| \tilde{I}_i - I_i \|_1}{\Gamma_t} + \log \Gamma_t,
\]  

(9)

where \( \Gamma_t \) denotes photometric uncertainty map.

### 3.5. Differentiable Gauss-Newton optimization

Furthermore, we use a differentiable Gauss-Newton [7] layer to minimize \( \mathcal{L}_{\text{pho}} \) for optimized depth \( D'_k \) and pose \( T'_k \). The predicted \( \Gamma_t \) in Eq. 9 improves the robustness to illumination change and occlusions. Specifically, starting with an initial depth and pose \( D_0, T_0 \), we compute the weighted photometric loss \( r_i(p) \) for each pixel \( p_i \) in all frames \( I_i \) among two keyframes \( I_{k_1}, I_{k_2} \):

\[
r_i(p) = \frac{\tilde{I}_i(p_i) - I_i(p_i)}{\gamma_i}, \quad \gamma_i \in \Gamma_t
\]  

(10)

The first order derivatives with respect to \( D_k \) and \( T'_k \) are:

\[
J_i^I = \frac{1}{\gamma_i} \frac{\partial \tilde{I}_i(p_i)}{\partial p_i} \frac{\partial p_i}{\partial D_k}, \quad J'_i = \frac{1}{\gamma_i} \frac{\partial \tilde{I}_i(p_i)}{\partial p_i} \frac{\partial p_i}{\partial T'_k}
\]  

(11)

Thus the increment \( \delta \) to the current estimation is:

\[
\delta = - (J^T J)^{-1} J^T r, \quad J = [J^I, J']
\]  

(12)

where \( J \) denotes the stack of Jacobians \( \{J_i(p)\} \) and \( r \) denotes the stack of weighted photometric residuals \( \{r_i(p)\} \). The Gauss-Newton algorithm is naturally differentiable and we implement it as a layer in neural network. In practice, we find that it converges within only 3 iterations.

### 3.6. Loss functions

We propose to use the following loss functions to online learn DepthNet and FlowNet in a self-supervised manner.

#### Smoothness loss

We introduce an edge-aware loss for depth and flow to enforce local smoothness:

\[
\mathcal{L}_{\text{smooth}}(G) = \frac{1}{N} \sum_{x,y} \| \nabla_x G(x,y) \| e^{-\| \nabla_x I(x,y) \|} + \| \nabla_y G(x,y) \| e^{-\| \nabla_y I(x,y) \|},
\]  

(13)

where \( G \) denotes optical flow or depth.

#### Depth loss

We derive a loss function of depth by evaluating the negative log-likelihood of the estimated inverse depth \( \tilde{z}_i \) with uncertainty \( \sigma_i \) defined in Eq. 6. This allows the network to attenuate the cost of difficult regions and to focus more on well explained parts. We assume a Laplacian distribution of inverse depth residuals:

\[
p(\tilde{z}_i | \mu_i, \sigma_i) = \frac{1}{2\sigma_i} \exp \left( - \frac{\tilde{z}_i - \mu_i}{\sigma_i} \right)
\]  

(14)

We use refined inverse depth \( z'_k \) as \( \mu_i \) for self-supervision. Thus the negative log-likelihood becomes:

\[
\mathcal{L}_d = \sum - \log p(\tilde{z}_i | \mu_i, \sigma_i) = \frac{1}{\Sigma_k} \| 1/D_k - 1/D_k' \|_1 + \log \Sigma_k
\]  

(15)

Intuitively, the network will tune the depth uncertainty \( \sigma_i \) that best minimize the depth loss \( \| 1/D_k - 1/D_k' \|_1 \) while being subject to the regularization term \( \log \Sigma_k \). In order to enforce depth continuity, we modify Eq. 15 to:

\[
\mathcal{L}_d = \frac{1}{\Sigma_k} \| 1/D_k - 1/D_k' \|_1 + \log \Sigma_k + \mathcal{L}_{\text{smooth}}(D_k)
\]  

(16)

