Deep Global-Relative Networks for End-to-End 6-DoF Visual Localization and Odometry

Yimin Lin, Zhaoxiang Liu, Jianfeng Huang, Chaopeng Wang, Guoguang Du, Jinqiang Bai*, Shiguo Lian, Bill Huang

AI Department, CloudMinds Technologies Co. Ltd, Beijing, China. *School of Electronic Information Engineering, Beihang University, Beijing, China

{anson.lin, robin.liu, jianfeng.huang.interns, chaopeng.wang, george.du, scott.lian, bill}@cloudminds.com. *baijinqiang@buaa.edu.cn

Abstract

Although a wide variety of deep neural networks for robust Visual Odometry (VO) can be found in the literature, they are still unable to solve the drift problem in long-term robot navigation. Thus, this paper aims to propose novel deep end-toend networks for long-term 6-DoF VO task. It mainly fuses relative and global networks based on Recurrent Convolutional Neural Networks (RC-NNs) to improve the monocular localization accuracy. Indeed, the relative sub-networks are implemented to smooth the VO trajectory, while global sub-networks are designed to avoid drift prob-All the parameters are jointly optimized using Cross Transformation Constraints (CTC), which represents temporal geometric consistency of the consecutive frames, and Mean Square Error (MSE) between the predicted pose and ground truth. The experimental results on both indoor and outdoor datasets show that our method outperforms other state-of-the-art learning-based VO methods in terms of pose accuracy.

1 Introduction

The problem of visual localization has drawn significant attention from many researchers over the past few decades. Solutions for overcoming this problem come from computer vision and robotic communities by means of Structure from Motion (SfM) and visual Simultaneous Localization and Mapping (vSLAM) [Cadena *et al.*, 2016; Özyeşil *et al.*, 2017]. Many variants of these solutions have started to make an impact in a wide range of applications, including autonomous navigation and augmented reality.

During the past few years, most of traditional visual localization techniques have been proposed and grounded on the estimate of the camera motion among a set of consecutive frames with geometric methods. For example, the feature-based method uses the projective geometry relations between 3D feature points of the scene and their projection on the image plane [Mur-Artal *et al.*, 2015; Klein and Murray, 2007], or the direct method minimizes the gradient of the pixel intensities across consecutive images [Engel *et al.*, 2014; Engel *et al.*, 2018]. However, these techniques are critical to

ideal and controlled environments, e.g., with a large amount of texture, unchanged illumination and without dynamic objects. Obviously, their performance drops quickly when facing those challenging and unpredicted scenarios.

Recently, a great breakthrough has been achieved in the Deep Learning (DL), through the application of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), e.g., for the object recognition and scene classification tasks. Therefore, learning-based visual odometry in the past few years has seen an increasing attention of the computer vision and robotic communities [Li et al., 2018]. This is due to its potentials in learning capability and the robustness to camera parameters and challenging environments. However, so far they are still unable to outperform most state-of-the-art feature-based localization methods. The drift from the true trajectory due to accumulation of errors over time is inevitable in those learning based VO system. This is due to the fact that such approaches cannot exploit high-capacity learning 3D structural constraints from limited training datasets. Recent work [Clark et al., 2017] concluded that global place recognition and camera relocalization plays a significant role in reducing these global drifts. As demonstrated in another relevant VLocNet [Valada et al., 2018], the global and Siamese-type relative networks are designed for inferring global poses with the great help of relative motion. Nevertheless, VO drift problem still exists since its global and relative networks are separately optimized and regressed by a multitask alternating optimization strategy.

