SELF-SUPERVISED LEARNING OF DEPTH AND POSE USING CYCLE GENERATIVE ADVERSARIAL NETWORK

Yunhe Tong, Anjie Wang, Songchao Tan, Shanshe Wang, Siwei Ma and Wen Gao

Institute of Digital Media, Peking University, Beijing, China

ABSTRACT

In recent years, large amount of ground truth data is typically required to feed into the supervised depth estimation models to produce satisfactory performance, while it is usually costly and impracticable to acquire depth ground truth. In this paper, we propose a self-supervised joint deep learning pipeline for depth and pose estimation of monocular video sequences, which uses the cycle generation adversarial network structure to extend the existing reconstruction loss function based on photometric consistency. The generation function of the algorithm learns to synthesize the adjacent image to predict the depth map and the relative target pose, and the discriminant function learns the dispersion of the monocular images to correctly classify the realism of the composite image. At the same time, a reconstruction loss function based on pose consistency is used to assist the generator function in training. Extensive experimental results on the KITTI dataset show superior performance of the proposed method.

Index Terms— Self-supervised learning, CycleGAN, depth estimation, pose estimation, monocular

1. INTRODUCTION

Both depth map and pose estimation are crucial in computer vision domain, which can not only enrich the representations of objects and environments, but also facilitate many further applications such as 3D reconstruction, virtual reality, autonomous driving, object recognition, tracking and so on.

Many works have been documented on estimating depth map in recent years. The previous works based on stereo images or structures from motion aimed to find corresponding characteristics so as to get relationship between the images nearby. Although the traditional methods are effective and efficient in many cases, they may fail when coming up against the areas with occlusions or lack of structures and texture. Furthermore, depth estimation can be regarded as an ill-posed problem due to the monocular scale ambiguity issue [1].

This work was supported in part by the National Natural Science Foundation of China under Grant 61902006, China Postdoctoral Science Foundation funded project 2019M650346, and High-performance Computing Platform of Peking University, which are gratefully acknowledged.

Benefited from the publication of large-scale datasets [2, 3, 4, 5], depth estimation methods based on Convolutional Neural Networks (CNN) have received a substantial interest in the past few years. Specifically, NYUD [2] presents indoor images while Make3D [3] presents outdoor images. Instead, KITTI [4] and Cityscapes [5] are both collected outdoor images with calibrated stereo cameras. All these datasets are composed of RGB images and corresponding depth maps e.g. ground truth. It is easy to compute loss function between predicted depth map and ground truth, which is referred as supervised learning.

However, large amounts of data are typically required to feed into the supervised learning models to produce satisfactory performance while depth map in dataset is usually costly and impracticable to be acquired. When less ground truth depth maps are used for learning, the depth estimation performance degrades significantly. Hence, Self-supervised learning of a single-view depth prediction has gained more attention. Self-supervised methods implicitly estimate the depth map by images reconstruction from multiple viewpoints, such as adjacent viewpoints in temporal domain or stereo pairs. We can use the photometric consistency between reconstructed views and original views to build the neural network to obtain reliable depth estimation results.

In the binocular stereo matching, by learning the photometric consistency error of the left and right views, these rectified stereo methods show comparable accuracy to the supervised method of learning on datasets with only sparse depth annotations. However, the assumption of using a calibrated binocular pair eliminates the use of monocular video which is easier to be obtained and more versatile.

Part of the work based on the monocular video sequences provides a strategy to learn the individual pose and depth CNN by minimizing the photometric consistency of the monocular video dataset during training. Although impressive results have been achieved, the accuracy has yet to be improved.

