Robust Pose Estimation for Large Displacement Trajectories Through Dual-Task Learning

Jeong-Wook Lee[1], Su-Ji Jang[2], and Ue-Hwan Kim[2]

[1] Interdisciplinary Program in Artificial Intelligence, Seoul National University
[2] AI Graduate School, GIST

Abstract. Visual Odometry (VO) is a crucial task for estimating the current pose of intelligent agents—particularly in applications such as autonomous driving and Simultaneous Localization and Mapping (SLAM). However, accurately estimating the pose becomes challenging when significant displacement occurs between consecutive image frames. Also, camera pose estimation and loop closure detection which are the main modules of the SLAM task have utilized separate networks in previous works—increasing the number of computation modules for resource-constrained environments. To overcome these limitations, we propose an architecture that robustly estimates relative pose between consecutive image frames with large displacement. By formulating dual-task learning of VO and VPR, the proposed architecture leverages both local and global contexts to handle large displacement—reducing the number of computation modules as well. The proposed network demonstrates notable performance enhancement for substantial displacement trajectories in TartanAir and DynaKITTI benchmark datasets—showcasing its effectiveness and potential for applications in real-world scenarios, such as autonomous navigation and mapping.

Keywords: visual odometry, visual place recognition, robot vision

1 Introduction

Simultaneous Localization and Mapping (SLAM) is a widely used technique that enables an intelligent agent to keep track of its location and map the surrounding environment at the same time. Typical SLAM systems include pose estimation and loop closure detection modules. The pose estimation module estimates the current location of the moving sensor while the loop closure detection module identifies if the current location is a previously visited area, to adjust the accumulated errors. Among various methods to predict the pose, it is resolved through the visual odometry (VO) task [32]. VO tasks fall into geometric based methods [15,8] and deep learning based methods [35,30]. With the development of high-performance computing devices and large-scale datasets, research on deep learning-based VO tasks has become prevailing.

However, previous deep learning-based methods show low performance on datasets with large camera motions. Particularly, VO networks fail to capture global information which leads to low performance in large displacement trajectories; conventional visual odometry and SLAM methods [24,28,22] exhibit from 30% to 150% performance gaps between small and large displacement trajectories. We assume that large displacement

trajectories require the analysis of a wide range of views, i.e., global contexts, and contemporary deep learning-based VO methods do not effectively incorporate global contexts for handling large displacements.

While VO modules handle camera pose estimation, the visual place recognition (VPR) task performs loop closure detection. Through the VPR task, agents can determine whether the current location is a previously visited place or not and resolve the accumulated estimation error. Conventional SLAM systems have employed separate VO and VPR networks to conduct camera pose estimation and loop closure detection [13]; this increased number of networks could turn into a significant computational burden as intelligent agents typically operate on mobile robots or edge environments.

To overcome the limitations, we propose a dual-task learning network capable of both the VO and VPR tasks; the proposed network robustly operates on large displacement trajectories and reduces the number of computation modules. In the proposed setup, the network can effectively utilize both global and local features since the VO and VPR modules extract the corresponding features from images. Thus, the proposed approach improves the performance of both VO and VPR tasks in large displacement environments by benefiting from rich features.

2 Related Works

2.1 Deep Visual Odometry (DVO)

DVO methods can be broadly categorized into two types: supervised [28,20], and self-supervised [18,14] learning. Supervised learning-based VO relies on ground truth poses for each image, thus displaying promising performance. However, supervised learning-based approaches demand significant time and cost for annotating ground-truth labels. Self-supervised learning-based VO methods have emerged to address the aforementioned issue, yet they often fall short of the performance achieved by supervised learning. To enhance the effectiveness of self-supervised learning-based VO methods, a set of studies proposed multi-tasking networks [13] that combine pose estimation with depth estimation or semantic segmentation tasks. However, both supervised and self-supervised methods frequently exhibit low performance and struggle to estimate camera pose on unseen trajectories, indicating poor generalization. In tackling this challenge, TartanVO [28] leverages intrinsic parameter information, resulting in robust generalization performance.

While previous VO approaches take advantage of multi-task learning with depth estimation and semantic segmentation tasks for performance [13], or intrinsic parameter information for generalization [28], they lack balanced global and local features for handling large displacements.

