Unsupervised Ego-Motion and Dense Depth Estimation with Monocular Video

Yufan Xu
School of Automation Science and Electrical Engineering
Beihang University
Beijing, China
e-mail: 936831611@qq.com

Yan Wang
School of Automation Science and Electrical Engineering
Beihang University
Beijing, China
e-mail: xuyufan@buaa.edu.cn

Lei Guo
School of Automation Science and Electrical Engineering
Beihang University
Beijing, China
e-mail: 936831610@qq.com

Abstract—**In recent years, Deep Learning based method for 3-Dimension (3D) geometry perception tasks, like dense depth recovery, optical flow estimation and ego-motion estimation, is attracting significant attention. Inspired by recent advances in unsupervised strategies to learning from video datasets, we present a reasonable combination of constrains and a finer architecture, used for unsupervised ego-motion and depth estimation. Specifically, we introduce our effective neural networks Depth-Net (for monocular depth estimation) and Pose-Net (for ego-motion estimation), which are trained with monocular images. Depth-Net is proposed by us, improving the accuracy of estimation with as few parameters as possible. Finally, extensive experiments are implement on the KITTI driving dataset, proving our method outperforms some state-of-the-art results in unsupervised even supervised method.**

Keywords-deep learning; depth; ego-motion; unsupervised; geometry perception.

I. Introduction

Traditional scene understanding techniques, like Simultaneous Localization and Mapping (SLAM) [1] have been effective and efficient in ideal environment: rich texture, simple geometry. However, under the circumstances of non-textured regions and changing environments, the performance of these approaches falls sharply. To this end, some deep learning based models attempt to complete these tasks, such as depth [2], optical flow [3], camera pose [4]. These supervised learning methods often consider those objectives as general regression problems, but ignore that these tasks have internal geometric consistency relations between each other instead of being independent. Moreover, all the above methods need ground truth data, which is expensive to obtain.

Recent works have begun to fuse all of those tasks together. [5-6] take advantage of the coupling of these targets, utilizing view synthesis as supervision instead of ground truth, learning the monocular depth and ego-motion even optical flow, obtaining impressive results. However, the accuracy of depth and pose estimation cannot exceed method with calibrated stereo images, although monocular video is easier to obtain than stereo. Inspired of that, we present our effective architecture and constrains to make full use of monocular images. Firstly, we present a reasonable loss function to make full use of image sequences to train. Then we try to modify the architecture in [5], with as few parameters as possible. Finally, we combine these two strategies and the experiment proves the effectiveness.

II. Related Work

Excellent solutions to estimating depth and camera pose have been proposed, which are mainly based on traditional matching method or deep models. Recently, unsupervised learning from unlabeled video sequences or stereo images are gaining increasing attention in 3-Dimension (3D) scene geometry, which employ self-supervision strategy to train, solving single or multi-tasks at the same time.

A. Supervised/Unsupervised Depth Estimation

Classical learning approaches regard this task as a supervised problem, training a convolutional neural network (CNN) to minimize a loss based on the scale invariant RMS [7], or the log RMS [8] of the depth estimations from ground truth. Last several years, unsupervised learning method is popularly used for depth estimation. Authors use real un-annotated imagery to train their network [9]. However, their loss function is not fully differentiable, which leads to training suboptimal. Then, a similar approach was taken by [10], with a left-right consistency constraint, and a differentiable loss that led to impressive performance. However, the assumption of using calibrated binocular image pairs excludes itself from utilizing monocular video, which is easier to obtain and richer in variability.

B. Pose Estimation

Traditional geometry based pose estimation is generally called Visual Odometry (VO), which involves two main directions, feature-based method [11] and direct method [12-13]. Some state-of-the-art algorithms based on this pipeline have shown excellent performance in ideal environments, but when it comes to non-textured regions and other challenging environments, these methods deteriorate sharply.

978-1-5386-7635-6/18/$31.00 ©2018 IEEE

Deep Learning which automatically learns suitable feature representation from large-scale dataset provides an alternative solution to the VO problem.

Given pre-processed optical flow, a CNN based VO system was proposed in [14]. Authors then presented a Recurrent Convolutional Neural Network (RCNN) based VO method [15]. Nevertheless, all of these methods demand the ground truth of pose or depth. Lately, authors [5], who associate tasks for learning depth and pose, utilize view synthesis as supervision instead of ground truth, learning ego-motion and depth from monocular images, although they cannot make full use of the information of monocular images.