In this paper, we obtain depth and pose information through view reconstruction of monocular video sequences. Considering the effectiveness of the generative adversarial network in image generation domain, we combine the consistency constraint of pose with cyclic adversarial learn-

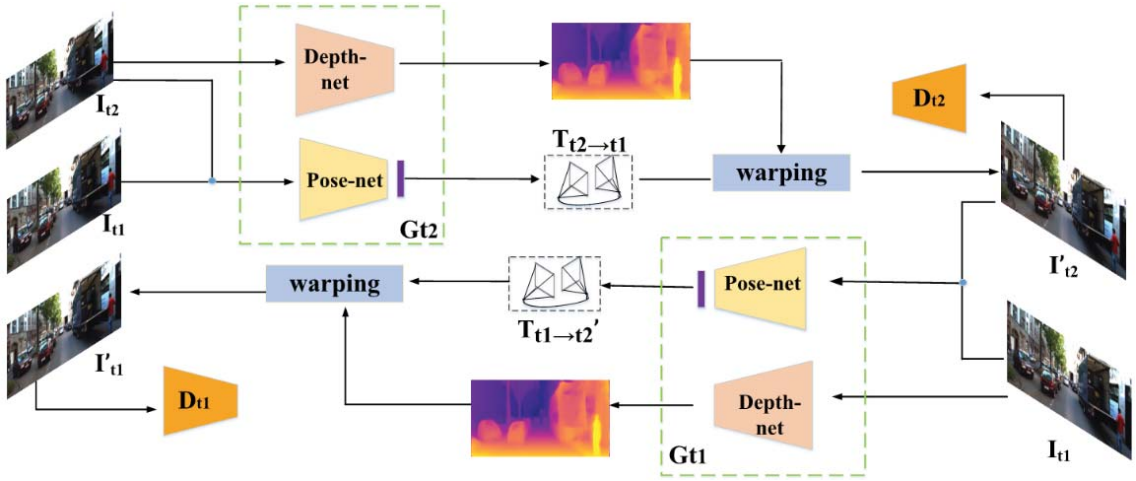

Fig. 1. The pipeline of the proposed depth and pose estimation framework.

-ing. Subjective and objective evaluation of pose and depth on the KITTI dataset shows the effectiveness. Specifically, the main contributions of our method are following: applying the cycle generative adversarial network to the depth and pose estimation of monocular video sequences and proposing the principle of cycle consistency of pose. Our method achieves a significant improvement in pose and depth estimation performance.

2. RELATED WORK

The emergence of neural networks has made rapid development in various fields of computer vision in recent years. In terms of depth estimation, Eigen et al. [6, 7] first propose a single multiscale convolutional network architecture to predict depth map, surface normal and semantic labeling simultaneously.

In binocular stereo pairs, Garg et al. [8] propose an unsupervised deep learning method by using stereo image pairs captured by two cameras. The goal of the network is to minimize the photometric error between the left image and the inverse warped right image which is generated by the predicted depth map and the known pose translation between the stereo cameras. Godard et al. [9] improve the estimation performance by adding the left-right consistency to the loss function of the network.

Then, zhou et al. [10] use single-view depth and multiview pose networks, with a loss based on warping nearby views to the target using the computed depth and pose. Mahjourian et al. [11] consider the inferred 3D geometry of the whole scene rather than only consider pixels in small local neighborhoods, and enforce consistency of the estimated 3D point clouds and ego-motion across consecutive frames. Pilzer et al. [12] present a novel unsupervised approach for predicting depth maps within an adversarial learning framework, which jointly trains two networks with adversarial learning to provide mutual constraints and supervision to each other.

Based on [10, 12], we combine the cyclic generative adversarial learning network with the depth and pose estimation of the monocular video sequences. Through the cyclic generation of the monocular views, the use of photometric consistency loss based on view reconstruction loss, pose consistency loss and cyclical adversarial loss improves the performance of the neural network, simultaneously enhance the performance of pose and depth estimation.

3. THE PROPOSED APPROACH

We propose a novel approach for self-supervised learning of depth and pose using cycle generative adversarial networks. Fig. 1 shows the framework of the proposed method. In this section, we first illustrate the self-supervised learning framework of depth and pose based on view reconstruction, and then combine the consistency constraint of pose to carry out cycle adversarial learning. Finally, we introduce the overall end-to-end optimization objective.

