2023 IEEE 26th International Conference on
Intelligent Transportation Systems (ITSC)
24-28 September 2023. Bilbao, Bizkaia, Spain

A Novel Place Recognition Network using Visual Sequences and LiDAR Point Clouds for Autonomous Vehicles

Huaiyuan Xu[1], Huaping Liu[2], Shiyu Meng[1], and Yuxiang Sun[1,*]

***Abstract*— Place recognition plays an important role in autonomous vehicles localization, particularly in GNSS-degraded environments. LiDAR-based place recognition (LPR) could achieve localization by comparing on-line LiDAR point clouds with a pre-built off-line point-cloud database. However, LiDAR sensors are expensive, which hinders their large-scale deployment on every vehicle. To alleviate this issue, we propose a novel cross-modal network, which replaces on-line point clouds with on-line images captured by a low-cost and lightweight monocular camera. We use image sequences instead of single images, which would be helpful to eliminate false matches since image sequences capture more environmental information. Furthermore, we propose an image sequence descriptor to represent the observed environment by learning multi-image integration and global representation. Experiments on 6 trajectories of the KITTI dataset demonstrate our effectiveness and superiority over single image-based methods.**

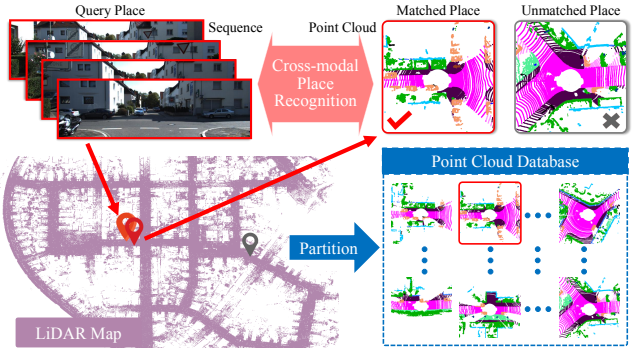

Fig. 1. The figure shows the motivation of our cross-modal place recognition method. Given a query place in the form of an image sequence captured by a monocular camera, our method is designed to find the corresponding place from a LiDAR point-cloud database.

I. Introduction

Place recognition is a fundamental component for autonomous vehicle localization [1], [2]. It generally consists of two stages: off-line stage and on-line stage. The off-line stage usually uses visual cameras [3] or 3-D LiDARs [4] to build environment maps or databases. Then, the on-line stage matches the current sensory data with the pre-built maps or databases to infer the vehicle location. However, this is still a kind of coarse localization [5]. To achieve precise localization, place recognition can be used to provide constraints (e.g., loop closure constraints) to odometry or simultaneous localization and mapping (SLAM) algorithms to estimate vehicle poses [6].

Place recognition is typically treated as a retrieval problem, that is, given a query place, the algorithm finds the corresponding image or point cloud of the matched place in a database. According to the used sensors, most previous works can be roughly divided into two categories: vision-based place recognition (VPR) [7]–[9] and LiDAR-based place recognition (LPR) [10]–[12]. VPR methods use traditional hand-crafted features or learnable descriptors to represent environments captured by visual cameras, and then retrieve places via descriptor matching. Although they can achieve localization in large-scale outdoor environments, the performance could be degraded when environment appearances change, for example, time of day, different weather or seasons [12]. Compared with VPR, LPR attracts more attention because LiDAR point clouds can effectively preserve the geometric structural context of the scene, meanwhile robust to different illumination and weather conditions. However, LPR requires every robot to be equipped with an expensive LiDAR, which limits its large-scale deployment [13]. In contrast, cross-modal place recognition can alleviate this problem by matching low-cost on-line visual images with off-line point clouds.

There are some pioneering works on cross-modal place recognition. One way is to convert cross-modal place recognition into a VPR problem by rendering point clouds into images [14]. Another way is to exploit a shared embedding space to reduce the modality gap between image and point-cloud modalities [13], [15]. However, all these methods ignore the insufficient discrimination of information in a single image, which could lead to false place matching. In this paper, we propose learning-based image sequences to point clouds place recognition, as shown in Fig. 1. It adopts image sequences rather than single images to enhance place discrimination, benefiting from considering more environmental information. We test our method on the KITTI dataset [16] to analyze the effectiveness of our method. Quantitative and qualitative experimental results prove that image sequences outperform single images in cross-modal place recognition. Our contributions are summarized as follows:

- We introduce a novel cross-modal place recognition network[1]. To our best knowledge, this is the first deep learning-based cross-modal method using on-line image sequences and off-line point clouds.
- We design an image-sequence descriptor that encodes the image sequence with rich appearance information

[1]The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (e-mail: huaiyuan.xu@polyu.edu.hk; shiyu.meng@connect.polyu.hk; yx.sun@polyu.edu.hk, sun.yuxiang@outlook.com).
[2]Tsinghua University, Beijing, China. (e-mail: hpliu@tsinghua.edu.cn).
*Corresponding author: Yuxiang Sun.

