Semi-Supervised Deep Learning Framework for Monocular Visual Odometry

Dengsheng Chen
AI Institute, Ecovacs Robotics
Nanjing, China
dense.chen@ecovacs.com

Yuanlong Yu
AI Institute, Ecovacs Robotics
Nanjing, China
jeff.yu@ecovacs.com

Xiang Gao
Ecovacs Robotics
Nanjing, China
shawn.gao@ecovacs.com

***Abstract*—In the last decade, supervised deep learning approaches have extensively employed in visual odometry (VO) applications, which is not feasible in environments where labelled data is not abundant. Therefore, many unsupervised deep learning approaches based on depth map in unknown environments from unlabeled data have been proposed. In this paper, we propose a more simplify framework (Encode and Regress Network, ERNet) to generate a robust 6-DoF pose of a monocular camera with Semi-supervised learning strategy. Compared to other state of the art unsupervised deep VO methods, our framework without any assistant of depth map information can also achieve a good performance in terms of pose accuracy on KITTI dataset.**

I. INTRODUCTION

Visual odometry (VO), as one of the most essential techniques for pose estimation and robot localization, has attracted significant interest in both the computer vision and robotics communities over the past few decades [1].

In the past few decades, model-based VO or geometric based VO has been widely studied on its two paradigms, feature-based method and direct method, which have both achieved great success. However, model-based methods tend to be sensitive to camera parameters and fragile in challenging settings, e.g., featureless places, motion blurs and lighting changes.

In recent years, data-driven VO or deep learning-based VO has drawn significant attention due to its potentials in learning capability, the robustness to camera parameters in challenging environments. Deep learning-based techniques have been adopted with precision to solve a lot of computer vision task, such as image classification [2], semantic segmentation [3] and object tracking [4]. Starting from the relocation problem with the use of supervised learning, Kendall et al. [5] proposed to use a Convolutional Neural Network (CNN) for 6-DoF poses regression with raw RGB-D images as its inputs, which named PoseNet. Video clips were employed in VidLoc to capture the temporal dynamics for relocation. Given pre-processed optical flow [6], a CNN based frame-to-frame VO system was reported in Wang et al. [7] then presented a Recurrent Neural Network (RCNN) based VO method resulting in a competitive performance against model-based VO methods, which named DeepVO. Ummenhofer [8] proposed "DeMoN" which can simultaneous estimate the camera's ego-motion, image depth, surface normal and optical flow. Visual inertial odometry with deep learning was also developed in [9] and [10]. However, all the above-mentioned methods require the ground truth of camera poses or depth maps conducting the supervised training. Currently, obtaining ground truth datasets in practice is typically difficult and expensive, and the amount of existing labeled datasets for supervised training is still limited. These limitations suggest people to look for various unsupervised learning VO schemes, such as [11] and [12]. In order to train the network without pose ground truth, most of them are based on depth map to calculate the loss of the estimated 6-DoF pose of camera, which makes the network quite redundancy.

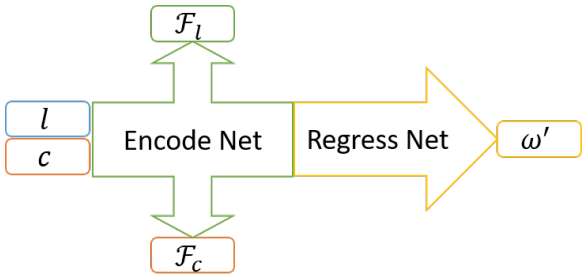


Fig. 1: Architecture overview. The Semi-supervised deep learning approach consists of an encode network and a regress network. Firstly, current frame c and last frame l are feed into encoding network to generate two feature vector $\mathcal{F}_c$ and $\mathcal{F}_l$. Then, $\mathcal{F}_c$ and $\mathcal{F}_l$ are feed into regresser net to generate the final estimated 6-DoF result ω'. The loss of ERNet will be calculated by $\mathcal{F}_c$, $\mathcal{F}_l$, ω' and 6-DoF ground truth ω while supervised training stage, and be calculated only by $\mathcal{F}_c$, $\mathcal{F}_l$ and ω' **without** 6-DoF ground truth ω while unsupervised training stage.

Different from most recently popular unsupervised visual odometry framework, in this paper, we propose a Semi-supervised encode and regress network (ERNet) to efficiently calculate the 6-DoF pose of camera which train on a small part of labeled sequences and a large part of unlabelled sequences **without** any need of depth message. In summary, the main contribution of our method are as follows:

- To the best of our knowledge, this is the first monocular VO method in literature, which uses Semi-supervised deep learning method without any assistant of depth

information.