Given a pair of adjacent images I_{t1} and I_{t2} in the temporal domain, the adjacent frames are reconstructed by the image warping process with the help of pose-net and depth-net. For each pixel point p_{t2} in the reference view I_{t2}, we first project it to the adjacent source view I_{t1} according to the predicted depth map and camera pose. Then, the pixel value of the reference view I_{t2} is reconstructed at position x_i using bilinear interpolation. The projected coordinates are obtained by:

$$I'_{t2} = f\left(K, T_{t_2 \to t_1}, D_2, I_{t1}\right), \tag{1}$$

where I'_{t2} is the reconstructed reference view of I_{t2}; K is the camera's intrinsic matrix; D_2 is the depth value of the pixel in the reference view I_{t2} ; $T_{t2 \to t1}$ is the camera coordinate transformation matrix from the reference view to the source view. We can synthesize the reference view I_{t2} from the source view I_{t1} using the estimated pose and spatial transform [13]. Therefore, the reconstruction loss of the syth-

Table 1. Quantitative results on KITTI raw dataset using Eigen-spilt.

Method	Dataset	Supervision	Error Metric (lower,better)				Accuracy Metrics (higher,better)		
			Abs Rel	SqRel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Depth: cap 80m									
Eigen et al. [7] (Fine)	**K**	Depth	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu et al. [18]	**K**	Depth	0.201	1.584	6.471	0.273	0.680	0.898	0.967
Zhou et al. [10]	**K**	Mono	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Kumar et al. [19]	**K**	Mono	0.211	1.980	6.154	0.264	0.732	0.898	0.959
Prasad et al. [20]	**K**	Mono	0.175	1.396	5.986	0.255	0.756	0.917	0.967
Mahjourian et al. [11]	**K**	Mono	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Ours	**K**	Mono	**0.154**	**1.235**	**5.861**	**0.243**	**0.802**	**0.921**	**0.971**

-esize image I'_{t2} can be expressed as:

$$\mathcal{L}_{rec}^{t2} = \sum_{x_i}\{\mathcal{L}_{s\_l1}[I_{t2}(x_i), I'_{t2}(x_i)]\}, \quad (2)$$

where $\mathcal{L}_{S-l1}$ is a combinational loss function which combines the SSIM [14] and L1 norm.

Unlike binocular stereo pair training with known baselines, depth and pose estimation for monocular video sequences have scale ambiguity. In order to solve this problem, we adopt a simple and effective method to normalize the output of the depth-net by applying a nonlinear operator before inputting it to the loss layer [15, 16], which can remove the scale sensitivity problem of the loss function.

Since adversarial learning shows strong capabilities in image generation tasks [12], we use adversarial learning for further optimization in order to improve the quality of image I'_{t2}. The entire synthesis network consists of a generator G_{t2} including a depth-net and a pose-net, and a discriminator D_{t2} which outputs a scalar value to tell whether the image I_{t2} or I'_{t2} is fake or not. Thus, the adversarial objective for the generative network can be formulated as follows:

$$\mathcal{L}_{adv}^{t2}(D_{t2,}I_{t2}, I'_{t2}) = E_{I_{t2}\sim p(I_{t2})}[log D_{t2}(I_{t2})] + E_{I_{t1}\sim p(I_{t1})}[log(1 - D_{t2}(I'_{t2})]. \quad (3)$$

We call the above process as half-cycle structure. In order to make the image reconstruction in the temporal domain implicitly constrain each other, we further adopt the cycle generation network structure in our framework. When the synthesized image I'_{t2} is obtained from the half-cycle network, we further use it as an input of the next cycle generative network G_{t1} that warps I_{t1} into the reconstructed view I'_{t1}. As shown in Fig. 1, the generators G_{t1} and G_{t2} in our framework have similar network structures, which means the network parameters of G_{t1} can be shared with the G_{t2} to build a more brief network model. Then the optimization goal of the two generators in the cycle loop is:

$$\mathcal{L}_{rec} = \mathcal{L}_{rec}^{t2} + \mathcal{L}_{rec}^{t1}. \quad (4)$$