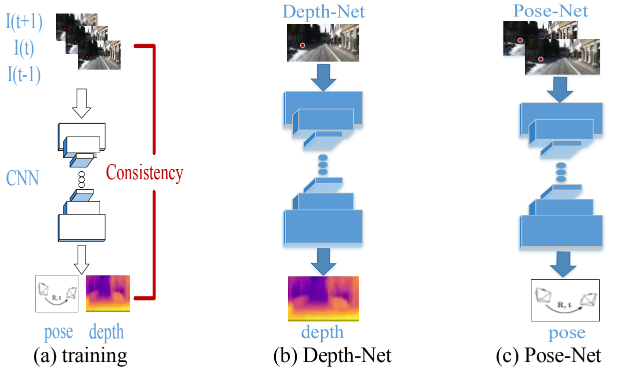
C. *Multi-Task Learning*

After [5] proposes their novel idea of using self-supervision, many works are done based on their work. [16] further proposes 'SfM-Net' to jointly predict depth, segmentation, optical flow, camera and rigid object motion. [17] proposes supervised 'DeMoN' for jointly estimating depth, ego-motion, surface normal and optical flow given two successive views, showing that learning these multiple-tasks jointly leads to better performance on each of the tasks compared to scenarios where each task is learnt in a disjoint fashion. A cascaded architecture consisting of two stages is used to solve the rigid optical flow and object motion in work of [6]. Inspired by related work in direct VO, [18] proposes a simple normalization strategy to recover the scale. However, most of these works focus on the loss function instead of the network.

Based on recent works, we try to learn single view depth and camera pose from monocular images in an unsupervised strategy. Then, we present a finer architecture and a reasonable loss function to learn depth and camera pose from monocular images. Further, the Depth-Net, proposed by us, brings no increase for the number of parameter, while the improvement of accuracy for depth and pose estimation is evident.

III. APPROACH

A. *Pipeline*



(a) training (b) Depth-Net (c) Pose-Net

Figure 1. System pipeline

The pipeline of our system is shown in Fig. 1. In every training session, the input of our approach is a sequence image. The length of the sequence could be 3, 5 or more. Here we use 3 frames for simplicity. Then with the depth for $I(t)$, pose for two image pair$(I(t-1), I(t); I(t+1), I(t))$, we combine loss for view synthesis and depth smoothness to optimize. In every test session, the two network can work individually. The network architecture is modified based on the architecture in [5].

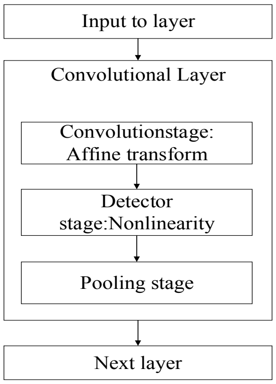

Input to layer
Convolutional Layer
Convolutionstage: Affine transform
Detector stage:Nonlinearity
Pooling stage
Next layer

Figure 2. The components of a typical convolutional neural network layer.

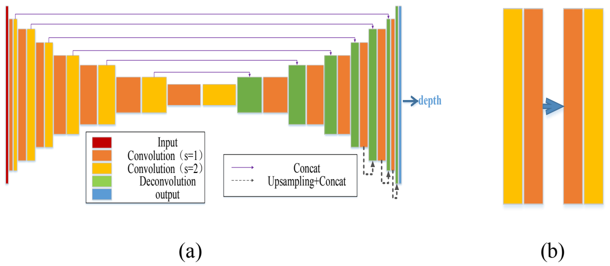


Figure 3. (a) shows the Depth-Net of us and (b) shows the change for the order of convolution pairs.

B. *Network Architecture*

The components of a typical convolutional neural network consist of three levels [19], convolution, activation and pooling. Among these levels, convolution performs as an affine transformation. Meanwhile, pooling has a role to reduce dimensions, expand the sensing field and achieve invariance. When the stride of convolution is two, which is often used as a role of "pooling". General order of these levels is shown in Fig. 2.

Inspired of this, we propose our Depth-Net, changing the convolutional pairs' order of DispNet [20] architecture, which is a single-view depth prediction network, employed by [5]. As shown in Fig. 3, in DispNet [20] architecture, the order of each pair of convolution is to implement convolution with two strides firstly, and then to implement convolution with one stride, while we inverse the order to make it work like the components of a typical CNN. With this change, the performance is evidently improved with zero parameter added.

