SAM-Net: semantic probabilistic and attention mechanisms of dynamic objects for self-supervised depth and camera pose estimation in visual odometry applications.

YANG, B., XU, X., REN, J., CHENG, L. GUO, L. and ZHANG, Z.

2022

Pattern Recognition Letters
journal homepage: www.elsevier.com

SAM-Net: Semantic probabilistic and Attention Mechanisms of dynamic objects for self-supervised depth and camera pose estimation in visual odometry applications

Binchao Yang[a], Xinying Xu[a,b,*], Jinchang Ren[a,c,*], Lan Cheng[a], Lei Guo[a], Zhe Zhang[a]

[a]*College of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan 030024, China*
[b]*Shanxi Key Laboratory of Advanced Control and Equipment Intelligence, Taiyuan 030024, China*
[c]*National Subsea Centre, Robert Gordon University, Aberdeen, U.K.*

ABSTRACT

3D scene understanding is an essential research topic in the field of Visual Odometry (VO). VO is usually built under the assumption of a static environment, which does not always hold in real scenarios. Existing works fail to consider the dynamic objects, leading to poor performance. To tackle the aforementioned issues, we propose a self-supervised learning-based VO framework with Semantic probabilistic and Attention Mechanism, SAM-Net, which can jointly learn the single view depth, the ego motion of camera and object detection. For depth estimation, semantic probabilistic fusion mechanism is employed to detect the dynamic objects and generate the semantic probability map as a prior before feeding it to the network to generate a more refined depth map, and attention mechanism is explored to enhance perception ability in spatial and channel view. For pose estimation, we present a novel PoseNet with the atrous separable convolution to expand receptive field. And the photometric consistency loss is employed to alleviate the impact of large rotations. Intensive experiments on the KITTI dataset demonstrate that the proposed approach achieves excellent performance in terms of pose and depth accuracy.

Keywords: Visual odometry, Self-supervised deep learning, Object detection, Semantic probabilistic map, Attention mechanism

1. Introduction

Object detection is the process of identifying object instances such as vehicles, animals and people in videos. It allows for the recognition, localization, and detection of multiple objects. It is generally utilized in applications such as advanced driver assistance systems. Camera motion and scene understanding are a fundamental research topic in machine perception and navigation [1]. Accurate self-localization is a prerequisite for reliable mobile autonomy and is especially significant in situations where Global Positioning Systems (GPSs) signals are unavailable [2].

Simultaneous Localization and Mapping (SLAM) and Vision Odometry (VO) serve as the basis for many emerging technologies such as Unmanned Aerial Vehicles (UAVs) [3]. Among various implementations that rely on diverse sensors, such as the Laser Radar-based (LiDAR-based), and Inertial Measurement Unit (IMU) [3, 4], vision-based self-localization is cheap and compact [2]. Relatively, monocular VO [5-11] has irreplaceable advantages in the low cost and applicability [3, 12].

Conventional solutions to estimate the position of objects and scene geometry involve accurate image correspondence between the corresponding frames, whereas they often fail in challenging environments when there are low texture features [13].

To achieve a robust VO, some well-studied and CNN-based methods have been proposed [3, 10, 12, 16] in recent years, aiming to use data-driven learning models to solve the shortcomings of conventional methods. In certain situations, such as a small amount of data, conventional algorithms may perform better than deep learning algorithms. However, when the amount of data is sufficiently large, they are not as robust as deep learning algorithms because of the insufficiency of the hand-crafted features.

In CNNs, the loss function usually plays a key role [13]. Most of the VO methods reconstruct images based on the photometric consistency [2, 8, 10, 16-18], and the loss function is constructed by the temporal or spatial photometric consistency, making self-supervised training possible. By learning directly from the data, these learning-based techniques have the potential to relax the assumptions that classical VO pipelines rely on, and as a result, to be robust to moving objects, and poor illumination.

Recently, to reduce the reliance on the ground truth, self-supervised methods that use the novel view synthesis as the principal supervisory signal have been presented. These methods explore the inherent redundancy among some sub-problems of 3D scene understanding, which can be constrained via the nature of world regularities [13]. Meanwhile, only when the intermediate

* Corresponding authors.
E-mail address: xuxinying@tyut.edu.cn (X. Xu), jinchang.ren@ieee.org (J. Ren)

predictions of the scene geometry and camera pose are consistent with the physical ground, the geometric view synthesis system can work consistently. However, the self-supervised learning requires the scene to be as static as possible without any moving objects, the modeling surface to be Lambert-like, and there is no occlusion between adjacent views [3, 10]. These requirements are usually difficult to be met in practical situations [19].

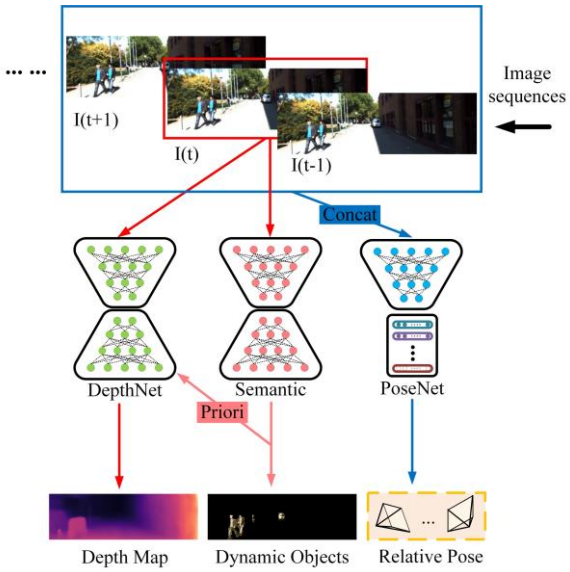

Fig. 1. Diagram of the proposed SAM-Net, which consists of the DepthNet and the PoseNet, taking respectively the current frame and the adjacent frames from a monocular image sequence as input for depth/pose estimation and object detection. After training the model in a totally self-supervised manner, we can estimate the pixelwise depth map, camera's ego-motion, and detect dynamic objects to generate the semantic probability map.

In this paper, we propose a novel VO system namely SAM-Net, in which the pose movement can be obtained from a continuous sequence of color images. By taking an end-to-end approach, the model can estimate camera's ego-motion with a parameterized 6-degrees-of-freedom (6-DoF) transformation matrix and the single view depth, as shown in Fig. 1. Our self-supervised learning does not require any manual interventions or additional information. In comparation to existing models that need a lot of labeled dataset [1, 4], our approach is applicable to a larger range of application scenarios.

Enlightened by the biological resemblance, we propose a Semantic Probability Fusion Mechanism (SPFM) and an Attention Mechanism (AM) in the proposed SAM-Net. SPFM employs image semantic information for geometric estimation, making the network to learn more inclined to static pixels rather than dynamic pixels. AM is established to ensure the network to focus more on significant areas during the training process.

The major contributions of our paper are highlighted as follows:

1) We propose a novel monocular VO system in a self-supervised manner. By harnessing temporal geometric constraints, the depth map and relative pose between the monocular image sequences can be jointly estimated.

2) We employ an object detection technique, SPFM, to detect dynamic objects and to generate a more refined depth map as a prior for depth estimation. AM is explored to enhance the perception ability in the spatial and channel view.

3) We present a novel pose estimation network with the atrous separable convolution to expand the receptive field and to reduce data uncertainty whilst strengthening the assumption of photometric consistency to allow the photometric consistency loss more robust to large rotations.

We evaluate the proposed SAM-Net on the KITTI dataset for depth and pose estimation and have achieved various encouraging findings. Firstly, SAM-Net obtains competitive accuracy with the state-of-the-art approaches. Secondly, we find that the atrous separable convolution and the residual structure are highly capable of predicting the pose from input images. Finally, SPFM and AM can significantly improve the quality of the disparity map.

The remaining of this paper is organized as follows. In Section 2, we present the development history of VO and introduce several representative works. Section 3 details our method and the architecture of the proposed network, including the network structure, the semantic prior probability and the loss function. Experimental results are given and analyzed in Section 4. Finally, conclusions are drawn in Section 5.