Similarly, we add the discriminator D_{t1} to distinguish the synthesize view I'_{t1} and source view I_{t1}, and use the adversarial learning strategy for the spatial sequences in the cyclic network. The adversarial learning objective of the complete cycle model can be expressed as:

$$\mathcal{L}_{adv} = \mathcal{L}_{adv}^{t2}(D_{t2}, I_{t2}, I'_{t2}) + \mathcal{L}_{adv}^{t1}(D_{t1}, I_{t1}, I'_{t1}). \quad (5)$$

Each half of the loop network produces a pose change corresponding to the movement of the view, namely $T_{t_2\to t_1}$ and $T_{t_1\to t'_2}$. To make them constrain each other, we add an L1-norm consistency loss between these two poses which can be expressed as:

$$\mathcal{L}_{con} = \left\| T_{t_2\to t_1} \right\| - \left\| T_{t_1\to t'_2} \right\|. \quad (6)$$

The entire optimization goal includes the reconstruction loss of the two generators, the adversarial loss of view synthesis and the loss of pose consistency during the loop, which can be written as:

$$\mathcal{L} = \alpha\mathcal{L}_{rec} + \beta\mathcal{L}_{adv} + \gamma\mathcal{L}_{con}, \quad (7)$$

where α, β, γ represent a set of weights for controlling the importance of different optimization parts. Since the loss is based on the above constraints rather than labelled data, the proposed network is a self-supervised learning network.

4. EXPERIMENTS AND ANALYSIS

In this section, we compare our method with state-of-the-art depth estimation and VO (visual odometry) methods on the benchmark KITTI dataset [4], and show the qualitative and quantitative evaluation results.

4.1. Network Architecture

Depth-net in our framework is composed of an encoder and a decoder, which is similar to the network structure in [10]. Skip-connections are designed to fuse the features from different lower layers of the encoder. To obtain a reasonable prediction, the rectified linear unit (ReLU) is used as the activation function after the last prediction layer in the depth-net. The pose-net has the similar network structure as the one described in [17]. The input of the pose-net is two consecutive monocular frames while the regression output is the six degrees of freedom (DOF) pose matrix , which is conversed to a 4x4 transformation matrix. For the discriminators D_{t1} and D_{t2}, we adopt the same network structure in [12]. Five consecutive convolutional operations are designed with a kernel size of 3, a stride size of 2 and a padding size of 1, and batch normalization is performed after each convolutional operation.