C. *Loss Function*

1) Vertically Forward-Backward Consistency : In overlapped views, each pair of image sequence is constrained by the depth, pose and intrinsic, which means that every pixel in one image have only one correspondence in other images, as shown in Fig. 4.

TABLE I. DEPTH EVALUATION(S=SUPERVISION, T= TRAINING DATASET, K=KITTI)

			Error Metric				Accuracy Metric		
	S	T	Abs Rel	Sq Rel	RMSE	RMSE log	Acc.1	Acc.2	Acc.3
Ours(N+L) w/o Mask	×	**K**	**0.177**	1.9817	**6.242**	**0.255**	**0.772**	**0.919**	**0.967**
Ours(N) w/o Mask	×	**K**	0.186	1.694	6.548	0.265	0.741	0.905	0.961
Ours(L) w/o Mask	×	**K**	0.182	2.199	6.457	0.264	0.7667	0.9164	0.963
Results with [8]	✓	**K**	0.202	1.614	6.523	0.275	0.678	0.895	0.965
Results with [7] (Fine)	✓	**K**	0.203	**1.548**	6.307	0.282	0.702	0.890	0.958
Results with [7](Coarse)	✓	**K**	0.214	1.605	6.563	0.292	0.673	0.884	0.957
Results with [5] w/o Mask	×	**K**	0.221	2.226	7.527	0.294	0.676	0.885	0.954

Denote I_t, I_{t+1} as the tth and $(t+1)$th image frame, respectively, and $p_t(u_t, v_t)$ as one pixel in I_t, and $p_{t+1}(u_{t+1}, v_{t+1})$ as the corresponding pixel in I_{t+1}. K denotes the camera intrinsic matrix, D_{dep} is the depth value of the pixel in the tth frame, $T_{t,t+1}$ is the camera coordinate transformation matrix from the tth frame to the $(t+1)$th frame. Then, we can derive p_{t+1} from p_t through:

$$p_{t+1} = KT_{t,t+1}D_{dep}K^{-1}p_t \tag{1}$$

Here we predicted the inverse of depth map by our Depth-Net. With this spatial constraint, we could synthesize one image from the other through "spatial transformer network" [21]. to evaluate similarity of two images. Therefore, We can synthesize I_t from I_{t+1} as I_t^{t+1}. Similarly, We can synthesize I_t from I_{t-1} as I_t^{t-1}. Here we use a robust function, combining an L1-norm and single scale SSIM term [22], to evaluate the similarity of two images, instead of directly applying the L1-norm, which is used in [5]. Here, we use a simplified SSIM with a 3×3 block filter instead of a Gaussian, and set α=0.85. L^{l1} is the L1-norm operation and L^{SSIM} is the SSIM operation, which have been used in depth estimation [2].Then the photometric losses between the image sequence are:

$$L_{pho}^{t-1} = \alpha L^{SSIM}(I_t^{t-1}, I_t) + (1-a)L^{l1}(I_t^{t-1}, I_t) \tag{2}$$

$$L_{pho}^{t+1} = \alpha L^{SSIM}(I_t^{t+1}, I_t) + (1-a)L^{l1}(I_t^{t+1}, I_t) \tag{3}$$

2) Smoothness Loss

We encourage disparities to be locally smooth with an L1-norm penalty on the disparity gradients ∂d. As depth discontinuities often occur at image gradients, we weight this cost with an edge-aware term using the image gradients ∂I_t, then the loss is defined as follows:

$$L_{smooth} = \partial_x D_p e^{-||\partial_x I_t||} + \partial_y D_p e^{-||\partial_y I_t||} \tag{4}$$

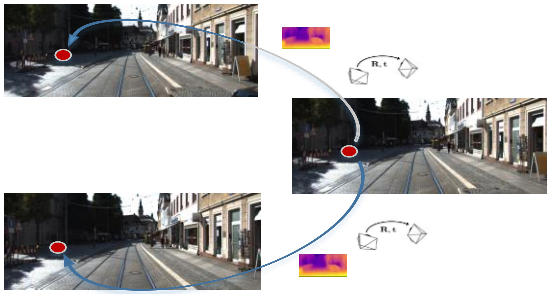


Figure 4. View Synthesis

In addition, all above losses are computed at four output scales.