- Combined supervised learning on a small part of labeled dataset with unsupervised learning on a large part of unlabelled dataset, our semi-supervised deep learning method achieves quite good results compared to the state-of-the-art unsupervised methods.
- Compared to most popular unsupervised deep learning methods which most using depth map to generate the loss of visual odometry, our ERNet can generate the loss of predicted 6-DoF poses without depth map more efficiently while unsupervised training process.

Since ERNet only requires a small part of stereo imagery for supervised training with the need of labeled datasets and a large part of stereo imagery for unsupervised training without the need of labeled datasets, it can train it with an extremely large number of unlabeled datasets to continuously improve its performance.

The rest of this paper is organized as follows. Section II gives an overview of the proposed approach named ER-Net(Encode and Regress Network). Section III describes the proposed Semi-supervised deep learning architecture and the different types of losses used in supervised and unsupervised training process on visual odometer system. Section IV presents experimental results on KITTI dataset. Finally, conclusion is drawn in Section V.

II. ARCHITECTURE OVERVIEW

As shown in Fig. **??**, our Semi-supervised ERNet is composed of an encode network and a regress network. Our ERNet takes two consecutive monocular images as inputs, and produces a 6-DoF pose.

For the encode network, we will train the network to generate two vectors $\mathcal{F}_c$ and $\mathcal{F}_l$, which we supposed they are satisfied the following relationship:

$$\mathcal{F}_c = \mathcal{F}_l \circ \omega$$

where $\circ$ means a specified operation, and ω is ground truth we needed to regress.

On supervised learning stage, we will use the ground truth of ω to train encode and regress net. On unsupervised learning stage, we will use the predicted ω' to calculate the loss for encoding and regresser net. Through $\mathcal{F}_c$ and $\mathcal{F}_l$, we are now able to train ours network on supervised and unsupervised mode. The next section will have a detail description about our framework on visual odometry problem.

III. SEMI-SUPERVISED VISUAL ODOMETRY DEEP LEARNING NETWORK

The network architecture of the proposed method is shown in Fig. **??**. The details of the architecture are explained in the following sections.

A. Encode and Regress Network Architectures

For the encode network, it takes two consecutive monocular images as input and concatenated with two paths which consist of three fully connected linear layers to generate two vectors with length of 3 as $\mathcal{F}_c$ and $\mathcal{F}_l$.

As for visual odometry problem, we designed the following relationship between $\mathcal{F}_c$ and $\mathcal{F}_l$:

$$\mathcal{F}_c = \mathcal{F}_l * \omega_r + \omega_t$$

where $*$ stands for dot production and ω_r, ω_t is rotation and translation value of 6-DoF pose.

For the pose regress network, we are staked two layers LSTM with 1000 hidden units to better handle the history trajectory information. Since rotation (represented by Euler angles) has high nonlinearity, it is usually difficult to train compared with translation, for supervised training, a popular solution is to give a bigger weight to the rotation loss as a way of normalization. In order to better train the rotation with unsupervised learning, we decouple the translation and the rotation with two separates paths of fully-connected layers after LSTM. This enables us to introduce a weight normalizing the rotation and the translation predictions for better performance.

In order to generate more meaningful vectors $\mathcal{F}_c$ and $\mathcal{F}_l$, in the last four convolution layers of encode network, we divide feature with two groups while doing convolution which indicates the current and last frame feature explicitly. And the relationship between $\mathcal{F}_c$ and $\mathcal{F}_l$ are more likely to doing a pose transformation between them. Different with currently popular unsupervised visual odometry framework which need to generate depth map for input scenes, then back-project the depth to three-dimension points and apply the predicted 6-DoF pose to get the loss of estimated pose, it is not necessary for us to estimate any depth information but just two vectors $\mathcal{F}_c$ and $\mathcal{F}_l$.

B. Objective Loss Function

We designed two loss functions for supervised and unsupervised training process.

1) Supervised Loss Function: While supervised learning process, which the ground truth of 6-DoF ω is given, we define the following loss function for ENet:

$$\mathcal{L}_{ENet} = \mathcal{MSE}(\mathcal{F}_c - \mathcal{F}_l * \omega_r - \omega_l)$$

and the following loss function for RNet:

$$\mathcal{L}_{RNet} = \mathcal{MSE}(\omega_r - \omega'_r) + \mathcal{MSE}(\omega_l - \omega'_l)$$

Then, we get the final loss function for our supervised training process:

$$\mathcal{L}_{supervised} = \mathcal{L}_{ENet} + \mathcal{L}_{RNet}$$

2) Unsupervised Loss Function: While unsupervised learning process, which the ground truth of 6-DoF ω is missed, we define the following loss function for ERNet:

$$\mathcal{L}_{unsupervised} = \mathcal{MSE}(\mathcal{F}_c - \mathcal{F}_l * \omega'_r - \omega'_l)$$