#### Flow loss

The optimized depth and pose \( D'_k, T'_k \) can be used to synthesize optical flow \( F_{k,t} \) by calculating the difference between warped coordinates \( p'_i \) and \( p_t \). We use \( F_{k,t} \) to supervise FlowNet during online adaptation:

\[
\mathcal{L}_{\text{flow}} = \| F_{k,t} - F'_{k,t} \|_1 + \mathcal{L}_{\text{smooth}}(F_{k,t})
\]  

(17)

#### Photometric loss

is defined in Eq. 9. Thus the total self-supervised loss is:

\[
\mathcal{L} = \mathcal{L}_{\text{pho}} + \mathcal{L}_d + \mathcal{L}_{\text{flow}}
\]  

(18)

### 4. Experiments

#### 4.1. Implementation details

##### Network Architectures

Since our method focuses on improving online adaptation of deep VO to achieve better generalization, we adopt similar networks with existing self-supervised VO methods. As for DepthNet, we use the same architecture as Monodepth2 [15] and add a 5 × 5 convolution layer at the output to predict depth uncertainty map \( \Sigma_k \). The optical flow network is based on RAFT [32]. We add a 5 × 5 convolution + Sigmoid layer at output to predict photometric uncertainty \( \Gamma_t \) at the same time.

##### Learning Settings

Our model is implemented by PyTorch [27] on a single NVIDIA GTX 2080Ti. The images are resized to 256 × 832 for KITTI [13] and Cityscapes [4] datasets while set 192 × 256 for TUM dataset [31]. The FlowNet and DepthNet are pretrained in a self-supervised manner for \( 1 \times 10^5 \) iterations according to [28]. The Adam [18] optimizer with \( \beta_1 = 0.9, \beta_2 = 0.99 \) is used. The learning objective (Eq. 18) is used for both pretraining and online adaptation with the learning rate of \( 1 \times 10^{-4} \). The uncertainty maps \( \Gamma_t, \Sigma_k \) are also jointly trained by minimizing Eq. 18. During online adaptation, we retrain FlowNet and DepthNet for 2 iterations in every time step.
4.2. Cityscapes to KITTI

Firstly, we try to test the generalization ability of our framework to different outdoor environments. We pretrain our method on Cityscapes [4] dataset and test on KITTI [13] dataset, which differ not only in scene contents and white balance but also in camera intrinsics. We compare with recent self-supervised VO baselines: GeoNet [39], Vid2Depth [24], Zhan et al. [41], SAVO [22] and Li et al. [21] as well as classic methods: ORB-SLAM2 [26] (with and without loop closure) and VISO2 [14]. Besides, we compare with Zhao et al. [43] and DF-VO [42] which are state-of-the-art methods that combine the output of pretrained networks with classic VO pipeline.

As for pose estimation, we evaluate on 11 KITTI sequences with ground truth poses [39]. It’s worthy to note that all the other VO baselines are pretrained on Cityscapes, while our method is only pretrained on Cityscapes and directly tested on KITTI dataset. Although in such unfair conditions, our method achieves state-of-the-art results even compared with ORB-SLAM2 (LC) (shown in Table 1 and Fig. 4). Meanwhile, different from most self-supervised VO baselines, our method maintains a consistent scale of the entire trajectory. Thus, instead of calculating absolute trajectory error (ATE) on short sequence as previous methods, we align trajectories with ground truth [13] by a single scaling factor and compute translation/rotation error $t_{err}/r_{err}$ on entire trajectory.

Our method outperforms all the other baselines (including end-to-end learning and combination of geometric computation methods) by a clear margin. The rotation and translation errors are an order of magnitude smaller than the other self-supervised baselines, indicating that pose, depth and scale estimation collaborated with probabilistic geometric computation is much better than learning-based inference. As for classic baselines, ORB-SLAM2 is implemented by a local map tracking with bundle adjustment (BA) and ORB-SLAM2 (LC) processes the entire sequence with loop closure, pose graph optimization and global BA to ensure good performance. Our method doesn’t use any optimization backend techniques but it still achieves comparable results with ORB-SLAM2 (LC).