IV. EXPERIMENT

Here we firstly introduce our training details. Then we will show qualitative and quantitative results in monocular depth, and camera pose estimation tasks respectively.

A. Implementation Details

We implement the system, using the publicly available TensorFlow [23] framework. All the experiments are employed to train the network for up to 20 epochs with parameter $\beta_1 = 0.9$, $\beta_2 = 0.99$, and mini-batch size of 4.We resize the images to 128 × 416 during training. The data is augmented with random scaling, cropping and horizontal flips.

B. Depth

In order to compare the accuracy of depth with state-of–the-art methods, we evaluate depth on the KITTI dataset. Fig. 5 compares sample depth estimates produced by our trained model to other unsupervised and supervised learning methods. We adopt the length of image sequence with 3. The metrics are computed over the Eigen [7] test set. When trained only on the KITTI dataset, our model reduces the mean absolute relative depth prediction error (in meters) from 0.221 [5] to 0.177, which is a significant improvement. We also visually compare depth estimates from our models with the DispNet used by [5]. As Fig. 5 and Table 1 shown, our proposed method achieves significant improvements.

In Table 1, N denotes that use our Depth-Net and loss function from [5], and L denotes that we adopt our loss function and the architecture of [5]. N+L denotes the combination of our loss function and architecture. For error metric, the smaller numbers are better and the bigger numbers are better in accuracy metric.

C. Camera Pose

For pose estimation, we train our model with popular KITTI Odometry Dataset [24], which is the same training and testing dataset as in [5], and the length of image sequences is 5. Table 2 shows the results compared with [5], ORB_SLAM (short and full) [1]. Obviously, our method outperforms both baselines like, ORB_SLAM (short) [1] and ORB_SLAM (full) [1], which leverages whole sequences for loop closure and re-localization. That shares the same input setting as ours. The metric is the Absolute Trajectory

Error (ATE) on the KITTI Odometry split averaged over all 5-frame snippets (lower is better).