2. Related work

2.1 Object Detection with VO

Although it is believed that the photometric uncertainty can be learned to capture the intensity changes of each image pixel, thereby enhancing the robustness of the VO system to any observation noise [12], the factors that affect the actual environments are still unresolved. Object detection can help to remove uncertain pixels from self-supervised photometric reconstruction loss, rather than a simple mask.

Object detection in visual SLAM (vSLAM) has received increased attention in recent years. This requires not only to obtain geometric structure in the environment, but also to identify independent individuals with their poses, attributes and other information [7, 38]. Current research on merging object detection in SLAM is usually to incorporate object detection and semantic segmentation as a multi-task learning process [39] or to apply semantic mapping into map building [40]. No previous work has investigated object detection assisted construction of the geometric information, such as the depth and pose. In our proposed SPFM, we can detect dynamic objects and generate semantic probability information into the network to improve the estimation of the depth and pose.

The attention mechanism, originally used for machine translation [41], is used as an extremely effective way to increase the representativeness of the network. The attention mechanism is mainly divided into soft attention [41], hard attention [42] and local attention [43]. It has improved the performance of computer vision tasks such as image classification, object direction [44-46] and semantic segmentation [47]. Atloc [12] associates the features in spatial domain with attention mechanism, which encourages the network to focus on parts of the image that are temporally consistent and robust. Therefore, we hypothesize that the attention mechanism can allow more complex environment modelling and improve depth estimation performance.

2.2 Classical Visual SLAM

Visual SLAM have a wide range of applications in UAVs, automatic pilot, Virtual Reality (VR) and Augmented Reality (AR) [20, 21]. In the past few decades, vSLAM approaches have been widely studied and various complex algorithms have been proposed, such as VO [11, 22-26], location and identification [27-29], complete SLAM system [6, 15, 22], Structure from Motion

(SfM) [30, 31] and others [32]. The state-of-the-art vSLAM approaches can be characterized into two categories: i.e. direct and indirect formulations.

Indirect methods solve the motion estimation problem by first computing certain reliable geometric representations such as the key points and optical flows [14, 33]. Geometric error is then minimized by using these stable geometric representations [11, 15, 33]. Direct methods can directly optimize the photometric error, which is corresponding to the light value received by a camera. The self-supervised learning framework is inspired by the classical direct method, which can eliminates the expensive sparse geometric computation. However, it is less robust than the indirect ones when there are dynamic moving objects, featureless places, and lighting changes in the scene [19].

Classical vSLAM methods still face many challenges due to their shortcomings, such as hand-crafted features, incorrect system modeling, and complex environmental dynamic constraints. In recent years, deep learning is not restricted by manually designed features and has great advantages in advanced features and understanding [12]. Data-driven vSLAM [9-11, 16, 18, 34-36] has drawn remarkable attention due to its potentials in terms of strong learning capability and the robustness to challenging environments and camera movements.

2.3 Data-driven Visual Odometry

One of the most significant development of vSLAM currently is the Visual Odometry (VO). The motion state of the camera is established by analyzing the associated multi-view geometry. Unlike the classical inter-frame estimation in which feature points are extracted and matched to calculate posture motion, new data-driven paradigms for VO replace all of the classical localization pipeline with a learned model. The data-driven VO outputs posture through a designed and trained model instead of the use of the image geometry and complex operations.

Data-driven VO approaches can be characterized into two categories, i.e. supervised and self-supervised. As a differentiable image warping tool, the spatial transformer [37] is applied to efficiently synthesize the reconstructed image, which allows gradients backpropagated from the reconstruction loss. Inspired by the spatial transformer, some researchers [9, 18, 37] recovered the absolute scale of pose estimation by using the binocular image pairs (stereo images). These methods train a network to regress the relative pose movement between a current view and a nearby view by minimizing the loss of photometric reconstruction. Unlike the supervised methods that rely on expensive ground truth data, the self-supervised pipeline deals with the 3D scene understanding tasks using a photometric reconstruction loss to replace the loss based on ground truth [13]. Our network is actually a self-supervised one for its advantages.

Unreliable pixels caused by moving objects in actual environment break the assumption of photometric consistency, which will lead to inaccurate predictions of the motion. To solve this problem, GeoNet [5] designed a cascaded architecture and adaptive geometric consistency loss to adaptively solve the scene rigid flow and object motion. Ricco et al. [17] propose to estimate the depth information from a single image frame, and estimate the pose and mask from a pair of image frames, similar to Zhou et al. [10], where unreliable pixels are ignored in the mask. However, these methods cannot satisfactorily reduce the errors caused by moving objects to photometric projection.

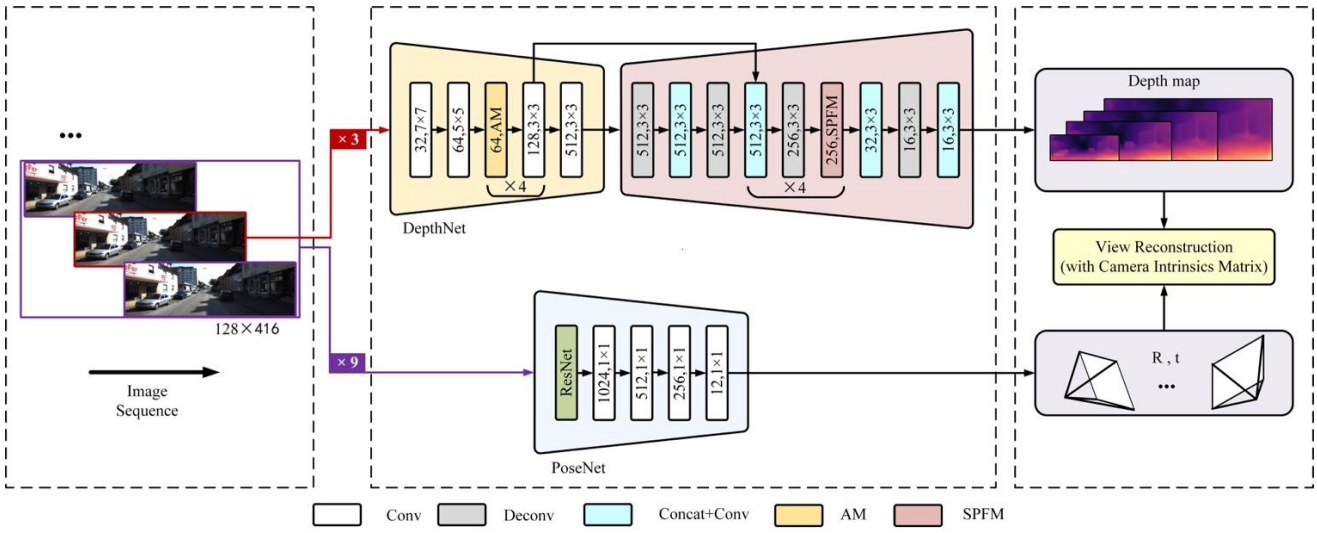

Fig. 2. The detailed network architecture of our SAM-Net. PoseNet outputs the 6-DoF relative poses. DepthNet outputs 4 scale depth maps. Except for the maximum scale depth map of the last layer, the depth maps of the other three scales are outputs from the last three SPFMs of the network.

3. Approach

We propose a VO system with the Semantic probabilistic fusion mechanism and the Attention Mechanism (SAM-Net), in which two sub-problems, monocular depth map prediction and camera's ego-motion estimation, are addressed. The proposed VO system is composed of a pose estimation network (PoseNet) and a depth estimation network (DepthNet). The Semantic Probabilistic Fusion Mechanism (SPFM) is used to detect dynamic objects as a prior for depth estimation, and the Attention Mechanism (AM) is employed to enhance the perception ability in spatial and channel view, as shown in Fig. 2.