To solve the drift problem completely, this paper extends VLocNet to fuse both relative and global networks, and considers more temporal sequences with LSTM incorporated in each networks for accurate pose prediction. Furthermore, we also employ a geometric consistency of the adjacent frames for regressing the relative and global networks at the same time. This proposed method brings two advantages: one is obviously that we leverage the camera re-localization to improve the accuracy of 6-DoF VO. On the other hand, relative motion information from odometry can also be used to improve the global pose regression accuracy. In summary, our main contributions are as follows:

(1) We demonstrate the architecture consisting of the CNN-based feature extraction sub-networks (CNN1), the RCNNs-type relative and global pose regression sub-networks (named RCNN1 and RCNN2 respectively),

and finally Fully-connected fusion layers (FCFL) fuse the global and relative poses by connecting these subnetworks to each other.

- (2) The training strategy: we firstly train the feature extraction and relative pose estimation sub-networks from a sequence of raw RGB images, and then the whole architecture is trained in an end-to-end manner to fill the rest of the pose regression sub-networks according to different scenes.
- (3) We design two loss functions to improve the accuracy of our networks. For training the relative sub-networks, the CTC is employed to enforce the temporal geometric consistency between each other within a batch of frames. For training the whole networks, we minimize both CTC and the pose MSE.
- (4) We evaluate our networks using 7-Scenes and KITTI datasets, and the results show it achieves state-of-the-art performance for learning-based monocular camera localization.

2 Related Work

Over the past years, there are numerous approaches that have been proposed for visual localization. In this section, we discuss traditional geometry-based and recent learning-based localization approaches.

2.1 Geometry-based Localization

Geometry-based localization estimates the camera motion among a set of consecutive frames with geometric methods. A variety of geometric methods can be classified into featurebased and direct methods.

Feature-based methods: most feature-based methods work by detecting feature points and matching them between consecutive frames. To improve pose accuracy, they minimize the projective geometry errors between 3D feature points of the scene and their projection on the image plane, e.g., PTAM [Klein and Murray, 2007] is a classical vSLAM system. However, it may suffer from drift since it does not address the principle of loop closing. More recently, the ORB-SLAM algorithm by Mur-Artal et al. [Mur-Artal et al., 2015] is state-of-the-art vSLAM system designed for sparse feature tracking and reached impressive robustness and accuracy. In practice, it also suffers from a number of problems such as the inconsistency in initialization, and the drift caused by pure rotation.

Direct methods: in contrast, direct methods estimate the camera motion by minimizing the photometric error over all pixels across consecutive images. Engel at al. [Engel et al., 2014]developed LSD-SLAM, which is one of the most successful direct approaches. Direct methods do not provide better tolerance towards changing lighting conditions and often require more computational costs than feature-based methods since they work a global minimization using all the pixels in the image.

2.2 Learning-based Localization

Even though Deep Neural Networks (DNNs) are not a novel concept, their popularity has grown in recent years due to a great breakthrough that has been achieved in the computer vision community. Inspired by these achievements, lots of learning-based visual relocalization and odometry systems have been widely proposed to improve the 6-DoF pose estimation.

Visual relocalization: Learning-based relocalization systems are designed to learn from recognition to relocalization with very large scale classification datasets. For example, Kendall et al. proposed PoseNet [Kendall et al., 2015], which was the first successful end-to-end pre-trained deep CNNs approach for 6-DoF pose regression. In addition, Clark et al. [Walch et al., 2017] introduced deep CNNs with Long-Short Term Memory (LSTM) units to avoid overfitting to training data while PoseNet needs to deal with this problem with careful dropout strategies.

Visual odometry: learning-based visual odometry systems are employed to learn the incremental change in position from images. LS-VO [Costante and Ciarfuglia, 2018] is a CNNs architecture proposed to learn the latent space representation of the input Optical Flow field with the motion estimate task. SfM-Net [Vijayanarasimhan et al., 2017] is a self-supervised geometry-aware CNNs for motion estimation in videos that decomposes frame-to-frame pixel motion in terms of scene and object depth, camera motion and 3D object rotations and translations. Recently, most state-of-the-art deep approaches to visual odometry employ not only CNNs, but also sequence-models, such as long-short term memory (LSTM) units [Iyer et al., 2018], to capture long term dependencies in camera motion.