4.3. Outdoor KITTI to indoor TUM

In order to further evaluate the generalization ability to more complex indoor environments, we test on TUM [31] dataset using networks pretrained on KITTI. TUM indoor dataset contains much more complicated motion patterns...
4.4. Depth evaluation on KITTI and NYUv2

We demonstrate the effectiveness of using optimized $D_k$ for self-supervision by evaluating different single-view depth estimation methods on KITTI [13] and NYUv2 [30] datasets. We only use triangulation and Bayesian updating for training. During test, our method predicts single-view depth without refinement. As for KITTI, we take Eigen et al. [6] split for training and test. As for NYUv2, we use the raw training set and evaluate depth prediction results on labeled test set. The predicted depth is multiplied by a scaling factor to match the median with ground truth [6].

Table 3, 4 and Fig. 6 show the depth evaluation results on KITTI and NYUv2 datasets. Benefiting from the patch-based depth triangulation and multi-frame refinement process, our method is able to synthesize refined depth for self-supervision. The learned depth is more accurate and preserves sharper edges with fine details than other methods. More qualitative results and analysis can be found in the supplementary materials.

4.5. Ablation studies

In order to demonstrate the effectiveness of each component, we present ablation studies on various versions of our method on KITTI, TUM and NYUv2 datasets (shown in Table 1, 2, 3, 4). ‘/w/o RDS’ means without the final step of retraining both DepthNet and FlowNet. It can be seen that the performance of pose and depth estimation shows a considerable improvement when the refined depth is used for online training the DepthNet. Besides, it can be noticed that KITTI contains many moving objects (cars, people) and all these datasets have many sequences with changing camera exposure time. The online learned
Table 3. Depth estimation results on KITTI dataset by Eigen et al. [6] split. The results are capped at 80 meters. As for error metrics Abs Rel, Sq Rel, RMSE and RMSE log, lower value is better; as for accuracy metrics δ, higher value is better. w/o RDS: without refined depth for online supervision. w/o PU: without online learned photometric uncertainty.

Table 4. Depth estimation results on NYUv2 dataset. Supervised methods are shown in the first rows. *The inference resolution of P2Net is 288 × 384 with 5-frame left-right flipping augmentation.

5. Conclusions

In this paper, we propose an online adaptation framework for deep VO with the assistance of scene-agnostic geometric computations and Bayesian inference. The predicted single-view depth is continuously improved with incoming observations by Bayesian depth filter. Meanwhile, we explicitly model depth and photometric uncertainties to deal with the observation noise. The optimized pose, depth and flow from differentiable Gauss-Newton layer are used for online self-supervision. Extensive experiments on various environment shifting demonstrate that our method has much better generalization ability than state-of-the-art learning-based VO methods.

Acknowledgments This work is supported by the National Key Research and Development Program of China (2017YFB1002601) and National Natural Science Foundation of China (61632003, 61771026).
References