In this section, we introduce the design of the network model and the construction of the loss function. Our system merges the self-supervised training procedure of end-to-end VO networks and jointly estimates the pose, depth prediction, and semantic probability map. Despite being jointly trained, the depth estimation network and the pose estimation network can work independently during the testing for flexibility.

3.1. SAM-Net

As depicted in Fig. 1, our approach consists of two independent networks, where the PoseNet take a three frames sequence ($I_{(t-1)}$, $I_{(t)}$, $I_{(t+1)}$) as input, and the DepthNet takes only the middle frame of the sequence as input.

A. SPFM and AM

SPFM: As photometric error is one of the major supervision signals, we also consider to decrease the systematic error caused in the optimization process. To this end, we introduce a novel solution that works well in the experiment. Since dynamic and occluded objects generally exist in images, previous work [10, 17] further train a network to mask out these erroneous regions. Another work [19] proposes a deterministic mask based on the distribution of image reconstruction loss. However, these approaches only bring tiny performance boost because of being entangled with the depth and motion networks and also limited by the loss function design.

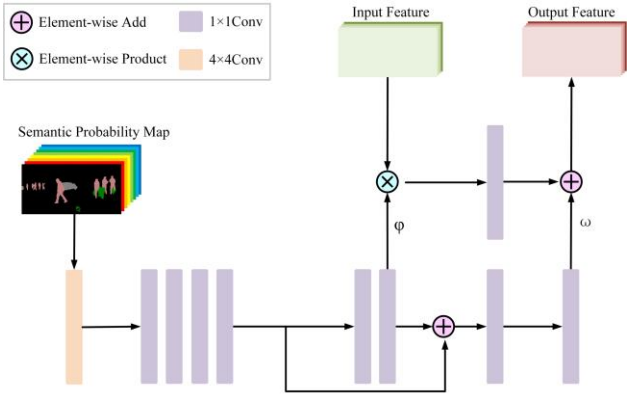

Fig. 3. The structure of semantic probabilistic fusion mechanism (SPFM).

To tackle the insufficiency of the aforementioned methods, we propose to leverage the semantic probability map of dynamic objects as the prior information. Existing methods usually train a CNN to generate a mask to represent the uncertainty of a pixel, where this uncertainty is used as a weight within [0,1] to affect the photometric error loss function. In our model, we employ the DeepLab [51] based training to generate frame-based semantic probability map of dynamic objects, which is then taken as the prior information P. Inspired by Wang et al. [49] which fuses the semantic information in superpixel level super-resolution reconstruction to improve the accuracy, our designed SPFM is based on the spatial feature transform [49], as shown in Fig. 3. Rather than simply predicting the uncertainty of pixels as conventional approaches, taking P as a priori can provide the network with more dynamic object information and solve its damage to the photometric consistency, which will also affect the intermediate feature layer in the form of an affine transformation. To adapt to our network, SPFM is added to the decoder of the DepthNet within the encoder-decoder structure, and context information and semantic information are fused, as shown in Fig. 2.

$$SPFM(F|\varphi,\omega) = \varphi \otimes F + F + \omega \tag{1}$$

$$\varphi = C_{\varphi}\big(\tau(P)\big) \tag{2}$$

$$\omega = C_{\omega}\big(\tau(P)\big) \tag{3}$$

where $\otimes$ is element-wise multiplication, C is convolution, τ is a semantic probability fusion network. φ, ω are the parameter pair generated after the calculation of P, which are used in SPFM to construct the output feature according to the input feature F.

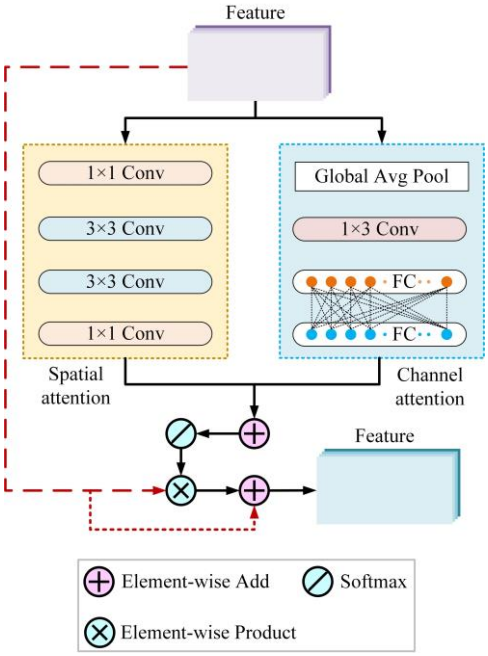

Fig. 4. The structure of the attention mechanism (AM).

Because the spatial dimensions are preserved, SPFM not only operates on the features, but also transforms the spaces [49].

AM: AM in networks has been investigated to estimate the geometric structure of environments [48]. Inspired by this, we explicitly investigate the use of attention as a way to improve DepthNet's representational power in an extremely efficient way. We utilize the Bottleneck Attention Module (BAM) [46] in the encoder to enhance the perception ability in the spatial and channel view.

The objects appearing in the image will bring rich texture information, which is helpful for depth prediction. Multiple AMs can construct a hierarchical attention, in a similar way as the human perception procedure. AM focuses on the exaction of target that is a high-level semantic. This also causes the DepthNet to be more sensitive to semantic targets, and has a clearer depth map at the edge. For efficiency, the receptive field is enlarged by using the atrous convolution. Two attention branches, i.e. the spatial attention branch and the channel attention branch, are used in channels of multiple dimensions to enhance or suppress features at different positions.

The detailed structure is shown in the Fig. 4. The feature map of a given input is $\dot{x} = Ax + Bu$, $F \in R^{C\times H\times W}$, the attention map is $Bam(F) \in R^{C\times H\times W}$, the processed feature map is F':

$$F' = F + F \otimes Bam(F) \tag{4}$$

$$Bam(F) = \sigma\big(B_c(F) + B_s(F)\big) \tag{5}$$

$$B_c(F) = BN(w_1(w_0 AvgPool(F) + b_0) + b_1) \tag{6}$$

$$B_s(F) = BN(f_3^{1\times1}(f_2^{3\times3}(f_1^{3\times3}(f_0^{1\times1}(f_0^{1\times1}(F)))))) \tag{7}$$

where $\otimes$ is element-wise multiplication, σ is the sigmoid function, B_c is channel attention branch, B_s is spatial attention branch, BN is normalization layer, f denotes a convolution operation and the superscripts denote the convolutional filter sizes, $w_0 \in R^{d\cdot C}, b_0 \in R^{c/d}, w_1 \in R^{c\cdot c/d}, b_1 \in R^{c}$, d is the reduction ratio [46].

B. DepthNet

For the depth estimation network, it is based on the encoder-decoder structure to generate the dense depth maps. The encoder architecture consists of stacked convolutional layers to convert images to a compact feature map. In the encoder, AM is employed to strengthen the representation ability in the spatial and channel view. The decoder has stacked upconvolutional layers with skip connections to produce multiscale depth predictions of the input image sequence. The SPFM is utilized to detect the dynamic objects and generate the prior map in the decoder.

Rather than directly generating the depth map as existing approaches, the designed DepthNet can estimate the disparity map. The maximum scale of the depth map is outputted by the convolution from the last feature layer. The depth maps of the other three scales are outputted by convolution from the SPFM with 32, 64, and 128 channels, respectively. The feature map produced by SPFM is up-sampled, connected to the context information, and convolved with the previous depth map, before feeding to the next SPFM.

C. PoseNet

Relative pose estimation is designed to take the three frames sequence concatenated along the color channels (height × width × channels × 3) as input, and the outputs include two 6-DoF poses, corresponding to 3D Euler angles and translation, representing the camera's ego-motion between the middle view and each of its adjacent views.

It is our aim to extract the relative pose between the target frame and each reference frame. If we only take 2 frames as input, i.e. a target frame and a reference frame, and do this for each reference frame, it means that the pose network has only two frames as the temporal context. More consecutive frames in the sample will not be exploited.