More recently, learning-based global and relative networks are designed for 6-DoF global pose regression and odometry estimation from consecutive monocular images. Clark et al. [Clark et al., 2017] have presented a CNNs+Bi-LSTMs approach for 6-DoF video-clip relocalization that exploits the temporal smoothness of the video stream to improve the localization accuracy of the global pose estimation. Brahmbhatt et al. [Brahmbhatt et al., 2018] proposed a MapNet that enforces geometric constraints between relative poses and absolute poses in network training. Our work is extended to VLocNet [Valada et al., 2018], which incorporated a global and a relative sub-networks. More precisely, even though it has the joint loss function designed for global and relative sub-networks, it is just used to improve the global predictions. Conversely, the global regression results are unable to totally benefit relative networks since its unshared weights are optimized independently without consider the global constraints. Moreover, it considers only a single image as global networks input, which greatly impedes the ability of CNNs to achieve accurate poses. In contrast, we fuse these two streams from both global and relative RCNNS-type sub-networks with joint optimization to benefit the pose prediction.

3 Proposed Model

In this section, we detail our learning-based global and relative fusion framework for jointly estimating global pose and

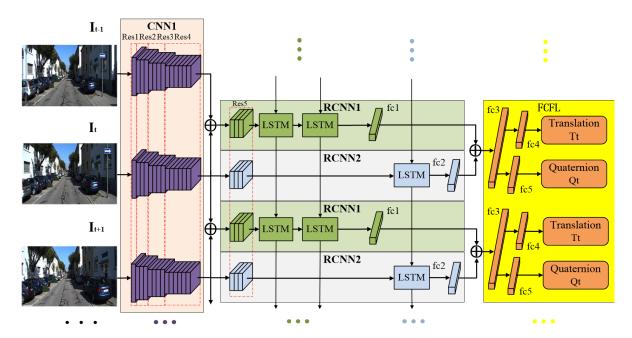


Figure 1: Architecture of the proposed learning-based monocular VO system. CNN1 determines the most discriminative feature as an input for the next two RCNNs; RCNN1 estimates the egomotion of the camera and constrict the motion space while regressing the global localization; RCNN2 is competent to model the 3D structural constraints of the environment while learning from the first two assistant networks; Fully-connected fusion layers (FCFL) fuse relative and global networks to improve the VO accuracy.

odometry from consecutive monocular images. The proposed networks are shown in Fig. 1.

3.1 Network Architecture

CNN-based feature extraction networks (CNN1)

In order to learn effective features that are suitable for the global and relative pose estimation problem automatically, CNN-based feature extraction networks are developed to perform feature extraction on the monocular RGB image. We build upon this networks using the first four residual blocks of the ResNet-50 (named from Res 1 to Res 4) [He $et\ al.$, 2016]. Each residual unit has a bottleneck architecture consisting of 1×1 convolution, 3×3 convolution, 1×1 convolution layers. Each of the convolutions is followed by batch normalization, scale and Exponential Linear Units (ELUs) [Clevert $et\ al.$, 2015].

RCNNs-type relative sub-networks (RCNN1)

Following the feature extraction networks, the deep RCNNs are designed to model dynamics and relations among a sequence of CNNs features. It takes CNNs features from a consecutive monocular RGB images as input, and then the concatenate features from them are fed into the last residual blocks of the ResNet-50 (Res 5). Note that the output dimension of this layer is W×H×1024. As described in DeepVO [Wang *et al.*, 2018], two Long Short-Term Memory (LSTMs) [Zaremba and Sutskever, 2014] are employed as RNNs to find and exploit correlations among images taken in long trajectories and each of the LSTM layers has 1000 hidden states. The RCNNs output pose estimation at each time step with a fully-connected layer fc1 whose dimension is 1024.