[1]Our code is available at: https://github.com/lab-sun/VSeq2PC

2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC) | 979-8-3503-9946-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ITSC57777.2023.10421887

into a global representation, thereby describing the observed environment better than a single image.

- We combine multi-image information to extract geometric features of the observed environment by Transformers, which can effectively capture the temporal relationship between images.

II. Related Work

A. Sequence-based Place Recognition

Sequence matching is a general paradigm for sequence-based place recognition [17], [18], which compares each frame of the input sequence with all images in the database, and finds the place corresponding to the sequence through similarity score aggregation. However, this paradigm would be inefficient because its computational cost could be increased with the database size and sequence length.

Several improved methods directly extract sequence descriptors and conduct sequence retrieval in a database [19]–[21]. Facil *et al.* [19] first introduced sequence descriptors in VPR, which integrated image descriptors by three modes, namely concatenation, fully-connected operation, and LSTM networks. Later, researchers studied methods to convolve CNN features of multi-image images to summarize image sequences [20], [21]. Differently, we extract the geometric features of image sequences, and then encode them into global descriptors for sequences.

B. LiDAR-based Place Recognition

Point clouds obtained from LiDARs are robust to illumination and weather changes compared with visual cameras. In LiDAR-based place recognition, learnable descriptors and empirical descriptors are two types of commonly-used LiDAR point-cloud representations. The former benefits from a powerful neural network and is data-driven. PointNetVLAD [22] is a representative work, which extracts point features by PointNet, and uses a NetVLAD aggregator to form a global descriptor for a scene. The latter provides geometric distribution of the data in a more intuitive way. One latest work is Scan Context [4], which divides a point cloud into blocks in the radius and azimuth directions, then counts the maximum height in each block. Furthermore, there are some hybrid algorithms such as MinkLoc3D-SI [23] and RINet [12], combining both advantages of learnable and handcrafted descriptors.

C. Cross-Modal Place Recognition

Reducing the modal gap of input data is the key to cross-modal place recognition. Mithun *et al.* [14] proposed to render LiDAR point clouds into depth images, then used VPR methods to realize place recognition. Besides, some works study a shared embedding space for images and point clouds [13], [15], that is, using neural networks to map images and point clouds to the same high-dimensional space, where data from the same place are close to each other. In contrast, we study place recognition by matching image sequences to point clouds, which utilizes multi-image information and is thus more robust than previous approaches that match single images to point clouds.

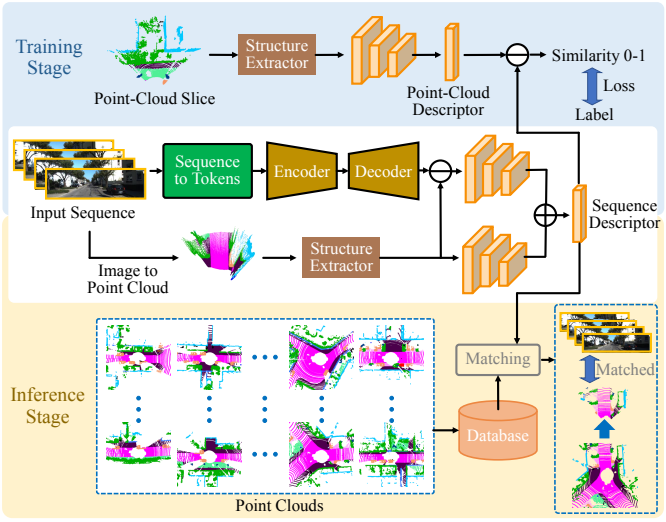

Fig. 2. The pipeline of our network. During the training stage, the network learns to judge whether the input image sequence and point-cloud slice come from the same place. During inference, the network retrieves the point cloud whose slice corresponds to a scene that is consistent with the input sequence. The structure extractor produces point-cloud geometric features [12]. We propose the sequence-to-tokens and encode-decoder modules to extract geometric features of image sequences.

III. The Proposed Method

A. Method Overview

Given a query place and its image sequence, our place recognition network aims to retrieve the corresponding point cloud using visual sequence information, where the retrieved point cloud is consistent with the query sequence in place.

To achieve this goal, we have to extract point-cloud descriptors [12], and integrate multi-image data to generate image-sequence descriptors. Specifically, we encode multi images to generate the structure feature for the image sequence. This structure feature records the geometric characteristics of the observed environment. Afterward, we encode structure features to be a compact image-sequence descriptor. Then, cross-modal place recognition can be realized by comparing image-sequence descriptors with point-cloud descriptors.