The thumb rule is that the lower the number of frames, the easier the algorithm will converge. This is because the displacement of pixels between frames is very small, and thus the photometric loss is always meaningful. However, it will make both the networks less precise because the parallax will also be very small. Finally, there are diminishing returns in adding more frames in a snippet, because at some points the pixel displacement can be so high that quite a few become invisible on both frames. We have also carried out tests with 5-frames and 7-frames snippets, but they only have marginally gain than the 3-frames solution yet with much increased computational cost. As a result, the 3-frame solution is chosen in our experiment.

Pose estimation can continuously track the motion of the camera and generate the relative poses. The global motion trajectory is reconstructed by integrating the relative pose under the existing initialization conditions. For the pose estimation network, it takes the ResNet [50] as the backbone. Chen et al. [51] found that the atrous convolutions can significantly reduce the computational complexity of the model while maintaining similar (or better) performance. Inspired by this, we utilize the atrous convolution in the pose estimation network to expand the receptive field and reduce the computation time. For the input feature map F_{input}, for each position i, the filter is w, after atrous convolution, the output feature map F_{output} is determined by:

$$F_{output}(i) = \sum_j F_{input}(i + s \bullet j)w(j) \qquad (8)$$

where the atrous rate s determines the stride for sampling the input signal.

Since the rotation representation is highly nonlinear, it is relatively difficult to train compared to the translation. Some scholars have proposed to utilize differently fully connected layers for the rotation and the translation of the output layer [18]. Here, we employ the 1 × 1 convolution kernel to map the output feature layer into the pose vector. This not only reduces the amount of calculation, but also achieves good results.

Photometric consistency means the constraint between the corresponding points in two consecutive monocular images of geometric projection. By using this constraint to construct and minimize the loss function, the network can learn the 6-DoF pose and depth maps based on the self-supervised learning.

3.2 Loss Function of Monocular Image

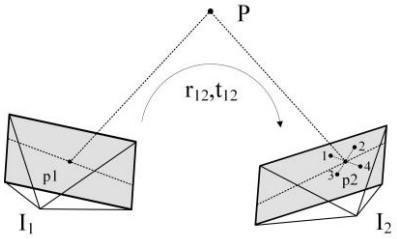

Fig. 5. An illustration of the differentiable image warping process of adjacent frames (pairwise matching and epipolar geometry). We use the differentiable bilinear sampling mechanism proposed by [37], which linearly interpolates the values of the 4-pixel neighbors of p_2 to approximate the photometric of p_2.

As mentioned above, in our framework, the main self-supervised signal comes from the geometric constraints and view reconstruction. The SAM-Net is trained with losses through backpropagation. In this section, the detailed loss function will be introduced. Our approach is distinctive in considering the large rotation in view reconstruction between the middle image and its adjacent images. We present a novel loss function that can establish photometric consistency between temporal view reconstruction, thus improving the accuracy of the joint estimation of the depth map and relative poses. Furthermore, we adopt a depth smoothness loss to ensure the smoothness of the predicted depth.

Here we take two adjacent frames as an example (the three frames are the same two by two). The photometric error loss comes from two adjacent monocular images. The depth estimation network takes the target frame I as input and output its depth map, denotes as D. The pose estimation network takes the concatenated image to generate a 6-DoF relative pose $[r|t]$.

Let us denote a pixel of the first image frame I_1 as p_1, which has a corresponding pixel in the second image frame I_2 as p_2. The pose vector between I_1 and I_2 is $[r_{12}|t_{12}]$. The depth map that I_2 feeds to the DepthNet is D_2, and the depth corresponding to p_2 is $D_2(p_2)$, then p_1 can be estimated from p_2 by:

$$p_1 = K_1[r_{12}|t_{12}]D_2(p_2)K_2^{-1}p_2 \qquad (9)$$

where K_1 and K_2 are the intrinsic matrix for the corresponding two images. We can get the generated image $\hat{I}_2$ through the image I_1, as shown in Fig. 5.

The depth estimation network architecture adopts the idea of multiscale output. In our work, the predicted depth maps are outputted with four scales. Some researchers [19] have pointed out that small-scale depth estimation can produce more photometric error. Therefore, we propose to use different weights for the depth maps of the four scales to reduce such errors. We also verify the impact of small scales on depth prediction in the experiment section.

Early works usually utilize the L1 loss of the corresponding pixels while the Structured Similarity (SSIM) [52] is introduced to

evaluate the quality of predicted images. Furthermore, similar to [18, 19], we adopt the combination of the both L1 loss and SSIM loss as the photometric error loss L_{pho}:

$$L_{pho} = \alpha L^{SSIM}\left(I_k, \widehat{I_k}\right) + (1-\alpha)L^{L1}\left(I_k, \widehat{I_k}\right) \quad (10)$$

$$L^{SSIM}\left(I_k, \widehat{I_k}\right) = \frac{1-SSIM(I_k-\widehat{I_k})}{2} \quad (11)$$

$$L^{L1}\left(I_k, \widehat{I_k}\right) = \left\|I_k - \widehat{I_k}\right\| \quad (12)$$

$$SSIM(x,y) = \mu_y^2 + c_1\frac{(2\mu_x\mu_y+c_1)(2\delta_{xy}+c_2)}{(\mu_x^2+)(\sigma_x^2+\sigma_y^2+c_2)} \quad (13)$$

where α is a balancing factor [5, 19, 34], I_k is a frame in the image sequence, $\widehat{I_k}$ is the synthesis image generated by the source frames I_k by bilinear sampling. μ_x is the average of x, μ_y is the average of y, σ_x^2 is the variance of x, σ_y^2 is the variance of y, and δ_{xy} is the covariance of x and y, $c_1 = (k_1L)^2$, $c_2 = (k_2L)^2$ are constants used to maintain stability, L is the dynamic range of pixel values, $k_1 = 0.01$, $k_2 = 0.03$ are constants [52] .

In order to solve the gradient locality problem in motion estimation [53] and eliminate the discontinuity of learning depth in low-texture regions, we consider a smoothing term in the loss formula. Zhou et al. put forward two ideas to solve this problem: i) using a convolutional encoder-decoder architecture with a small bottleneck for the depth network that implicitly constrains the output to be globally smooth and facilitates gradients to propagate from meaningful regions to nearby regions; ii) explicit multi-scale and smoothness loss [8, 16] that allows gradients to be derived from larger spatial regions directly [10]. Some studies have adopted the method of smoothing loss, using the image gradient to weight the depth gradient [5, 10, 19]. We adopt the second strategy and smooth the item by minimizing the L1 norm of the second gradient of the predicted depth map [17] as follows:

$$L_{smo} = \beta\left(\left|\partial_m\widehat{D}_K\right|e^{-\|\partial_m I_K\|} + \left|\partial_n\widehat{D}_K\right|e^{-\|\partial_n I_K\|}\right) \quad (14)$$

where β is the weight of depth smoothing loss, $\widehat{D}_K$ is the estimated depth map, m and n are gradient direction.

Finally, the samples incorporate large rotations are less common than samples with smaller rotations, though they are obviously more important for pose estimation. Inspired by Wagstaff et al. [2], we utilize the large rotations loss to increase their relative weight compared to other samples. The loss term L_{rot} is the same as for photometric reconstruction, but is set to zero for all samples except those with large rotations, see below.

$$L_{rot} = \begin{cases} \gamma L_{pho}, & \|log(r)^v\| > \Omega \\ 0, & \|log(r)^v\| < \Omega \end{cases} \quad (15)$$

where γ is the weight of the photometric error loss during large rotation, r is the amount of rotation, Ω is the threshold of rotation.

Therefore, our total loss consists of three components, i.e. "Photometric Error Loss" L_{pho}, "Depth Smoothness Loss" L_{smo}, "Large Rotation Loss" L_{rot}. Then, the final loss function L_{all} can be formulated as:

$$L_{all} = L_{pho} + L_{smo} + L_{rot} \quad (16)$$

4. Experiments

In this section, we compare our SAM-Net with several state-of-the -art methods [3, 9, 10, 16, 54-56], and the results are given and analyzed below for valeting its efficacy.