RCNNs-type global sub-networks (RCNN2)

We also feed the previous CNNs features to the last residual blocks of the ResNet-50 (Res 5) and reshape LSTM's output to a fully-connected layer fc2, whose dimension is 1024. It corresponds in shape to the output of the relative RCNNs unit before the fusion stage. Note that the cell of LSTM stores the past few global poses and therefore it is able to improve the predicted pose accuracy of current image.

Fully-connected fusion layers (FCFL)

Finally, the following fusion stage concatenates features from the two relative and global sub-networks, and reshapes its output to 1024, namely fc3. We also add two inner-product layers for regressing the translation T_k and quaternion Q_k , namely fc4 and fc5. Obviously, the dimensions of fc4 and fc5 layers are 3 and 4, respectively.

3.2 Temporal Geometric Consistency Loss

Here, we introduce CTC that are based on the fundamental concepts of composition of rigid-body transformations. Fig. 2 shows a sequential set of frames $F=(I_0,I_1,I_2,I_3,I_4)$, where we note that temporal length K=5. Note that $P_i=(Q_i,T_i)$ is a 6-DoF predicted pose, where T_i and Q_i denote the translation and quaternion of frame i, respectively. We train the networks to predict the transforms between each other: $[P_{01},P_{12},P_{23},P_{34},P_{02},P_{24},P_{04}]$. As an example, the predicted transform P_{01} from P_{01} to P_{01} should be equal to the product of the two P_{01} and P_{01} transforms, where P_{01} indicates the ground truth of frame P_{01} thus:

$$P_{01} = \widehat{P}_1 \widehat{P}_0^{-1} = \widehat{P}_{01} \tag{1}$$

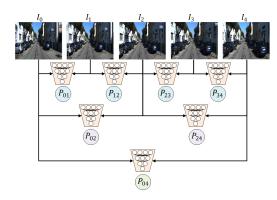


Figure 2: Architecture of CTC. It represents temporal geometric consistency of the consecutive frames.

Note that using Eq.(1) in practice, there exist errors in the predicted and ground truth, so we have CTC functions:

$$L_{0} = \|P_{01} - \widehat{P}_{01}\|_{2}^{2}, L_{1} = \|P_{12} - \widehat{P}_{12}\|_{2}^{2}$$

$$L_{2} = \|P_{23} - \widehat{P}_{23}\|_{2}^{2}, L_{3} = \|P_{34} - \widehat{P}_{34}\|_{2}^{2}$$

$$L_{4} = \|P_{02} - \widehat{P}_{02}\|_{2}^{2}, L_{5} = \|P_{24} - \widehat{P}_{24}\|_{2}^{2}$$

$$L_{6} = \|P_{04} - \widehat{P}_{04}\|_{2}^{2}$$
(2)

where $\left\|\cdot\right\|_2^2$ is MSE. So the relative loss function which consists of Eq.(2) are shown as:

$$\theta = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{6} (L_k^i)$$
 (3)

where θ is the relative or global RCNNs parameters and N is the number of samples. We use this optimization Eq.(3) to train our RCNNs sub-networks. Note that, these constrains can be equal to a Local Bundle Adjustment in traditional vS-LAM system [Mur-Artal *et al.*, 2015], also known as windowed optimization. It is an efficient way to maintain a good quality pose over a local number of frames. So the CTC here are better strategies to learn about spatial relations of the environment. To train our 6-DoF end-to-end pose regression system, we can jointly use the global and relative loss function as follows:

$$w = \underset{w}{\operatorname{arg\,min}} \frac{1}{N} \sum_{i=1}^{N} \left\{ \sum_{k=0}^{6} (L_{k}^{i}) + \sum_{j=0}^{4} \left\| P^{i}_{j} - \widehat{P}^{i}_{j} \right\|_{2}^{2} \right\}$$
(4)

where ω is the networks parameters. It is obvious that Eq.(4) tries to minimize the Euclidean distance between the ground truth pose and estimated one while enforcing the geometric consistency between each other within a batch of frames.