B. Place Recognition Network

Our network pipeline is illustrated in Fig. 2. It mainly includes the training stage and inference stage.

1) Training Stage: In the training stage, the network receives the point cloud and image sequence, and outputs their similarity score ranging from 0 to 1. For the input point cloud, we extract its structure features with the structure extractor, which divides the point cloud into S sectors in bird-eye-view, then counts the closest distance of each semantic category in each sector to the point-cloud center [12]. The output of the structure extractor is a feature $F \in \mathbb{R}^{C \times S}$, where C represents the number of semantic classes. This feature is further converted to a compact point-cloud descriptor D_{pc} by applying 1-D convolution and maximum pooling.

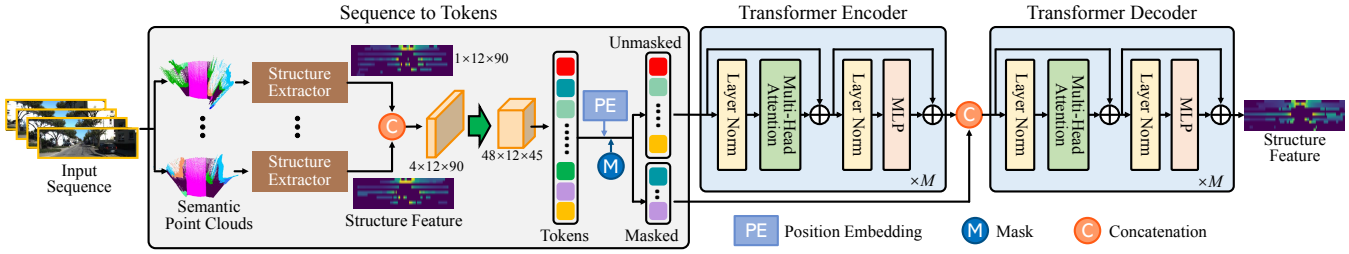

Fig. 3. Learning structure features from image sequences. The input sequence is converted into point clouds, which undergoes structure extractors and convolution process, and is then tokenized. A mask with a 10% masking ratio is employed to enhance the robustness. The visible-token subset is fed to a Transformer-based encoder [24]. The Transformer-based decoder processes the full set of encoded tokens and masked tokens, and then outputs a structure feature that represents the input sequence geometrically.

Similarly, we can compute image descriptors by transforming images to point clouds and then extracting features. For image sequences, considering that the direct concatenation of image descriptors would ignore the intrinsic correlation among different frames, we first integrate the multi-image information to generate the sequence structure feature, and then embed it to obtain the final sequence descriptor (see sec. III-C).

After extracting the image-sequence and point-cloud descriptors, our network calculates their similarity score by performing fully-connected layers on their element-wise difference. Ideally, the similarity of the positive pair is 1, while that of the negatives is 0.

2) Inference Stage: In the inference stage, the network retrieves the 3-D point cloud corresponding to the input query sequence. To do so, the network compares the descriptor of the input sequence with the descriptors of the point clouds from the database, and gives their similarity scores. The network ranks all scores and determines the target point cloud with the maximum similarity.

Notably, considering the view-point inconsistency between the front-looking image sequence and the omni-directional (i.e., 360°) point cloud, we cut the point cloud into slices with different views. Thus, the network compares sequences and point-cloud slices during training, and finds the most similar slice to the sequence during inference. Then, the point cloud of the found slice is our target.

C. Image-sequence Descriptor

Similar to generating point-cloud descriptors, we generate image-sequence descriptors by first extracting sequence structure features and then learning global representations.

1) Sequence Structure Feature: Given an image sequence, we convert each image into a semantic point cloud. Specifically, we estimate the dense depth for each image, then reconstruct a 3-D point cloud using the depth, and finally label the class for each point with semantic image segmentation. Afterward, we extract the structure feature for each image, and then combine multi-image structure features to reconstruct a better one corresponding to the observed scene. We refer to it as the sequence structure feature. The whole process consists of a sequence-to-tokens module, a Transformer encoder, and a Transformer decoder, as shown in Fig. 3.

Specifically, we concatenate the multi-image structure features, and apply a 1×3 convolution kernel on it to fuse the neighborhood information in space. To reduce computational and memory cost, we set the convolution stride to 2 to reduce the number of features. We define that each feature vector corresponds to a token, which is fed to the Transformer-based encoder (i.e., the encoder consisting of Transformer blocks [24]) to be projected to a latent representation. Then, the Transformer-based decoder reconstructs the target data from the latent representations.