4.1 Datasets and Experimental Settings

A. Datasets

Like many prior works, we evaluated the proposed SAM-Net on the KITTI dataset [20], so far the world's largest dataset for computer vision algorithm evaluation in autonomous driving scenario. It contains real image data collected in scenes such as urban areas, villages and highways, including the original input raw image, LiDAR 3D point cloud data and camera movement trajectory. KITTI Odometry dataset is used for pose testing. This dataset includes a sequence of 11 driving scenes with the ground truth of pose and depth. We adopted the Eigen split [54] of the raw KITTI dataset with 28 scenes and 697 test frames for the evaluation of depth prediction. These test frames do not include training frames. We also excluded all static sequence frames with average optical flow value less than one pixel from the test scene for training. The images of every three frames were used as an independent training sequence for the network. There are 34384 sequences in total. We utilized 26652 sequences for training and 7732 sequences for validation testing.

Cityscapes [57] is another well-known benchmark comprising a large set of video sequences recorded in streets from 50 different cities. PASCAL VOC 2012 [58] provides standardized image data sets for object class recognition. After training the Deeplab on PASCAL VOC 2012 dataset, a fine-tuning is done on Cityscapes to make the trained model generate 8 dynamic categories in streets and more suitable to other situations.

B. Training Details

We implemented the SAM-Net in PyTorch [59]. During the training, we utilized the Adam [60] optimizer with $\beta1$ = 0.9, $\beta2$ = 0.999, the initial learning rate of 0.0002 and decreases by 90% every 50 epoch, and the mini-batch size of 4. The training typically converges after about 200K iterations. For the parameters in the loss function, we selected $\alpha = 0.85$, $\beta = 0.1$, γ =4. All the experiments were performed with image sequences captured with a monocular camera. We resized the images to 128 × 416, but both the depth and pose estimation networks can be run fully-convolutionally for images of arbitrary size at test time. We randomly cropped and rotated the images of the dataset to improve generalization ability. We calculated the ground-true of depth map through LiDAR and calibration, because LiDAR and fixed calibration are very reliable. In order to reduce the difficulty of model calculation, we adopted an ingenious method of transfer learning [61] to achieve the training of the pose estimation network. Based on the pre-training model, the ResNet-101 was trained on the KITTI dataset.

C. Evaluation Index

We evaluated our method using the same evaluation criteria as in [19], which include the Absolute Relative Error (Abs Rel), the Absolute Difference Error (Abs Diff), the Root Mean Squared Error (RMSE), the Square Relative Error (Sq. Rel), and the prediction accuracy δ (A1,A2,A3), as defined below.

$$Abs\ Rel = \frac{1}{n}\sum_k \frac{|y^{pred}-y^{gt}|}{y^{pred}} \quad (17)$$

$$Abs\ Diff = \frac{1}{n}\sum_k |y^{pred} - y^{gt}| \quad (18)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_k \|y^{pred} - y^{gt}\|^2} \quad (19)$$

$$Sq\ Rel = \frac{1}{n}\sum_k \left\|\frac{y^{pred}-y^{gt}}{y^{pred}}\right\|^2 \quad (20)$$

$$RMSE\ Log = \sqrt{\frac{1}{n}\sum_k \|\log(y^{pred}) - \log(y^{gt})\|^2} \quad (21)$$

$$\delta = max\left(\frac{y^{pred}}{y^{gt}}, \frac{y^{gt}}{y^{pred}}\right) < \text{thr} \quad (22)$$

where n is the total number of pixels in image k, y^{pred} and y^{gt} are the estimated depth value and the label. For δ, three different thresholds (1.25^1, 1.25^2 and 1.25^3) are utilized as a convention in the literature [10], which actually represents the accuracy of depth estimation. It counts the percentage of pixels whose ratio of the predicted depth value to the true value is less than the threshold. The closer to 1 the better the predicted result.

4.2 Ablation Study

Regarding the selection and optimization of the DepthNet and the PoseNet, we compare the accuracy differences caused by different backbones, as shown in Table 1. As seen, the encoder-decoder structure can improve the depth prediction. The deeper residual network structure also has better performance than the shallower feature extraction network.

We show that adding AM improves the depth estimation (the first three items of Fig. 6), namely Base, Am-Encoder, Am-Decoder, it means that the network without AM, and AM is respectively added to the encoder and AM is added to the decoder. And then we compared the performance of the introducing SPFM in three positions of the depth estimation network (on the basis of Am-Encoder), namely Spfm-Encoder, Spfm-Decoder, Spfm-Context. In order to verify whether the small-scale depth map causes greater error to the photometric error loss. We directly set the minimum scale loss weight of the four-scale predicted depth map to zero (Small-Depth-Zero).

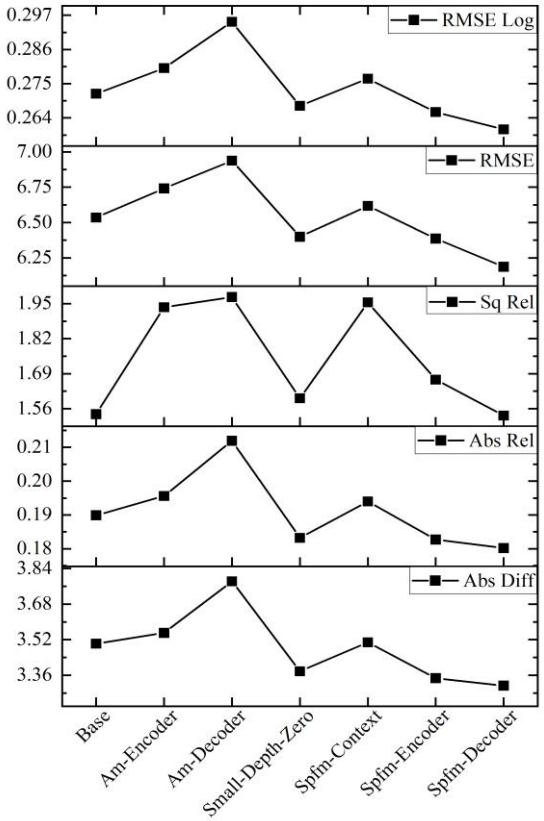

Fig. 6. Three ways to incorporate geometric constraints, compared with baseline method with AM & SPFM and without AM & SPFM. Single-view depth estimation performance. The experiment was conducted under self-supervision. The depth estimation is limited to 80m.

First, we found that adding AM and SPFM with a smaller scale could not improve the training apparently, so we added AM and SPFM on 4 scales. From Table 2 we can clearly see that, after introducing the SPFM and AM, the depth estimation and pose estimation results have been improved. Experiments have shown that the small-scale does cause small errors in the prediction results. This also inspired us to leverage multi-scale weights (using different weights on the 4 depth prediction scales). Note that the SAM-Net proposed here is denoted as Spfm-decoder.

Second, although the depth estimation network and the pose estimation network adopt an end-to-end manner, sometimes the two networks cannot produce the best model at the same time. This is because the pose estimation network can be converged earlier than the depth estimation network. This phenomenon also explains why the pose evaluation results and depth evaluation results sometimes cannot be the best at the same time when performing the model evaluation (see row 3 of Table 2.). But in most cases, the network reaches convergence almost simultaneously.

Table 1. Depth/Pose estimation performance of the ablation study. Absolute Trajectory Error (ATE) and Relative pose Error (RE) on the 09 sequences of the KITTI odometry dataset. The best results are shown as black bold.