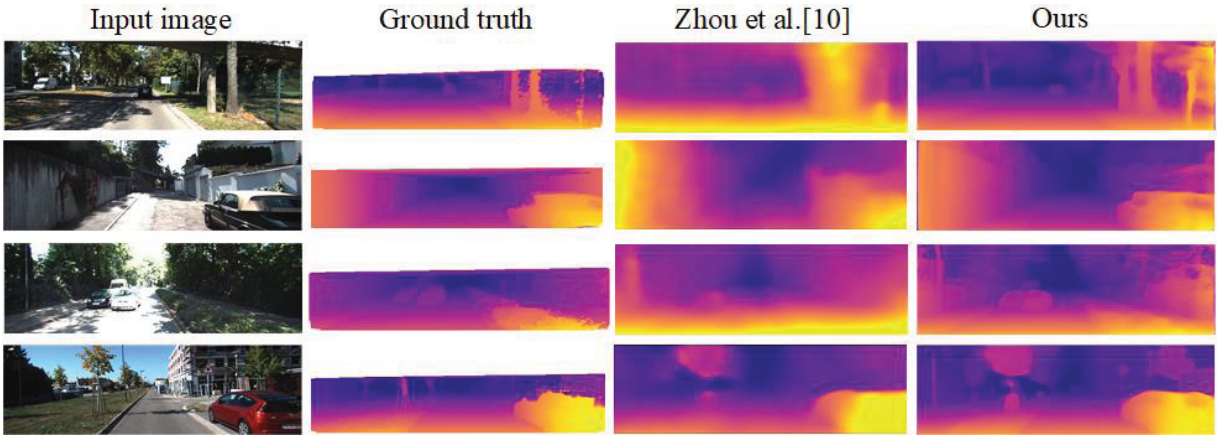

Fig. 2. Compared with the state-of-the-art methods. Here, ground truth depth maps are interpolated for visualization purpose.

4.2. Training details

The initial learning rate of all training networks is set to be 10^{-5} and the weights in the loss function are set as $[\alpha, \beta, \gamma] = [1, 0.1, 0.1]$. In the mixed loss function $\mathcal{L}_{s-l1}$, the ratio of SSIM to L1 is 0.85:0.15. We choose the Eigen-spilt which selects 697 images from the 28 sequences as the test datasets for monocular depth estimation. The original image resolution is down sampled from 1242x375 size to 512x256 size for computational efficiency. The batch size for training is set to 16. Firstly, we train the first half branch with generator G_{t2} and discriminator D_{t2} for a 25k iteration steps. After that we train the second half branch with generator G_{t1} and discriminator D_{t1} for another 25k iterations. Finally we jointly train the whole network with all the losses embedded for a final round of 100k iterations.

4.3. Depth estimation results

To quantitatively evaluate the proposed approach, we follow several standard evaluation metrics used in previous works, as shown in Table 1. To obtain a fair comparison, 80 meters is used as the maximum depth threshold value for metric evaluation. Table 1 and Fig. 2 present the error measurements of different methods and the visualization results of estimated depth maps, respectively.

For the algorithms trained with depth ground truth [7, 18], the direct learning from input image to depth depends entirely on the fitting ability of the neural network. Unlike the monocular depth estimation algorithms [11, 19, 20] which separately use geometric constraints or generative adversarial network, our approach combines the consistent loss of pose with cyclical adversarial loss, and performs higher accuracy and lower error than these methods.

4.4. Pose Estimation Results

We use monocular images to test all the methods for comparison. Since monocular pose estimation methods have the scale ambiguity problem, we conduct post-processing to align their results with ground truth. By tuning the scale factor, each of the short videos is aligned to the ground truth independently.

Table 2. Pose estimation results on the KITTI odometry sequence 09 and 10.

Method	**Seq. 09**	**Seq. 10**
Mean Odom.	0.032 ± 0.026	0.028 ± 0.023
ORB-SLAM [21](full)	0.014 ± 0.008	0.012 ± 0.011
ORB-SLAM [21](short)	0.064 ± 0.141	0.064 ± 0.130
Zhou et al. [13]	0.021 ± 0.017	0.020 ± 0.015
Lu et al. [22]	0.018 ± 0.007	0.014 ± 0.008
Ours	0.014 ± 0.007	0.013 ± 0.010

The camera pose estimation performance can be quantified by Absolute Trajectory Error (ATE). Table 2 shows the comparison results between our proposed method and other advanced methods. Obviously, compared with feature matching method [21] and view photometric consistency constraint method [13, 22], Our approach that uses the cyclical consistency of pose to provide strong supervision for pose estimation can achieve better results.

5. CONCLUSION

We propose a novel approach on self-supervised deep learning for the depth and pose estimation task using the adversarial learning strategy in a cycle generative network structure. With the help of cycle generation of the monocular views, our approach integrates the photometric consistency loss, pose consistency loss and cyclical adversarial loss to improve the performance of the neural network. Firstly, we adopt a self-supervised learning framework based on view reconstruction to learn depth and pose, which is used as a generator to build a cycle generative adversarial network, and then use the adversarial training to improve the performance of the generator. Finally, the entire network is optimized by combining the pose consistency constraint with the cycle reconstruction process. The evaluation of pose and depth on the KITTI dataset shows the effectiveness of our method.