2.2 Optical Flow Estimation

The optical flow estimation task aims to estimate the per-pixel correspondence between a source and target image [11]. Optical flow estimation is an important component in deep learning-based VO methods. In TartanVO [28], optical flow map output gets concatenated with intrinsics and becomes the input of the pose network. It is crucial to get

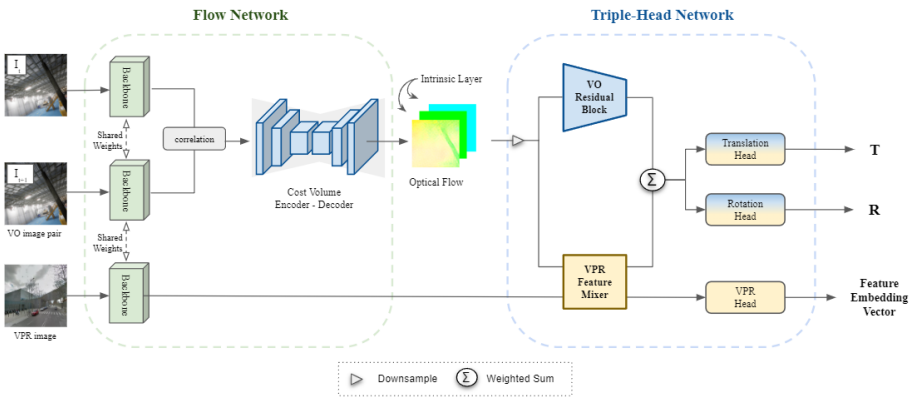

Fig. 1: Overall architecture of the proposed dual-task learning network.

an accurate optical flow map to obtain an accurate relative camera pose of consecutive input images. Traditionally, optical flow estimation networks have employed convolutional neural networks. FlowNet [7] has first proposed a deep learning-based optical flow method utilizing convolutional neural network layers. Then, PWC-Net [21] has enhanced FlowNet by introducing coarse-to-fine and iterative optical flow estimation methodology. RAFT [23] has introduced a coarse-and-fine approach to effectively detect fast-moving objects. Recently, transformer and attention-based optical flow estimation methods [11] have appeared to model long-range relations and showed state-of-the-art performance.

2.3 Visual Place Recognition

The VPR task identifies previously visited locations in the physical world based on visual cues [5]. The VPR task handles a pair of images having large displacement and time gap—making learning global context crucial. In recent years, research on visual place recognition tasks has shifted towards deep learning-based methods, due to their superior performance and ability to handle challenging environments [2]. Most VPR tasks are formulated as image retrieval, representing images as descriptors using a backbone network or a pooling method. Recently, a group of image retrieval methods using learnable pooling [16,34] have shown promising performance. In image retrieval problems, the contrastive loss is commonly used for training. Among them, the triplet loss [10] is mainly used to learn embeddings. Moreover, MixVPR [2] has used the multi-similarity loss [31] to learn visual feature embeddings. MixVPR has outperformed other visual place recognition methods in various benchmarks and real-world datasets.

In this work, we propose to learn VPR along with VO to enhance the performance of VO in large displacement environments by employing global features from VPR.

3 Methodology

Figure 1 illustrates the overall architecture of the proposed network. The proposed network consists of two consecutive modules: the flow network and the triple-head network [28,26]. The flow network receives two consecutive VO image frames $I_t, I_{t+1} \in \mathbb{R}^{H \times W \times C}$ and estimates the optical flow map $F \in \mathbb{R}^{H \times W \times 2}$ as the output (Sec 3.1). The triple-head network processes the estimated flow map and the intrinsic layer $I \in \mathbb{R}^{H \times W \times 4}$ and evaluates the relative translation and rotation as the final output (Sec 3.2). Moreover, the VPR input image only goes through the backbone of the flow network, and the VPR Feature Mixer in the triple-head network generates the final feature embedding vector for the VPR task.

3.1 Flow Network

Architecture Figure 2 illustrates the overall architecture of the optical flow network. We design the optical flow network utilizing a transformer block since the transformer block effectively captures global information—which is crucial for enhancing the performance of both pose estimation in large displacement scenarios and the VPR task. Concretely, we employ the local self-attention and the global cross-attention [6] in the flow network [11] for both the cost volume encoder and decoder. On one hand, the local self-attention process divides feature maps into sub-windows and calculates self-attention only within the sub-window. On the other hand, the global cross-attention sequentially extracts the global information. During the global cross-attention process, the local self-attention output acts as the key and value vectors, and the randomly initialized codeword acts as the query vector. They get updated throughout the training process. The application of local self-attention and global cross-attention enables the learning of both local and global information from the cost maps which enhances the performance of estimating the camera pose for both small and large displacement trajectories.