[1] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunjua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised Scale-Consistent Depth and Ego-Motion Learning From Monocular Video. In NeurIPS, 2019. 4328
[2] Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael I Jordan. Streaming Variational Bayes. In NeurIPS, 2013. 4322
[3] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth Prediction without the Sensors: Leveraging Structure for Unsupervised Learning from Monocular Videos. In AAAI, 2019. 4321
[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In CVPR, 2016. 4322, 4325, 4326
[5] John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. Journal of Machine Learning Research, 12(Jul):2121–2159, 2011. 4322
[6] David Eigen, Christian Puhrsch, and Rob Fergus. Depth Map Prediction From a Single Image Using a Multi-Scale Deep Network. In NeurIPS, 2014. 4327, 4328
[7] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct Sparse Odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(3):611–625, 2018. 4321, 4322, 4325, 4327
[8] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large-Scale Direct Monocular SLAM. In ECCV, 2014. 4321, 4322
[9] 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. 4323
[10] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO: Fast Semi-Direct Monocular Visual Odometry. In ICRA, 2014. 4321, 4322, 4324
[11] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep Ordinal Regression SVO: Fast Semi-Direct Monocular Visual Odometry. In Communications of the ACM, 2014. 4321, 4322, 4324
[12] Jakob Engel, Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are Depth and Surface Normal Estimation from Monocular Images Using Regression on Deep Features and Hierarchical CRFs. In CVPR, 2015. 4328
[13] Andreas Geiger, Julius Ziegler, and Christoph Stiller. StereoScan: Dense 3D Reconstruction in Real-Time. In 2011 IEEE intelligent vehicles symposium (IV), 2011. 4326, 4327
[14] Richard I Hartley. In Defense of the Eight-Point Algorithm. PAMI, 19(6):580–593, 1997. 4322
[15] Jakob Engel, Hannes Risse, and Daniel Cremers. Epnp: An Accurate O (n) Solution to the PnP Problem. IJCV(85):157–165, 2010. 4321, 4322, 4323
[16] Bo Li, Chunjua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and Surface Normal Estimation from Monocular Images Using Regression on Deep Features and Hierarchical CRFs. In CVPR, 2015. 4328
[17] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints. In CVPR, 2018. 4326, 4327, 4328
[18] 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. 4321, 4322, 4324, 4326, 4327
[19] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. PyTorch. https://github.com/pytorch/pytorch. 2017. 4325
[20] Vamip Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation. In CVPR, 2019. 4322, 4325, 4328
[21] Ashutosh Saxena, Min Sun, and Andew Y Ng. Make3D: Learning 3D Scene Structure from a Single Still Image. PAMI, 31(5):824–840, 2008. 4328
[22] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In ECCV, 2012. 4322, 4327
[23] Jürgen Sturm, Nikolaos Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A Benchmark for the Evaluation of RGB-D SLAM Systems. In IROS, 2012. 4322, 4325, 4326
[24] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In ECCV, 2020. 4323, 4325
[25] Sebastian Thrun. Lifelong learning algorithms. In Learning to learn, pages 181–209. Springer, 1998. 4322
[34] George Vogiatzis and Carlos Hernández. Video-Based, Real-Time Multi-View Stereo. Image and Vision Computing, 29(7):434–441, 2011. 4324
[35] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. DeepVO: Towards End-to-End Visual Odometry with Deep Recurrent Convolutional Neural Networks. In ICRA, 2017. 4322
[36] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-Scale Continuous CRFs as Sequential Deep networks for Monocular Depth Estimation. In CVPR, 2017. 4328
[37] Fei Xue, Xin Wang, Shunkai Li, Quyuan Wang, Junqiu Wang, and Hongbin Zhu. Beyond Tracking: Selecting Memory and Refining Poses for Deep Visual Odometry. In CVPR, 2019. 4322
[38] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry. In CVPR, 2020. 4327
[39] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. In CVPR, 2018. 4321, 4326, 4327, 4328
[40] Zehao Yu, Lei Jin, and Shenghua Gao. P2Net: Patch-match and Plane-regularization for Unsupervised Indoor Depth Estimation. In ECCV. 4328
[41] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction. In CVPR, 2018. 4326, 4327, 4328
[42] Huangying Zhan, Chamara Saroj Weerasekera, Jiawang Bian, and Ian Reid. Visual Odometry Revisited: What Should be Learnt? ICRA, 2020. 4323, 4326, 4327
[43] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards Better Generalization: Joint Depth-Pose Learning without PoseNet. In CVPR, 2020. 4322, 4326, 4327, 4328
[44] Yiran Zhong, Hongdong Li, and Yuchao Dai. Open-World Stereo Video Matching with Deep RNN. In ECCV. 2018. 4321, 4322
[45] Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. DeepTAM: Deep Tracking and Mapping. In ECCV. 2018. 4322
[46] Junsheng Zhou, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. Moving Indoor: Unsupervised Video Depth Learning in Challenging Environments. In ICCV, 2019. 4328
[47] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised Learning of Depth and Ego-Motion from Video. In CVPR, 2017. 4321, 4322, 4324, 4328