3.3 Training Strategy

We firstly initialize the CNN1 and RCNN1 from a sequence of raw RGB images using the optimization Eq.(3). In partic-

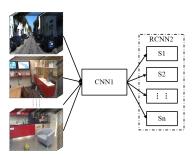


Figure 3: Illustration of training for the global networks from different scenes. CNN1 determines the most discriminative feature and RCNN2 learns from different scenes for saving their landmark Si.

ular, RCNN1 directly replace fc3 with fc1 and estimate the pose from fc4 and fc5 for the time being.

For initializing RCNN2, we observe that the most global pose regression [Kendall *et al.*, 2015] can only be determined in a known training environment. So it is time consuming to retrain the whole networks according to different scenes. As shown in Fig. 3, in order to retrain our deep model faster, different scenes are fed into the common CNN1 to produce an effective feature in the monocular image, which is then passed through individual RCNN2 to learn for saving their landmark Si. Thereby, we only need to retrain the RCNN2 for different scenes and each image still yields an accurate pose estimate at each Si through the networks. Note that, RCNN2 also directly replaces fc3 with fc2 and regress the pose using the optimization Eq.(4).

Up to now, we achieve pretrained weights for CNN1, RCNN1 and RCNN2. Finally, the whole architecture is trained and refined in an end-to-end manner via the optimization Eq.(4).

4 Experimental Evaluation

In this section, we evaluate our proposed networks in comparison to the state-of-the-art algorithms on both indoor and outdoor datasets, followed by detailed analysis on the architectural decisions and finally, we demonstrate the best temporal length.

4.1 Evaluation Datasets

We evaluate our networks on two well-known datasets: Microsoft 7-Scenes [Shotton *et al.*, 2013] and KITTI Visual Odometry benchmark [Geiger *et al.*, 2012]. We follow the original train and test splits provided by other literatures to facilitate comparison and benchmarking.

Microsoft 7-Scenes

It is a dataset that collect RGB-D images from seven different scenes in an indoor office environment. All scenes were recorded from a handheld Kinect RGB-D camera at 640×480 resolution. The dataset provides the ground truth poses extracted using KinectFusion. Each sequence was recorded with motion blur, perceptual aliasing and textureless features in the room, thereby making it a challenging dataset for relocalization and tracking.

Table 1: MEDIAN ERRORS ON MICROSOFT 7-SCENES

Scene -	PoseNet		DeepVO		VLocNet		Ours	
	t_{rel}				t_{rel}			r_{rel}
Chess	0.32	8.12	0.06	2.61	0.036	1.70	0.016	1.72
Fire	0.47	14.4	0.10	4.33	0.039	5.33	0.011	2.19
Heads	0.29	12.0	0.35	7.11	0.046	6.64	0.017	3.56
Office	0.48	7.68	0.10	3.11	0.039	1.95	0.024	1.95
Pumpkin	0.47	8.42	0.11	3.30	0.037	2.28	0.022	2.27
RedKitchen	0.59	8.64	0.10	2.58	0.039	2.20	0.018	1.86
Stairs	0.47	13.8	0.45	9.18	0.097	6.47	0.017	4.79
Average	0.44	10.4	0.18	4.60	0.048	3.80	0.018	2.62

KITTI Visual Odometry benchmark

It consists of 22 stereo sequences and they provide 11 sequences (00-10) with ground truth trajectories for training and 11 sequences (11-21) without ground truth for evaluation. This high-quality dataset was recorded with long sequences of varying speed, including a set of 41000 frames captured at 10 fps and a total driving distance of 39.2 km with frequent loop closures which are of interest in SLAM. So it is very popular for the monocular Visual Odometry algorithms.