Computational complexity analysis When the size of key and value is $H_1 \times W_1$, and query $H_2 \times W_2$, each having d channels, the original computational cost of cross-attention becomes $O(H_1 H_2 W_1 W_2 d)$. However, the application of the local self-attention and global cross-attention allows for a reduction in computational complexity [6]. Local self-attention operates on $H_1 \times W_1$-sized query, key, and value. The input features are divided into sub-windows of size $m \times n$, with each sub-window containing $\frac{H_1}{m} \times \frac{W_1}{n}$ pixels. Local self-attention is applied within each window, and the computational cost for each window is $O((\frac{H_1 W_1}{mn})^2 d)$. With $m \times n$ sub-windows, the overall computational cost of local self-attention becomes [6]: $O(\frac{H_1^2 W_1^2}{mn} d)$. Next, global cross-attention is performed on $H_1 \times W_1$-sized key and value and $H_2 \times W_2$-sized query. Instead of computing attention between every pair of pixels in key and value, a single value represents each window of key and value. Consequently, $m \times n$ values are used for key and value. The final computational cost of global cross-attention is as follows [6]: $O(mnH_2W_2d)$. Finally, the total computational cost of local self-attention and global cross-attention is then expressed as $O(\frac{H_1^2 W_1^2}{mn} d + mnH_2W_2d)$. Applying the arithmetic-geometric mean to the total computational cost yields $O(\frac{H_1^2 W_1^2}{mn} d + mnH_2W_2d) \geq$

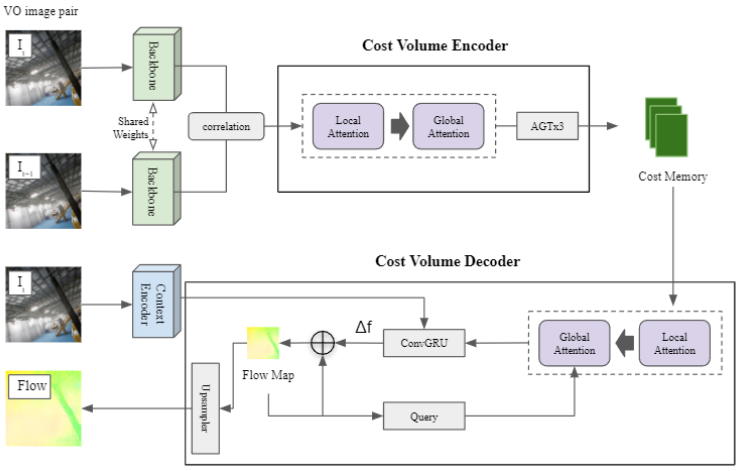

Fig. 2: Overall architecture of the proposed FlowNet. The proposed FlowNet effectively captures global information—crucial for handling large displacements.

$O(2H_1W_1\sqrt{H_2W_2}d)$. Therefore, the computational complexity for local self-attention and global cross-attention becomes [6]: $O(H_1W_1\sqrt{H_2W_2}d)$.

3.2 Triple-Head Network

We design the triple-head network architecture to accommodate the execution of the VPR task as well as the pose estimation task. We embrace a combination of VO residual blocks and the VPR feature-mixer block [2]. The VO residual blocks consist of stacked residual layers and the VPR feature mixer block of four feature mixer layers; each feature mixer layer contains sequential normalization-MLP-activation-MLP layers. This sequential application of feature-mixer layers fosters the incorporation of spatial relationships within each feature map and accumulates features in an integrated manner [2]—culminating in a global descriptor as the final output. Furthermore, opting for the MixVPR feature mixer as opposed to a transformer-based architecture contributes to the reduction in the number of parameters and computational complexity. Finally, the outputs of the VO residual blocks and the VPR feature mixer get weighted summed with an equal ratio, and the weighted summed vector goes into the translation and rotation head to get the final relative translation and rotation output.