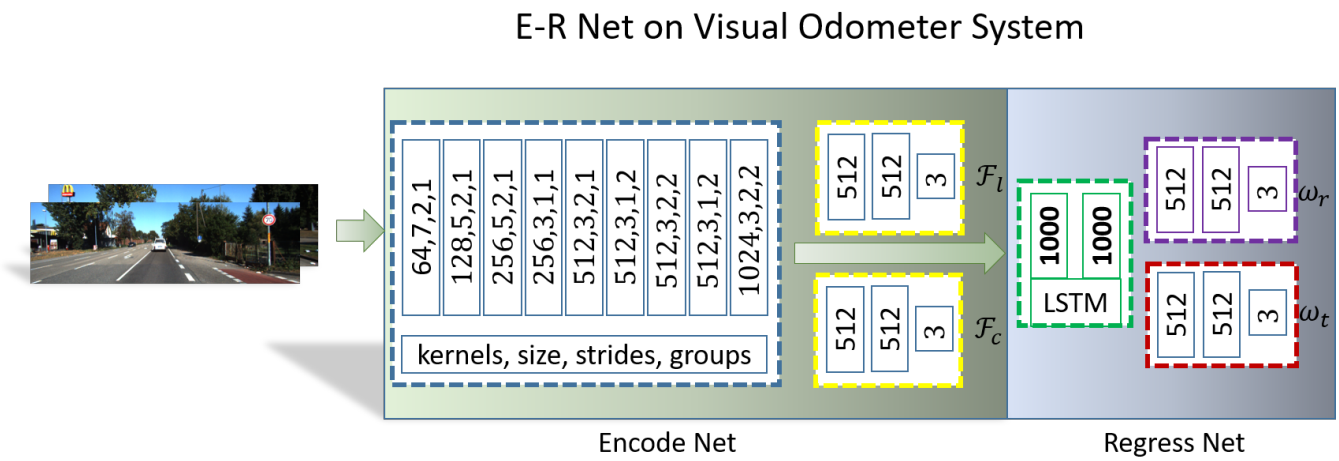


Fig. 2: ERNet on visual odometry system. It takes in two continuous frames every time, and output a 6-DoF pose estimation. In encode net, we first go through nine convolution layers and the number in the box is (kernels, size, strides, groups). Then, we stake three fully linear layers to generate $\mathcal{F}_c$ based on group one and $\mathcal{F}_l$ based on group two. At the same time, we will forward the encoded feature to regress net. In regress net, we stack two LSTM layers with 1000 units to better capture the trajectories information. In order to get a better pose estimation, we split 6-DoF into a 3-DoF translation and 3-DoF Euler rotation angles, then use two path Networks to regress them. Networks in the same color dash line box share the same parameters.

C. Training Strategy

In order to have a better network weight initialization, we first train ERNet with supervised learning, and then train it with unsupervised learning on unlabeled dataset. However, it is hard to make sure that $\mathcal{F}_c$ and $\mathcal{F}_l$ will perfectly satisfied the relationship, so we will slow down the learning rate while unsupervised training stage and then finetune our network with supervised learning to finish our training process.

IV. EXPERIMENTS AND RESULTS

We implemented the architecture with the publicly available PyTorch [13] framework. Batch normalization is employed for all the convolution layers except for the output layers. The weights of the network are optimized with Adam optimization to increase the convergence rate, with the parameters $\beta_1 = 0.9, \beta_2 = 0.999$, , learning rate of $l_{supervised} = 0.001, l_{unsupervised} = 0.0005, l_{finetune} = 0.001$, and mini-batch size of 8. For training purpose, the input tensors of the model are assigned to sequential images of size 184×608. Two consecutive images are stacked together to form the input batch. We use the KITTI [14] dataset for benchmarking. The model is trained on a NVIDIA TITAN XP model GPU. We compare the proposed method with standard training/test splits on the KITTI dataset for supervised odometry estimation task, and use the rest unlabeled sequences as training data on unsupervised training stage.

A. Pose estimation benchmark

We have evaluated the pose estimation performance of our ERNet on the standard KITTI visual odometry split. The dataset contains 11 driving sequences with ground truth odometry obtained through the IMU/GPS sensors and rest 11 driving sequences without ground truth. We use the sequences 00-08 with ground truth for supervised training, 11-21 without ground truth for unsupervised training and 09-10 for testing. The network regresses the pose predictions as 6-DoF relative motion (Euclidean coordinates for translation and rotation) between sequences. We compare the pose estimation accuracy with the existing unsupervised deep learning approaches with the same sequences length of 5, and monocular RGB SLAM. The results are evaluated using Absolute Trajectory Error (ATE) for five consecutive input frames with an optimized scaling factor to resolve scale ambiguity, which is reported to be the best length for the compared methods. As shown in Table **??**, our ERNet achieves good performance with all the competing unsupervised and traditional baselines, without any need of global optimization steps such as loop closure detection, bundle adjustment and re-localization, revealing that ERNet captures long-term high level odometry details in addition to short-term low level odometry features.