Method	RMSE	Sq Rel	A1
DepthNet-UNet	6.631	1.733	0.728
DepthNet-SegNet	7.023	1.950	0.699
DepthNet-E/Decoder	**6.187**	**1.535**	**0.769**
Method	Seq. 09-ATE		Seq. 09-RE
PoseNet-VGG	0.019±0.013		0.006±0.009
PoseNet-ResNet50	0.017±0.008		0.004±0.005
PoseNet-ResNet101	**0.016±0.007**		**0.003±0.003**

Table 2. Pose estimation performance of the ablation study. The best results are shown as black bold.

Method	Seq. 09		Seq. 10	
	ATE	RE	ATE	RE
Base	0.020±0.006	0.004±0.004	0.016±0.009	0.004±0.005
Am-Encoder	0.017±0.006	0.004±0.003	**0.013±0.008**	0.003±0.004
Am-Decoder	**0.015±0.011**	0.004±0.002	0.013±0.010	0.003±0.003
Small-Depth-Zero	0.018±0.008	0.004±0.003	0.014±0.010	0.004±0.005
Spfm-Context	0.019±0.007	0.006±0.006	0.015±0.010	0.005±0.008
Spfm-Encoder	0.019±0.007	0.004±0.003	0.015±0.009	0.004±0.004
Spfm-Decoder	0.016±0.007	**0.003±0.003**	**0.013±0.008**	**0.003±0.003**

4.3 Comparative Study

A. Depth Estimation

In depth estimation, we verified the prediction results of the maximum depth distance of 80m and 50m. Table 3 reports the depth accuracy of our framework on the KITTI dataset for quantitative evaluation. As seen, our proposed SAM-Net obtains the best performance on depth map and significantly outperforms other models. As shown in Fig. 7, after adding the SPFM and AM, the edge perception and object resolution capabilities of the network are significantly improved.

Table 4. ATE on KITTI Odometry. The fourth row of data is the dataset mean of car motion using ground-truth odometry. Our method has achieved better or similar performance than other methods, where the best results are shown in bold.

Method	Seq.09	Seq.10
ORB-SLAM [14](full)	0.014±0.008	0.012±0.011
ORB-SLAM [14](short)	0.064±0.141	0.064±0.130
ORB-SLAM2 [15]	**0.014±0.008**	0.012±0.011
Mean Odom	0.032±0.026	0.028±0.023
Zhou et al. [10]	0.021±0.017	0.020±0.015
Ours	0.015±0.018	**0.011±0.010**

B. Pose Estimation

Pose estimation also shows similar performance trends as depth estimation, and the quantitative results are shown in Table 4. The estimated trajectories of Seq.09 and Seq.10 on the KITTI

Odometry dataset produced by the ground truth, SfM-learner [10], and our methods are plotted in Fig. 8. We first measured the ATE over 3 or 5 frame snippets. Since the relative motion recovered by the monocular vision range system has an undefined scale, we utilized the similarity conversion of the evaluation package evo[1] to align the trajectory with the ground truth.

We generated and visualized the global pose trajectories by accumulating the predicted relative poses. Obviously, our method not only takes advantages in local trajectory estimation (see (a), (b), (c) in Fig. 8.), but also achieves an excellent improvement in the estimation of translation. For Euler angles (see (c) in Fig. 8.), our method also fits the ground truth better than other methods. After introducing the SPFM and AM, the improved network has greatly improved the estimation ability of pose. Our method is superior to other self-supervised monocular VO.

Table 3. Depth estimation results on KITTI using the split of Eigen at al. [54]. The best results among the methods are highlighted in bold.

Method	Supervision	Cap(m)	Abs Rel	RMSE	RMSE Log	Sq Rel	A1	A2	A3
Eigen et al. [54] Fine	Depth	80	0.203	6.307	0.282	1.548	0.702	0.890	0.958
Eigen et al. [54] Coarse	Depth	80	0.214	6.563	0.292	1.605	0.673	0.884	0.957
Liu et al. [55]	Depth	80	0.202	6.523	0.275	1.614	0.678	0.895	0.965
Mahjourian et al. [34]	No	80	**0.163**	6.220	**0.250**	1.240	0.762	0.916	0.968
Zhou et al. [10]	No	80	0.213	6.814	0.292	1.905	0.681	0.884	0.951
Zhou et al. [10] Mask	No	80	0.221	7.527	0.294	2.226	0.676	0.885	0.954
Yang et al. [56]	No	80	0.182	6.501	0.267	1.481	0.725	0.906	0.963
Li et al. [3, 4, 17, 18]	No	-	0.183	6.570	0.268	1.730	-	-	-
Ours	No	80	0.180	**6.187**	0.260	**1.535**	**0.769**	**0.928**	**0.974**
Garg et al. [8]	Pose	50	0.169	5.104	0.273	1.080	0.740	0.904	0.962
Zhou et al. [10]	No	50	0.201	5.181	0.264	1.391	0.696	0.900	0.966
Ours	No	50	0.177	5.125	0.254	1.366	0.762	0.927	0.971

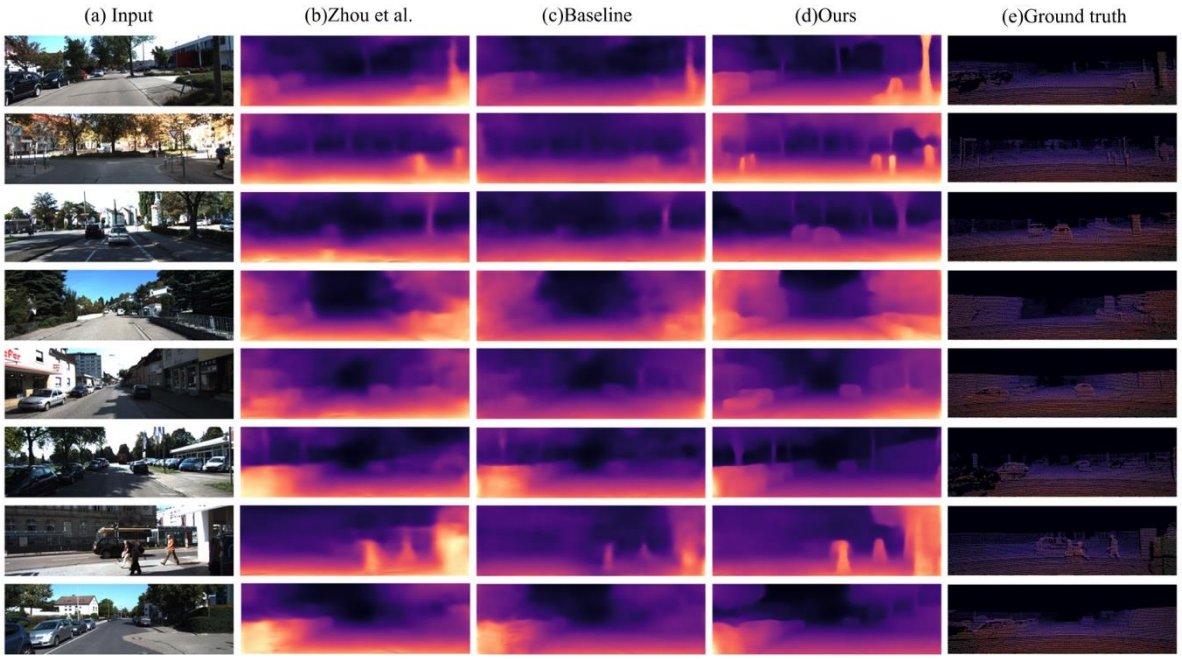

Fig. 7. Comparison of single-view depth estimation between Zhou et al. [10] , ours and the ground truth. These pictures are not in the training dataset. (a) is the raw input image, (b) is the depth map estimated by Zhou et al., (c) is our method without SPFM and AM (d) is our method (SAM-Net), (e) is the ground truth. The depth map estimated by our method has clearer boundaries and is relatively less affected by light.