3.3 Loss Functions

To make the training procedure capable of dual-task learning between the VO and VPR tasks, we employ various losses suitable for each task. The total loss function consists of three distinctive loss functions as follows:

$$L_{total} = L_{flow} + \lambda_1 L_{pose} + \lambda_2 L_{vpr}, \tag{1}$$

where L_{flow}, L_{pose} and L_{vpr} are the flow loss [23], the pose estimation loss [28] and the VPR loss [1], respectively and λ_1 and λ_2 are weighting coefficients between three different losses.

3.4 Training Scheme

Given the data-intensive nature of the transformer-based flow network, the proposed approach begins with pre-training the flow network utilizing the ground truth optical flow map of the TartanAir training dataset [29] for 25 epochs [28]. After the pre-training stage, we jointly learn the flow network and the triple-head network using the TartanAir VO training dataset and the GSVCities [1] VPR training dataset. Throughout this second training phase, each batch follows a sequential process wherein TartanAir source and target images traverse the entire flow network and the triple-head network to evaluate the final relative translation and rotation values. Following this, GSVCities training images pass through the backbone of the flow network and the VPR feature mixer of the triple-head network and become the final feature embedding vectors. Finally, the supervision signal derived from each batch backpropagates through the whole network.

4 Experiments

4.1 Datasets

Training. We train the overall network using the TartanAir VO dataset [29] and the GSVCities VPR dataset [1]. The TartanAir dataset contains images collected through the AirSim simulator and aids research in various robot navigation tasks. The TartanAir dataset includes thirty different environments while only eighteen environments are open to the public. The GSVCities dataset is a challenging image dataset depicting a total of 67,000 places which entails severe conditions such as illuminiation, seasonal changes, and occlusion.

Evaluation. For the evaluation, we use two benchmark datasets for both the VO and VPR tasks. For VO evaluation, we utilize the MH sequences of the official TartanAir test dataset [29] and the dynamic sequences of the KITTY Odometry dataset [9,19]. For both the TartanAir and DynaKITTI test datasets, we divide the sequences into large displacement trajectories and small displacement trajectories. Out of eight trajectories of the TartanAir test dataset, we classify MH002, MH005, and MH007 trajectories as large displacement trajectories (max translation of 0.50m) while the other five trajectories as small displacement trajectories (max translation of 0.29m). For the DynaKITTI dataset, we classify the 01 trajectory as a large displacement trajectory (max translation of 2.67m), and the 00, 03, 08, and 10 trajectories as small displacement trajectories (max translation of 0.78m). For VPR evaluation, we evaluate models on the Pitts250k-test dataset [25] and the Pitts30k-test dataset [25]. The Pitts250k-test dataset consists of 8,000 queries and 83,000 reference images; the Pitts30k-test dataset contains 8,000 queries and 8,000 reference images which is the subset of the Pitts250k-test dataset.

Data Imbalance Problem The eighteen environments of TartanAir [29] used for training amount to approximately 300,000 image pairs and the GSVCities dataset [1] to

Dataset	TartanAir-test										
Trajectory	Large Displacement (1) MH002	MH005	MH007	(1) Average	Small Displacement (2) MH000	MH001	MH003	MH004	MH006	(2) Average	Total Average
ORB SLAM †	2.37	-	2.73	-	1.30	**0.04**	2.45	-	21.47	-	-
DeepV2D	4.54	11.55	3.76	6.62	6.15	2.12	3.89	2.71	5.53	4.08	5.86
TartanVO	2.00	3.19	2.04	2.41	4.88	0.26	**0.94**	**1.07**	**1.00**	1.63	1.92
Ours	**1.85**	**1.90**	**1.91**	**1.88**	**3.55**	0.28	1.05	1.36	1.86	**1.62**	**1.72**

Table 1: Quantitative result of the visual odometry task on the official TartanAir test dataset. † indicates a SLAM method which has additional optimization processes. The proposed method achieves state-of-the-art performance in both large and small displacement trajectories.