Method	**Seq.09**	**Seq.10**
ORB-SLAM[15]	0.014±0.008	0.012±0.011
SfM-Learner[16]	0.016±0.009	0.013±0.009
GeoNet[17]	**0.012±0.007**	0.012±0.009
ERNet	0.018±0.018	**0.011±0.016**

TABLE I: Absolute Trajectory Error (ATE) on KITTI odometry dataset. We also report the results of the other methods for comparison that is taken from [16], [17]. Compared to other supervised methods, our approach can also achieve a good result on KITTI dataset with a more efficient network **without** any assistant of depth map.

B. Supervised Compare to Semi-Supervised ERNet

In order to better demonstrating novel function of our Semi-supervised ERNet, we have visualized the ERNet with unsupervised training and without in Fig. 3. With unsupervised training and finetune, our ERNet shows a better odometry estimation in terms of bath rotational and translational motions compared to ERNet only with supervised training.

V. CONCLUSION

In this study, we proposed a Semi supervised deep learning method for pose estimation without the needed of any depth information for a monocular video sequences, demonstrating the effectiveness of Semi supervised learning in these tasks. The proposed method can achieve a good performance compared to the competing unsupervised and traditional baselines in terms of pose estimation. In a further work, we would like to explicitly address another task which is similar with visual odometer with ERNet framework.

REFERENCES

[1] F. Fraundorfer and D. Scaramuzza, "Visual odometry: Part ii - matching, robustness, and applications," *IEEE Robotics Automation Magazine - IEEE ROBOT AUTOMAT*, vol. 19, pp. 78–90, 06 2012.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 06 2016, pp. 770–778.

[3] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2017.

[4] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," 06 2018, pp. 8971–8980.

[5] A. Kendall, M. K. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2938–2946, 2015.

[6] A. Dosovitskiy, P. Fischery, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *IEEE International Conference on Computer Vision*, 2015.

[7] S. Wang, R. Clark, H. Wen, and A. Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2043–2050, 2017.

[8] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5622–5631, 2017.

[9] W. Xu, D. Choi, and G. Wang, "Direct visual-inertial odometry with semi-dense mapping," *Computers Electrical Engineering*, vol. 67, 02 2018.

[10] V. Usenko, J. Engel, J. Stckler, and D. Cremers, "Direct visual-inertial odometry with stereo cameras," 05 2016, pp. 1885–1892.

[11] R. Li, S. Wang, Z. Long, and D. Gu, "Undeepvo: Monocular visual odometry through unsupervised deep learning," 09 2017.

[12] Y. Almalolu, M. R. U. Saputra, P. Gusmao, A. Markham, and N. Trigoni, "Ganvo: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks," 09 2018.

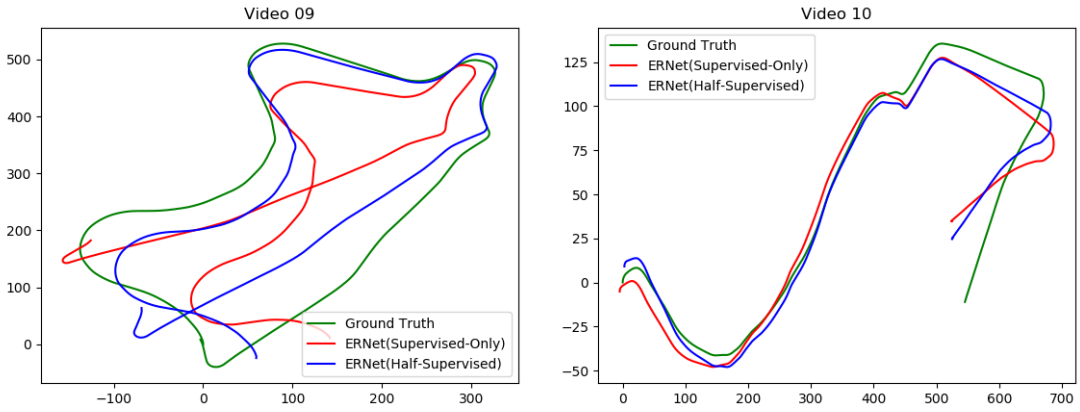
[13] N. Ketkar, *Introduction to PyTorch*, 2017.

[14] A. Geiger, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision Pattern Recognition*, 2012.

[15] R. Mur-Artal, J. M. M. Montiel, and J. D. Tards, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[16] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," 2017.

[17] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," 2018.



(a) (b)

Fig. 3: Visualization of testing result on sequences 09-10. With unsupervised training and finetune, our ERNet shows a better odometry estimation in terms of bath rotational and translational motions compared to ERNet only with supervised training.