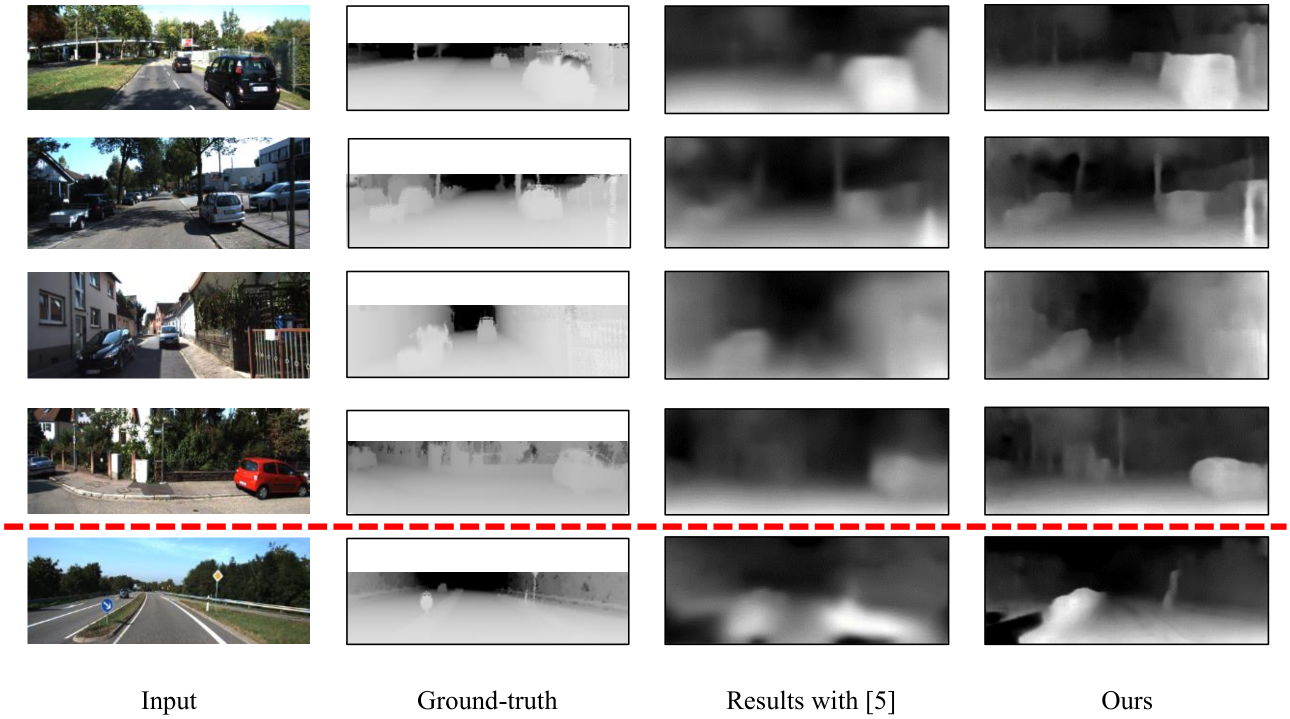


Figure 5. Depth esimation

TABLE II. POSE ESTIMATION(S=SHORT,F=FULL)

Abs	N+L	N	L	Zhou et al.	SLAM(s)	SLAM(f)
seq. 09	**0.0135±0.0073**	0.0143±0.0078	0.0143±0.0078	0.0218±0.0167	0.064±0.141	0.014±0.008
seq. 10	**0.0127±0.0096**	0.0131±0.0099	0.0128±0.0099	0.202±0.052	0.064±0.130	0.012±0.011

V. CONCLUSION

In this paper, we introduce an unsupervised learning method for depth and ego-motion estimation from monocular images. For single view depth estimation, we change the order of inner components (convolution pair), which cause no increase of parameter but evident improvement of accuracy. For the loss function, we take the gradient of source image and the scale problem into the loss. The environment proves the efficiency of our hybrid loss function. Finally, experiment prove that our strategy is effective. In general, unsupervised learning based multi-tasks for 3D perception approaches have the potential to improve their performance with the increasing size of training datasets. In future work, we will investigate how to make full use of stereo images, better combing the horizontally left-right consistency and vertically forward-backward constraints.

ACKNOWLEDGMENT

This work is supported by the National Nature Science Foundation of China [Grant numbers 61320106010, 61573019, 61627810].

REFERENCES

[1] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. ORBSLAM: a versatile and accurate monocular SLAM system. IEEE Transactions on Robotics, 31(5), 2015.

[2] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In CVPR, 2017.

[3] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks.In ICCV, 2015.

[4] A. Kendall and R. Cipolla. Geometric loss functions for camera pose regression with deep learning. CVPR, 2017.

[5] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, July 2017.

[6] Z. Yin and J. Shi. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[7] Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in Neural Information Processing Systems 27, 2014

[8] Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016

[9] R. Garg, V. K. BG, G. Carneiro, and I. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In European Conf. Computer Vision, 2016.

[10] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In Computer Vision and Pattern Recognition, 2017.

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

[12] J. Engel, T. Schops, and D. Cremers, "LSD-SLAM: Large-scale direct ¨monocular SLAM," in European Conference on Computer Vision (ECCV). Springer, 2014, pp. 834–849.

[13] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[14] G. Costante, M. Mancini, P. Valigi, and T. A. Ciarfuglia, "Exploring representation learning with CNNs for frame-to-frame ego-motion estimation," IEEE robotics and automation letters, vol. 1, no. 1, pp.18–25, 2016.

[15] S. Wang, R. Clark, H. Wen, and N. Trigoni, "DeepVO: Towards endto-end visual odometry with deep recurrent convolutional neural networks," in Robotics and Automation (ICRA), 2017 IEEE International Conference on. IEEE, 2017, pp. 2043–2050.

[16] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki. Sfm-net: Learning of structure and motion from video. CoRR, abs/1704.07804, 2017.

[17] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. Demon: Depth and motion network for learning monocular stereo. CVPR

[18] Wang C, Buenaposada J M, Zhu R, et al. Learning Depth from Monocular Videos using Direct Methods. 2017.

[19] Goodfellow I, Bengio Y, Courville A. Deep Learning[M]. The MIT Press, 2016.

[20] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4040–4048, 2016.

[21] M. Jaderberg, K. Simonyan, A. Zisserman, et al., "Spatial transformer networks," in Advances in Neural Information Processing Systems, 2015, pp. 2017–2025.

[22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. Transactions on Image Processing, 2004.

[23] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467, 2016.

[24] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in Conference on Computer Vision and Pattern Recognition (CVPR), 2012