Dataset	DyanKITTI						
Trajectory	Large Displacement (1) 01	Small Displacement (2) 00	03	08	10	(2) Average	Total Average
DeepVO †	1.2896	(0.0206)	0.0783	(0.6547)	0.1042	0.2144	0.4295
TrianFlow †	(8.2127)	0.6966	1.6862	(3.8984)	0.2545	1.6339	2.9497
CC †	(0.3060)	0.0253	0.0505	(1.0411)	(0.0346)	0.2879	0.2915
TartanVO	4.7080	**0.0345**	0.2832	0.9776	0.1024	0.3494	1.2211
Ours	**1.9579**	0.0859	**0.1172**	**0.9430**	**0.0653**	**0.3029**	**0.6338**

Table 2: Quantitative result of camera pose estimation on the official dynamic sequences of KITTI test dataset. † means the model is fine-tuned on the KITTI dataset and (·) indicates that the sequence is part of the training dataset for the methods of the respective approaches [19]. Both the TartanAir baseline and the proposed network have not been fine-tuned on the KITTI dataset.

roughly 500,000 images; there exists a discrepancy in the numbers of images between the two different training datasets. To solve the data imbalance problem between the TartanAir and GSVCities datasets during the training session, we downsample the number of the GSVCities dataset to that of the TartanAir dataset.

4.2 Quantitative Result

Visual Odometry Result TartanAir. Table 1 shows the overall quantitative result of camera pose estimation; the metric is ATE. We compared the proposed method against three established approaches: ORB SLAM [15], a SLAM method with additional optimization processes, DeepV2D [22] and TartanVO [28,20], deep learning-based VO methods. The proposed method consistently demonstrates superior performance on both small and large displacement trajectories, with a notable margin. This evidence emphasizes the robustness of the proposed network in effectively handling diverse displacement scenarios, validating its suitability for real-world applications.

DynaKITTI. Table 2 [19] shows the overall quantitative result of camera pose estimation on the DynaKITTI [9,19] VO test dataset; the metric is ATE. We compared the proposed method against four conventional approaches: DeepVO [27], TrainFlow [33], and CC [18], which have been fine-tuned on the KITTI dataset and TartanVO [28]. Tar-

tanVO and the proposed network are never fine-tuned on the KITTI dataset. The proposed approach demonstrates superior performance on both small and large displacement trajectories, with a particularly significant margin in the case of large displacement scenarios—while maintaining competitive performance in small displacement scenarios.

Visual Place Recognition Result Table 3 presents the quantitative results of the VPR task on the Pittsburgh 30k and 250k test datasets; the metric is recall@k [12]. We compared the proposed method against four VPR baseline methods as follows: GeM [17], NetVLAD [3], CosPlace [4], and MixVPR [2]. The proposed network exhibits comparable performance in the VPR task. We attribute this comparable performance to the usage of a small batch size. Instead of employing a batch size of 120 as in the MixVPR baseline, we reduced the batch size to 4 due to GPU memory constraints. Furthermore, we employed the multi-similarity loss to learn the VPR task, which can be highly influenced by the batch size. The multi-similarity loss segregates input images into positive and negative pairs within the same batch. We expect that we could enhance the VPR performance by devising a training scheme or exploiting more computational resources for employing larger batch sizes.

Dataset		Pitts30k-test			Pitts250k-test		
Method	Output Dimension	R@1	R@5	R@10	R@1	R@5	R@10
GeM	2048	-	-	-	82.9	92.1	94.3
NetVLAD	32768	86.0	-	-	90.5	96.2	97.4
CosPlace	2048	90.0	95.7	96.7	91.5	96.9	97.9
MixVPR	2048	-	-	-	94.1	98.2	98.8
MixVPR	4096	91.5	95.9	93.6	94.6	98.3	99.0
Ours	2304	84.9	91.8	94.1	84.9	93.1	95.0

Table 3: Quantitative result of the visual place recognition task on Pittsburg 30k and 250k test dataset.

4.3 Qualitative Result

Figure 4 presents the qualitative results of camera pose estimation for large displacement trajectories; the proposed network demonstrates more accurate trajectory estimations. The error along the z-axis is relatively higher in comparison to the errors along the x- and y-axes. We attribute this performance discrepancy between the axes to a potential lack of image depth information. The absence of actual depth information of the input images during the training session could pose challenges in accurately estimating the original three-dimensional information from two-dimensional input images. In future research, concurrently addressing both depth estimation and VPR tasks through multi-task learning would have the potential to significantly enhance camera pose estimation performance in all directions.