Different from the ORB-SLAM [14] and ORB-SLAM2 [15] with long-term sliding window optimization, our network is short-term evaluation hence more accurate. As seen in Fig. 8, the ATE is relatively large at the start and end positions. But in the middle part, the ATE is very small, and the relative estimation error is relatively small in the global range. This may be caused by the short sequence we fed to the network, which can estimate the relative position very well, but it is difficult to estimate the global pose (especially the start and end positions). Of course, this is also related to the difficulty of monocular estimation in scale. Considering our network only takes the adjacent three frame pictures as input with a smaller image size, the end-to-end VO system still has great potential for future improvement.

C. Semantics and Attention Mechanism

The attention mechanism can represent the remote spatial location and context information between different feature maps, so that the network can better estimate the depth map. We employed the DeepLab [51] as a dynamic object detection network to generate semantic prior information. The eight categories used for training include person, rider, car, truck, bus, train, motorcycle, and bicycle. There are movable objects that provide a priori information about dynamic objects for the network. The output semantic probability map is used as the input of the SPFM. We visualized the produced probability map and shown the ability of semantic probability to understand the dynamic objects, see in Fig. 9.

[1] https://github.com/MichaelGrupp/evo.

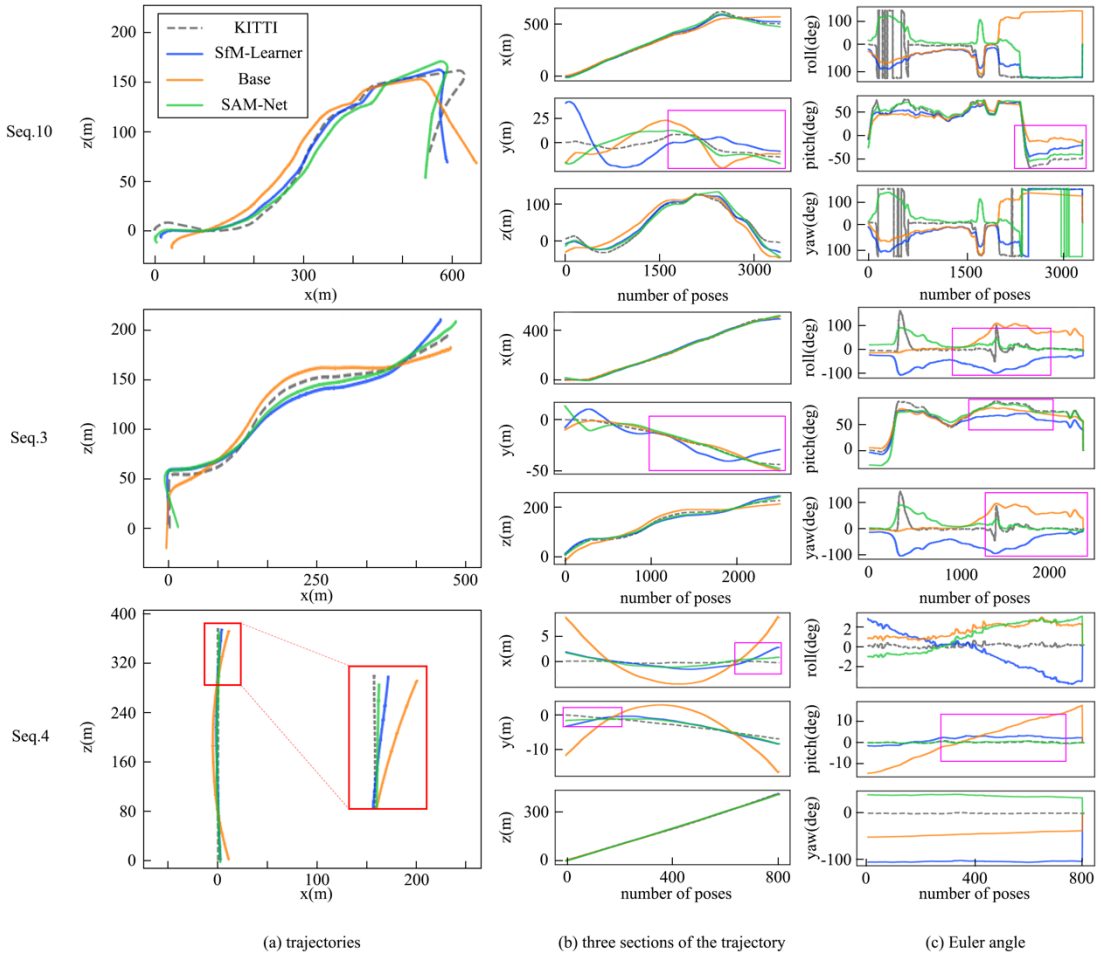

Fig. 8. Trajectories of KITTI sequences #03, #04 and #10 from our model (SAM-Net), SAM-Net without SPFM and AM (Base), SfM-learner [10] and ground-truth trajectories (KITTI) (a). As shown in (b) the three sections of the 3-D trajectory and (c) the Euler angle, our pose prediction is closer to the ground truth even when the SfM-learner failed [10] (purple rectangle). Moreover, after introducing the semantic prior information, pose estimation has achieved better results.

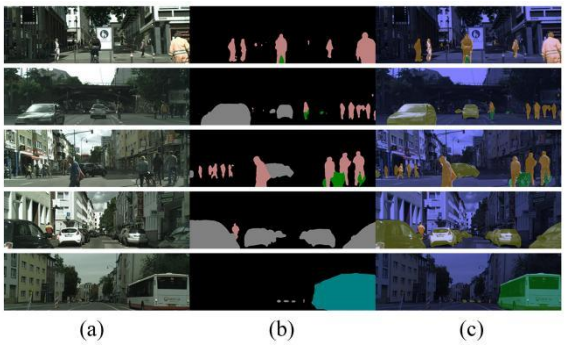

Fig. 9. The semantic probability prior: (a) raw images, (b) semantic probability maps, and (c) unstable dynamic pixels highlighted, showing different movable objects.

5. Conclusion

In this paper, we propose SAM-Net, a self-supervised learning-based VO framework to address the limitations of VO, based on the photometric consistency. Different from the traditional VO, the neural network directly connects the input raw data and the output target, and no manual intervention is required. Specifically, SPFM and AM are utilized to alleviate the impact of photometric inconsistency by dynamic objects. This paper is a preliminary exploration of object detection in the learning problem of self-supervised VO.

Currently, the depth and pose are estimated between consecutive monocular frames. Future research directions include estimating depth and pose problems under long-term distance conditions, multi-modal information fusion neural networks, and realizing the deployment of edge computing devices, etc. Due to the high expressive ability of deep learning, the learning-based SLAM algorithm provides another solution for future robot space environment perception.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China [grant numbers 62073232]; the Shanxi Provincial Natural Science Foundation of China [grant number 201801D12114 and grant number 201801D221190]; and Shanxi Hundred Talents Plan and the Foundation of Shanxi Key Laboratory of Advanced Control and Intelligent Information System [grant number ACEI202101].

References

[1] Y. Wang, et al. Anytime Stereo Image Depth Estimation on Mobile Devices, C. International conference on robotics and automation. (2019)5893-5900.

[2] B. Wagstaff, V. Peretroukhin, and J. Kelly, Self-Supervised Deep Pose Corrections for Robust Visual Odometry, C. Proceedings of the IEEE International Conference on Robotics and Automation, 2020.

[3] N. Yang, et al., D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry, C. Computer Vision and Pattern Recognition, 2020.

[4] S. Wang, et al. DeepVO: Towards end-to-end visual odometry with deep Recurrent Convolutional Neural Networks, C. International conference on robotics and automation. 2017.

[5] Z. Yin and J. Shi. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose, C. Computer vision and pattern recognition. (2018)1983-1992.

[6] A.J. Davison, et al., MonoSLAM: Real-Time Single Camera SLAM, J. IEEE Transactions on Pattern Analysis and Machine Intelligence. 29:6(2007)1052-1067.

[7] Davison. Real-time simultaneous localisation and mapping with a single camera, C. Int. Conf. on computer vision. (2003)1403-1410.

[8] R. Garg., et al. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue, C. European conference on computer vision. 9912(2016)740-756.