4.2 Network Training

The network models were implemented with the TensorFlow framework and trained with NVIDIA GTX 1080 GPUs and Intel Core i7 2.7GHz CPU. Adam optimizer was employed to train the networks for up to 2000 epochs with parameter β_1 = 0.9 and β_2 = 0.999. The learning rate started from 0.001 and decreased by half for every 1/5 of total iterations. The temporal length K fed to the relative and global pose estimator is 5. The size of image used by the networks is 224×224 pixel. Thus, our per-frame runtime for each pose inference is between 45 ms and 65 ms.

4.3 Microsoft 7-Scenes Datasets

In this experiment, we compare the performance of our networks with other state-of-the-art deep learning-based relocalization and tracking methods, namely PoseNet [Kendall $et\ al.$, 2015], DeepVO [Mohanty $et\ al.$, 2016] and VLocNet [Valada $et\ al.$, 2018]. In order to implement fair qualitative and quantitative comparison, we use the same 7-Scenes datasets for training and testing as described in [Valada $et\ al.$, 2018]. For each scene, we show their median translational t_{rel} : (m) and rotational t_{rel} : (\circ) errors in Table 1, respectively.

It shows that our networks outperform previous CNN-based PoseNet by 95.9% in positional error and 74.8% in orientation error. Taking Pumpkin as an example, we achieve a positional error reduction from 0.47m for PoseNet to 0.022m for our method. The reason is that PoseNet always results in noisy predictions on single image. In contrast, the RCNNs in our networks constrict the motion space while using sequential images to improve global relocalization accuracy. Therefore, this experiment results validate that our networks have the effectiveness of using geometric constraints from consecutive images for improving relocalization accuracy. Furthermore, it can be seen that the proposed networks significantly

Table 2: RESULTS ON KITTI SEQUENCES

Seq.	Dee	pVO	L-V	/O3	Ours	
	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
03	6.72	6.46	3.18	1.31	1.93	1.95
04	6.33	6.08	2.04	0.81	0.10	0.25
05	3.35	4.93	2.59	0.99	2.51	0.91
06	7.24	7.29	1.38	0.95	0.30	0.99
07	3.52	5.02	2.81	2.54	1.53	2.11
10	9.77	10.2	4.38	3.12	3.63	3.00
Average	6.15	6.66	2.73	1.62	1.67	1.54

outperform the DeepVO approach in all of the test scenes, resulting in a 90% and 43% boost in position and orientation accuracy, respectively. The DeepVO network tries to regress the VO but probably suffers from high drifts. The reason is that the orientation changes in the training data are usually small and orientation is more prone to overfitting. However, our system reduces the drift over time due to the global pose regression strategy as done in the traditional visual SLAM system. In addition, our networks also perform better than VLoc-Net, and the orientation and positional errors are reduced by more than 31% and 62%, respectively. The main reason we find from VLocNet is that their global pose regression and visual odometry networks are predicted independently. But in our framework, we do fuse the results from global regression and relative pose estimation. In summary, these experimental results validate that our strategy is able to filter out the noises by fusing a series of measurements observed from global and relative networks over time.

4.4 KITTI Datasets

Next, we additionally deploy experiments in an outdoor environment for analyzing the large-scale VO performance. KITTI is much larger than typical indoor datasets like 7-Scenes, where sequence 00, 02, 08 and 09 are used for training the RCNNs-type relative sub-networks. As described in [Wang et al., 2018], the trajectories are segmented to different lengths to generate almost 7410 samples in total for training. The trained models are tested on the sequence 03, 04, 05, 06, 07 and 10. As shown in Table 2, the performance of the our networks is analyzed according to the KITTI VO/SLAM evaluation metrics, where t_{rel} : (%) and r_{rel} : (°/100m) are averaged Root Mean Square Errors (RMSE) of the translational and rotational drifts for all subsequences of lengths ranging from 100 to 800 meters with different speeds.