6. REFERENCES

[1] Saxena, S. H. Chung, and A. Y. Ng. “Learning depth from single monocular images,” in Advances in Neural Information Processing Systems (NIPS). IEEE, pp. 1161-1168, 2006.

[2] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. “ Indoor segmentation and support inference from rgbd images,” in European Conference on Computer Vision (ECCV). IEEE, pp. 746-760, 2012.

[3] Saxena, M. Sun, and A. Y. Ng. “Make3d: Learning 3d scene structure from a single still image,” in IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 5, pp. 824-840, 2008.

[4] Geiger, P. Lenz, and R. Urtasun. “Are we ready for autonomous driving? the kitti vision benchmark suite,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 3354-3361, 2012.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. “The cityscapes dataset for semantic urban scene understanding,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 3213-3223, 2016.

[6] D. Eigen and R. Fergus. “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in European Conference on Computer Vision (ECCV). IEEE, pp. 2650-2658, 2015.

[7] D. Eigen, C. Puhrsch, and R. Fergus. “Depth map prediction from a single image using a multi-scale deep network,” in Advances in Neural Information Processing Systems (NIPS). IEEE, pp. 2366-2374, 2014.

[8] R. Garg, V. K. B. G, G. Carneiro, and I. Reid. “Unsupervised cnn for single view depth estimation: Geometry to the rescue,” in European Conference on Computer Vision (ECCV). IEEE, pp. 740-756, 2016.

[9] Godard, O. M. Aodha, and G. J. Brostow. “Unsupervised monocular depth estimation with left-right consistency,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 270-279, 2017.

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

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

[12] Pilzer , D. Xu, M. Puscas, E. Ricci, and N. Sebe. “Unsupervised Adversarial Depth Estimation Using Cycled Generative Networks,” in 2018 International Conference on 3D Vision (3DV). IEEE, pp. 587-595, 2018.

[13] M. Jaderberg, K. Simonyan, A. Zisserman, and K.Kavukcuoglu, “Spatial transformer networks , ” in Advances in Neural Information Processing Systems (NIPS). IEEE, pp. 2017-2025, 2015.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P.Simoncelli. “Image quality assessment: from error visibility to structural similarity,” in IEEE transactions on image processing, vol. 13, no. 4, pp. 600-612, 2004.

[15] Wang, J. Buenaposada, R. Zhu, and S. Lucey. “Learning Depth from Monocular Videos using Direct Methods， ” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 2022-2030, 2018.

[16] J. Engel, T. Schöps, and D. Cremers. “LSD-SLAM: Large-scale direct monocular SLAM， ” in European Conference on Computer Vision (ECCV). IEEE, pp. 834-849, 2014.

[17] R. Li, S. Wang, Z. Long, and D. Gu. “UnDeepVO: Monocular Visual Odometry through Unsupervised Deep Learning， ” in 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 7286-7291, 2014.

[18] F. Liu, C. Shen, G. Lin, and I. D. Reid. “Learning depth from single monocular images using deep convolutional neural fields,” in IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 10, pp. 2024–2039, 2016.

[19] A C Kumar, S M Bhandarkar, M Prasad. “Monocular depth prediction using generative adversarial networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 300-308, 2018.

[20] V. Prasad, B. Bhowmick. “SfMLearner++: Learning Monocular Depth & Ego-Motion using Meaningful Geometric Constraints,” in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 2087-2096, 2019.

[21] R. Mur-Artal, J. Montiel, and J. D. Tardos. “ORB-SLAM: a versatile and accurate monocular SLAM system,” in IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147–1163, 2015.

[22] Y. Lu and G. Lu. “Deep Unsupervised Learning for Simultaneous Visual Odometry and Depth Estimation”, in IEEE International Conference on Image Processing (ICIP). IEEE, pp. 2571-2575, 2019.