[9] H. Zhan, et al. Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction, C. Computer vision and pattern recognition. (2018)340-347.

[10] T. Zhou, et al. Unsupervised Learning of Depth and Ego-Motion from Video, C. Computer vision and pattern recognition. (2017)6612-+.

[11] T. Qin, P. Li, and S. Shen, VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator, J. IEEE Transactions on Robotics. 34:4(2018)1004-1020.

[12] C.W. Chen, Bing & Lu, Chris & Trigoni, Niki & Markham, Andrew, A Survey on Deep Learning for Localization and Mapping: Towards the Age of Spatial Machine Intelligence, C. Computer Vision and Pattern Recognition . 2020.

[13] S. Bompas, B.G., D. Guéry-Odelin, Accuracy of neural networks for the simulation of chaotic dynamics: Precision of training data vs precision of the algorithm, J. Chaos: An Interdisciplinary Journal of Nonlinear Science. 30:11(2020)113118.

[14] R. Murartal, J.M.M. Montiel, and J.D. Tardos, ORB-SLAM: A Versatile and Accurate Monocular SLAM System, J. IEEE Transactions on Robotics. 31:5(2015)1147-1163.

[15] R. Murartal and J.D. Tardos, ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras, J. IEEE Transactions on Robotics. 33:5(2017)1255-1262.

[16] C. Godard, O.M. Aodha, and G.J. Brostow. Unsupervised Monocular Depth Estimation with Left-Right Consistency, C. Computer vision and pattern recognition. (2017)6602-6611.

[17] S. Vijayanarasimhan, et al., SfM-Net: Learning of Structure and Motion from Video, C. Computer Vision and Pattern Recognition, 2017.

[18] R. Li, et al. UnDeepVO: Monocular Visual Odometry Through Unsupervised Deep Learning, C. International conference on robotics and automation. (2018)7286-7291.

[19] T. Shen, et al. Beyond Photometric Loss for Self-Supervised Ego-Motion Estimation, C. International conference on robotics and automation. (2019)6359-6365.

[20] Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite, C. Computer vision and pattern recognition. 2012.

[21] G. Klein and D.W. Murray. Parallel Tracking and Mapping on a camera phone, C. Int. Symp. on mixed and augmented reality. (2009)83-86.

[22] J. Engel,, J. Sturm, and D. Cremers. Semi-dense Visual Odometry for a Monocular Camera, C. Int. Conf on computer vision. (2013)1449-1456.

[23] C. Forster, M. Pizzoli, and D. Scaramuzza. SVO: Fast semi-direct monocular visual odometry, C. Int. Conf. on robotics and automation. (2014)15-22.

[24] M. Li and A.I. Mourikis, High-precision, consistent EKF-based visual-inertial odometry, J. The Int. J. of Robotics Research. 32:6(2013)690-711.

[25] S. Leutenegger, et al., Keyframe-based visual-inertial odometry using nonlinear optimization, J. The International Journal of Robotics Research. 34:3(2015)314-334.

[26] C. Forster, et al., On-Manifold Preintegration for Real-Time Visual--Inertial Odometry, J. IEEE Transactions on Robotics. 33:1(2017)1-21.

[27] W. Zhang and J. Kosecka. Image Based Localization in Urban Environments, C. International symposium on 3d data processing visualization and transmission. (2007)33-40.

[28] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2D-to-3D matching, C. Int. Conf. on computer vision. (2011)667-674.

[29] S. Lowry, et al., Visual Place Recognition: A Survey, J. IEEE Transactions on Robotics. 32:1(2016)1-19.

[30] H.C. Longuethiggins, A computer algorithm for reconstructing a scene from two projections, J. Nature. 293:5828(1987)61-62.

[31] C. Wu, Towards Linear-Time Incremental Structure from Motion, C. International conference on 3d vision. (2013)127-134.

[32] J. Yuan, et al., A Novel Approach to Image-Sequence-Based Mobile Robot Place Recognition, J. IEEE Transactions on Systems, Man, and Cybernetics. (2019)1-15.

[33] C. Campos, et al., ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM, J. IEEE Transactions on Robotics. 2020.

[34] X. Gao and T. Zhang, Unsupervised learning to detect loops using deep neural networks for visual SLAM system, J. Autonomous Robots. 41:1(2017)1-18.

[35] E. Parisotto, et al. Global Pose Estimation with an Attention-Based Recurrent Network, C. Computer vision and pattern recognition. (2018)350-359.

[36] X. Gao and T. Zhang, Unsupervised learning to detect loops using deep neural networks for visual SLAM system, J. Autonomous Robots. 41:1(2017)1-18.

[37] M. Jaderberg, et al. Spatial transformer networks, C. Neural information processing systems. 28(2015).

[38] K. Doherty, et al., Probabilistic Data Association via Mixture Models for Robust Semantic SLAM, C. IEEE Int. Conf. on Robotics and Automation, 2020.

[39] V. Nekrasov, et al., Real-Time Joint Semantic Segmentation and Depth Estimation Using Asymmetric Annotations, C. Computer Vision and Pattern Recognition. (2018)7101-7107.

[40] J. Mccormac, et al., Fusion++: Volumetric Object-Level SLAM, C. Computer Vision and Pattern Recognition. (2018)32-41.

[41] D. Bahdanau, K. Cho, and Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, C. Computation and Language, 2014.

[42] K. Xu, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, C. Int. Conf. on machine learning. 2015.

[43] M. Luong, H. Pham, and C.D. Manning, Effective Approaches to Attention-based Neural Machine Translation, C. Computation and Language, 2015.

[44] L. Bello, et al. Attention Augmented Convolutional Networks, C. International conference on computer vision. (2019)3285-3294.

[45] N. Parmar, et al. Stand-Alone Self-Attention in Vision Models, C. Neural information processing systems. 32(2019).

[46] J. Park, et al., BAM: Bottleneck Attention Module, C. Computer Vision and Pattern Recognition, 2018.

[47] Y. Yuan and J. Wang, OCNet: Object Context Network for Scene Parsing, C. Computer Vision and Pattern Recognition, 2018.

[48] J. Jiao, et al., Look Deeper into Depth: Monocular Depth Estimation with Semantic Booster and Attention-Driven Loss, C. Springer, Cham 2018.

[49] X. Wang, et al. Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform, C. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018)606-615.

[50] K. He, et al. Deep Residual Learning for Image Recognition, C. Computer vision and pattern recognition. (2016)770-778.

[51] L. Chen, et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, C. Euro. Conf. on computer vision. 2018.

[52] Z. Wang, et al., Image quality assessment: from error visibility to structural similarity, J. IEEE Trans. on Image Processing. 13:4(2004)600-612.

[53] J.R. Bergen, et al. Hierarchical Model-Based Motion Estimation, C. European conference on computer vision. 1992.

[54] D. Eigen, C. Puhrsch, and R. Fergus. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network, C. Neural information processing systems. 27(2014).

[55] F. Liu,, et al., Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields, J. IEEE Transactions on Pattern Analysis and Machine Intelligence. 38:10(2016)2024-2039.

[56] Z. Yang, et al., Unsupervised Learning of Geometry with Edge-aware Depth-Normal Consistency, C. Computer Vision and Pattern Recognition, (2018)7493-7500.

[57] M. Cordts, et al., The Cityscapes Dataset for Semantic Urban Scene Understanding, C. Computer Vision and Pattern Recognition. (2016)3213-3223.

[58] M. Everingham, et al., The Pascal Visual Object Classes (VOC) Challenge, J. Int. J. of Computer Vision. 88:2(2010)303-338.

[59] Paszke, et al., Automatic differentiation in PyTorch, C. NIPS 2017 Workshop Autodiff Submission. 2017.

[60] D.P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, C. Learning, 2014.

[61] S.U.H. Dar, et al., A Transfer-Learning Approach for Accelerated MRI using Deep Neural Networks, J. Magnetic Resonance in Medicine. 84:2(2017)663-685.