Table 2 shows quantitative comparison against two state-of-the-art VO approaches including L-VO3 [Zhao et al., 2018] and RCNNs-type DeepVO [Wang et al., 2018]. The proposed method significantly outperforms the DeepVO approach in all of the test sequences, resulting in a 71% and 76% boost in translation and rotation accuracy, respectively. As shown in Fig. 4, DeepVO suffers from high drifts as the length of the trajectory increases and the errors of the rotation significantly increase because of significant changes on rotation during car driving. Unlike that, our networks produce relatively accurate and consistent trajectories against to

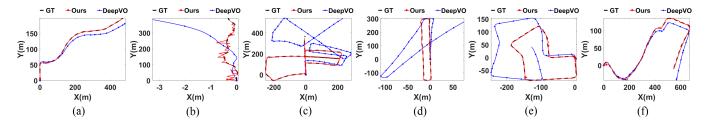


Figure 4: Trajectories of results on the testing Sequence (a) 03, (b) 04, (c) 05, (d) 06, (e) 07 and (f) 10 of the KITTI VO benchmark.

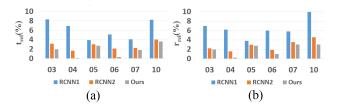


Figure 5: Comparative analyses of average (a) translation and (b) rotation RMSE using various architectures on KITTI sequences.

the ground truth. These owe to the global and relative architecture with the proposed CTC loss. In addition, it is able to overcome the performance of state-of-the-art learning-based L-VO3. Although some errors are slightly worse than that of the L-VO3, this may be due to the fact that our networks are trained without enough data to cover the velocity and orientation variation. Finally, we can see that the absolute scale to each sequence is completely maintained during the end-to-end training.

4.5 Ablation Studies

In this section, we present additional ablation studies on performances with respect to considering various architectural components and temporal length K.

In order to validate the effectiveness of our joint architecture, we compare our networks against relative-only (RCNN1) and global-only (RCNN2) architectures. The quantitative rules can be found in Section 4.4. In particular, RCNN1 directly replaces fc3 with fc1 and estimates the pose from consecutive images. While RCNN2 also directly replaces fc3 with fc2 and regresses the pose. They are trained using the loss function Eq.(3) and Eq.(4) respectively, and temporal length K equals to 5 as well. It is observed that compared with the RCNN1, the pose generated from RCNN2 is more accurate. A possible explanation is that the global networks reduce the serious drift since it has the ability to relocalization with previous observation for the long-term prediction. While the relative networks only focus on motion from 2D or 3D optical flow, which is hard to efficiently model 3D structural constraints with limited training samples in complex environments. Compared to the RCNN1 and RCNN2 as shown in Fig. 5, our approach predicts more precise pose. This is more evident if we fuse streams from both global and relative sub-networks to benefit the long-term pose prediction.

In addition, we provide a performance comparison with respect to various temporal lengths. Fig. 6 shows our net-

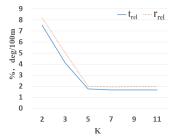


Figure 6: Comparation of average translation t_{rel} and rotation r_{rel} RMSE with respect to various temporal length K.

works with different K values changed from 2 to 11 and their corresponding average translation and rotation RMSE of the six KITTI sequences. We observe that the localization errors descend as the length of the sequential frames increases. However, the accuracy seems to be stable when K is larger than 5. This phenomenon is due to the fact that the covisible constraint between 1-th frame and K-th frame become weak when K is large enough. Furthermore, we find that training such networks, especially when K is larger than 5, requires more training data to generalize well in unseen data and avoid overfitting. Therefore, we conclude that the temporal length k=5 is the best configuration for VO task.