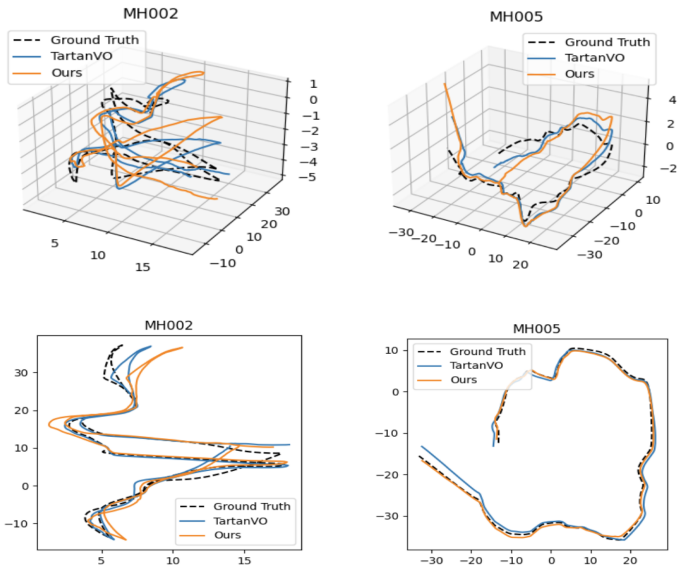

Fig. 3: Comparison of qualitative results of camera pose estimation on large displacement trajectories of official TartanAir test dataset between ground-truth, TartanVO baseline, and ours.

Dataset	TartanAir-test			
Trajectory	Large Displacement (1) MH002	MH005	MH007	(1) Average
TartanVO †	2.00	3.19	2.04	2.41
Ours †	2.15	2.90	3.27	2.77
Ours	**1.85**	**1.90**	**1.91**	**1.88**

Table 4: Ablation study on the training scheme. † indicates the method is trained with the conventional training scheme.

4.4 Ablation Study

Dual Task Learning Table 4 shows the result of the ablation study on the training scheme. The results clearly demonstrate the superiority of the proposed dual-task learning training scheme over the conventional training approach. This observation underscores the effectiveness of the proposed training scheme in enhancing camera pose estimation performance, particularly in scenarios involving large displacement trajectories. The proposed network, when trained using the previous TartanVO training scheme, exhibits slightly lower performance compared to the baseline. We attribute this discrepancy to the utilization of only eighteen environments in our experiments, as opposed to the entire TartanAir training dataset, which the Tartan VO baseline utilized.

Triple-Head Network Table 5 shows the result of the ablation study on the triple-head network. The results indicate that the combination of the proposed flow network with

TartanVO's pose network exhibits a modest improvement compared to the TartanVO baseline. Moreover, the proposed triple-head network, featuring a dedicated VPR head, showcases a substantial enhancement in performance. The superior performance, particularly on large displacement trajectories, validates the efficacy of incorporating the VPR information into the network architecture.

Dataset	TartanAir-test			
Trajectory	Large Displacement (1)			(1) Average
	MH002	MH005	MH007	
TartanVO †	2.00	3.19	2.04	2.41
Our flownet + TartanVO posenet	3.33	**1.81**	1.98	2.37
Ours	**1.85**	1.90	**1.91**	**1.88**

Table 5: Ablation study on the triple-head network. † indicates the method is trained with the conventional training scheme.

5 Conclusion

In this research, we proposed a dual-task network demonstrating robust performance across both small and large displacement trajectories—to the best of our knowledge, the proposed approach represents the first attempt to address dual-task learning between VO and VPR. Our contribution includes the introduction of a transformer-based flow network and a new triple-head network featuring a VPR aggregation module. The utilization of these components enables the incorporation of local and global contexts extracted from the two tasks—enhancing the network's robustness, particularly in addressing challenges posed by large displacement VO trajectories. Through extensive experiments, we verified the effectiveness of the proposed architecture and achieved new state-of-the-art performance. We expect our approach to enlighten the research on deep-learning-based VO by providing a fresh perspective: the significance of extracting global context in handling large displacement and the formulation of dual-task learning for rich feature extraction and computational efficiency.

Acknowledgement This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00926, Development of Self-directed Visual Intelligence Technology Based on Problem Hypothesis and Self-supervised Methods; No. 2022-0-00907, Development of AI Bots Collaboration Platform and Self-organizing AI; No. 2019-0-01842, Artificial Intelligence Graduate School Program (GIST)), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1C1C1009989), and 'Project for Science and Technology Opens the Future of the Region' program through INNOPOLIS FOUNDATION founded by Ministry of Science and ICT (No. 2022-DD-UP-0312-02-101)