5 Conclusion

In this paper, we addressed the challenge of learning-based visual localization of a camera or an autonomous system with the novel networks. It mainly consists of CNN-based feature extraction sub-networks that determine the most discriminative feature as an input for the next two RCNNs, RCNNs-type relative sub-networks that estimate the egomotion of the camera and constrict the motion space while regressing the global localization, and RCNNs-type global sub-networks that are competent to model the 3D structural constraints of the environment while learning from the first two assistant networks. Finally, it fuses and jointly optimizes the relative and global networks to improve VO accuracy. Furthermore, we employ the CTC loss function for training the relative and global RC-NNs. The indoor and outdoor experimental evaluations indicate that our networks can produce accurate localization and be adopted to maintain a large feature map for drift correction under long range pose estimation. In the next step, we plan to extend the ability of global networks to work under any unknown environment and promote the robustness of place recognition in cases where illumination and appearance change dramatically.

References

- [Brahmbhatt *et al.*, 2018] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2616–2625, 2018.
- [Cadena *et al.*, 2016] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016.
- [Clark et al., 2017] Ronald Clark, Sen Wang, Andrew Markham, Niki Trigoni, and Hongkai Wen. Vidloc: A deep spatio-temporal model for 6-dof video-clip relocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, pages 2652–2660, 2017.
- [Clevert *et al.*, 2015] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv* preprint arXiv:1511.07289, 2015.
- [Costante and Ciarfuglia, 2018] Gabriele Costante and Thomas Alessandro Ciarfuglia. Ls-vo: Learning dense optical subspace for robust visual odometry estimation. *IEEE Robotics and Automation Letters*, 3(3):1735–1742, 2018.
- [Engel et al., 2014] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In European Conference on Computer Vision (ECCV), pages 834–849, 2014.
- [Engel et al., 2018] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence* (*TPAMI*), 40(3):611–625, 2018.
- [Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012.
- [He et al., 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pages 770–778, 2016.
- [Iyer *et al.*, 2018] Ganesh Iyer, J Krishna Murthy, K Gunshi Gupta, and Liam Paull. Geometric consistency for self-supervised end-to-end visual odometry. *arXiv preprint arXiv:1804.03789*, 2018.
- [Kendall *et al.*, 2015] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2938–2946, 2015.

- [Klein and Murray, 2007] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In 6th IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR), pages 225–234, 2007.
- [Li *et al.*, 2018] Ruihao Li, Sen Wang, and Dongbing Gu. Ongoing evolution of visual slam from geometry to deep learning: Challenges and opportunities. *Cognitive Computation*, 10(6):875–889, 2018.
- [Mohanty et al., 2016] Vikram Mohanty, Shubh Agrawal, Shaswat Datta, Arna Ghosh, Vishnu Dutt Sharma, and Debashish Chakravarty. Deepvo: A deep learning approach for monocular visual odometry. arXiv preprint arXiv:1611.06069, 2016.
- [Mur-Artal *et al.*, 2015] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.
- [Özyeşil *et al.*, 2017] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion. *Acta Numerica*, 26:305–364, 2017.
- [Shotton et al., 2013] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), pages 2930–2937, 2013.
- [Valada *et al.*, 2018] Abhinav Valada, Noha Radwan, and Wolfram Burgard. Deep auxiliary learning for visual localization and odometry. *arXiv preprint arXiv:1803.03642*, 2018.
- [Vijayanarasimhan *et al.*, 2017] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017.
- [Walch et al., 2017] Florian Walch, Caner Hazirbas, Laura Leal-Taixe, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using lstms for structured feature correlation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 627–637, 2017.
- [Wang et al., 2018] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks. *The International Journal of Robotics Research*, 37(4-5):513–542, 2018.
- [Zaremba and Sutskever, 2014] Wojciech Zaremba and Ilya Sutskever. Learning to execute. *arXiv preprint arXiv:1410.4615*, 2014.
- [Zhao *et al.*, 2018] Cheng Zhao, Li Sun, Pulak Purkait, Tom Duckett, and Rustam Stolkin. Learning monocular visual odometry with dense 3d mapping from dense 3d flow. *arXiv preprint arXiv:1803.02286*, 2018.