References

1. Ali-bey, A., Chaib-draa, B., Giguère, P.: Gsv-cities: Toward appropriate supervised visual place recognition. Neurocomputing 513, 194–203 (2022)
2. Ali-Bey, A., Chaib-Draa, B., Giguere, P.: Mixvpr: Feature mixing for visual place recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2998–3007 (2023)
3. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5297–5307 (2016)
4. Berton, G., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4878–4888 (2022)
5. Camara, L.G., Pivoňka, T., Jílek, M., Gäbert, C., Košnar, K., Přeučil, L.: Accurate and robust teach and repeat navigation by visual place recognition: A cnn approach. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6018–6024. IEEE (2020)
6. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C.: Twins: Revisiting the design of spatial attention in vision transformers. Advances in Neural Information Processing Systems 34, 9355–9366 (2021)
7. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2758–2766 (2015)
8. Forster, C., Pizzoli, M., Scaramuzza, D.: Svo: Fast semi-direct monocular visual odometry. In: 2014 IEEE International Conference on Robotics and Automation. pp. 15–22. IEEE (2014)
9. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research 32(11), 1231–1237 (2013)
10. Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14141–14152 (2021)
11. Huang, Z., Shi, X., Zhang, C., Wang, Q., Cheung, K.C., Qin, H., Dai, J., Li, H.: Flowformer: A transformer architecture for optical flow. In: European Conference on Computer Vision. pp. 668–685. Springer (2022)
12. Jang, S., Kim, U.H.: On the study of data augmentation for visual place recognition. IEEE Robotics and Automation Letters (2023)
13. Kim, U.H., Kim, S.H., Kim, J.H.: Simvodis: Simultaneous visual odometry, object detection, and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(1), 428–441 (2020)
14. Lai, L., Shangguan, Z., Zhang, J., Ohn-Bar, E.: Xvo: Generalized visual odometry via cross-modal self-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10094–10105 (2023)
15. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: Orb-slam: a versatile and accurate monocular slam system. IEEE Transactions on Robotics 31(5), 1147–1163 (2015)
16. Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B.: Large-scale image retrieval with attentive deep local features. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3456–3465 (2017)
17. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence 41(7), 1655–1668 (2018)

18. Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M.J.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12240–12249 (2019)
19. Shen, S., Cai, Y., Wang, W., Scherer, S.: Dytanvo: Joint refinement of visual odometry and motion segmentation in dynamic environments. In: 2023 IEEE International Conference on Robotics and Automation. pp. 4048–4055. IEEE (2023)
20. Shuang Ma, Sai Vemprala, W.W.J.K.G.Y.S.D.M.A.K.: Compass: Contrastive multimodal pretraining for autonomous systems (2022)
21. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8934–8943 (2018)
22. Teed, Z., Deng, J.: Deepv2d: Video to depth with differentiable structure from motion. arXiv preprint arXiv:1812.04605 (2018)
23. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision. pp. 402–419. Springer (2020)
24. Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. Advances in neural information processing systems 34, 16558–16569 (2021)
25. Torii, A., Sivic, J., Pajdla, T., Okutomi, M.: Visual place recognition with repetitive structures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 883–890 (2013)
26. Truong, P., Danelljan, M., Timofte, R.: Glu-net: Global-local universal network for dense flow and correspondences. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6258–6268 (2020)
27. Wang, S., Clark, R., Wen, H., Trigoni, N.: Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In: 2017 IEEE International Conference on Robotics and Automation. pp. 2043–2050. IEEE (2017)
28. Wang, W., Hu, Y., Scherer, S.: Tartanvo: A generalizable learning-based vo. In: Conference on Robot Learning. pp. 1761–1772. PMLR (2021)
29. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4909–4916. IEEE (2020)
30. Wang, X., Maturana, D., Yang, S., Wang, W., Chen, Q., Scherer, S.: Improving learning-based ego-motion estimation with homomorphism-based losses and drift correction. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 970–976. IEEE (2019)
31. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5022–5030 (2019)
32. Younes, G., Asmar, D., Zelek, J.: A unified formulation for visual odometry. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6237–6244. IEEE (2019)
33. Zhao, W., Liu, S., Shu, Y., Liu, Y.J.: Towards better generalization: Joint depth-pose learning without posenet. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9151–9161 (2020)
34. Zhong, Y., Arandjelović, R., Zisserman, A.: Ghostvlad for set-based face recognition. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II 14. pp. 35–50. Springer (2019)
35. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1851–1858 (2017)