We employ a random mask with a masking ratio of 10% on the tokens, so that the full token set can be divided into a visible subset and a masked subset. The input of the encoder is the visible subset, which forces the encoder to learn effective data representations from incomplete signals. The input of the decoder is the encoded tokens and masked tokens. Note that the latter only retains the position embedding to prompt the decoder where the tokens are removed during the encoding process. Finally, the decoder generates an integrated structure feature, which is further used to create a global sequence representation.

2) Global Sequence Representation: The global sequence representation is a compact G-dimensional vector to summarize the scene depicted by the image sequence. To do so, we consider the structure features of the sequence and its first frame (F_{seq}, F_{img}), and generate the sequence representation in a manner of learning a basic representation and its residual. Specifically, we feed F_{img} to a subnetwork f_θ composed of convolutional layers and max pooling layers to obtain a basic descriptor. Then, we compute the difference between F_{seq} and F_{img} and apply another convolution-pooling subnetwork f_ξ on it to estimate the complement. The final global sequence representation D_{seq} is a combination of the basic descriptor and its complement:

$$D_{seq} = f_\theta\left(F_{img}\right) + f_\xi\left(F_{seq} - F_{img}\right). \tag{1}$$

D. Losses

To have better initial weights, we utilize a reconstruction loss to pre-train the part of the sequence structure feature in the network. Then, a place recognition loss is used to train the whole network. The reconstruction loss $\mathcal{L}_{recons}$ computes the mean square error between the sequence structure

TABLE I

STATISTICS OF THE EVALUATION DATASET. K-00 DENOTES THE 00 TRAJECTORY IN THE KITTI ODOMETRY DATASET.

	K-00	K-02	K-05	K-06	K-07	K-08
Images	9082	9322	5522	2202	2202	8142
Scans	4541	4661	2761	1101	1101	4071
Pos. pairs	7399	1691	4773	1412	1353	1970
Neg. pairs	10051955	10696682	3654081	544034	502155	8046762
Loops	804	315	448	270	57	345

feature F_{seq} and the LiDAR point-cloud structure feature F_{pc} corresponding to the first frame in the sequence:

$$\mathcal{L}_{recons} = \frac{1}{\Lambda}\sum_{i=1}^{\Lambda}\left(F_{seq}(i) - F_{pc}(i)\right)^2, \quad (2)$$

where Λ is the feature size. The place recognition loss penalizes recognition errors. Given a place-pair sample, our network outputs a place similarity score s_t. The place recognition loss $\mathcal{L}_{place}$ calculates the average cross-entropy between the place similarity scores of all input samples and their ground-truth labels:

$$\mathcal{L}_{place} = \frac{1}{\Pi}\sum_{t} -\left[y_t \log\left(s_t\right) + \left(1-y_t\right)\log\left(1-s_t\right)\right], \quad (3)$$

where Π is the number of input samples. The label $y_t = 1$ for positive samples, and $y_t = 0$ for negative samples.

IV. EXPERIMENTAL RESULTS

A. Implementation Details

1) Dataset: We evaluate our cross-modal place recognition on the KITTI odometry dataset [16]. It contains 11 trajectories which provide RGB images, LiDAR point clouds, and ground-truth poses. We select 6 trajectories that contain loop closures to evaluate our network. When testing on a certain trajectory, we use the remaining 10 trajectories as the training set and validation set. According to ground-truth poses, we construct the positive and negative place-pair samples for network training. The two places in a positive sample are within 3 meters apart, while those in negative samples are more than 20 meters apart. The statistics of the dataset are displayed in Tab. I.

2) Metrics: We adopt three metrics for evaluation, namely precision-recall curve [1], maximum F1 score [25] and recall@N [7]. The precision-recall curve qualitatively presents the change of precision and recall varied with the threshold of place recognition. The maximum F1 score is the harmonic mean of precision and recall, as a quantitative metric. Recall@N is the percentage of cases where the correct match is ranked within the top N retrievals.

3) Network and Training: Both Transformer-based encoder and decoder have 4 Transformer blocks. Each block contains 3 heads, and the embedding dimension is set to 48. Our 1-D convolution and maximum pooling process in descriptor extraction follows RINet [12]. We train our network with two processes. The first process is pre-training the sequence structure feature to obtain good initial weights.

TABLE II

QUANTITATIVE COMPARISON IN TERMS OF F1 MAXIMUM SCORES (%). K IS SHORT FOR KITTI.

Method	K-00	K-02	K-05	K-06	K-07	K-08	Mean
NetVLAD-based	37.8	22.9	25.2	17.9	47.0	25.1	29.3
RINet-based	53.8	47.8	38.7	17.6	43.3	31.6	38.8
Ours	**67.4**	**51.6**	**44.9**	**18.4**	**56.0**	**35.5**	**45.6**

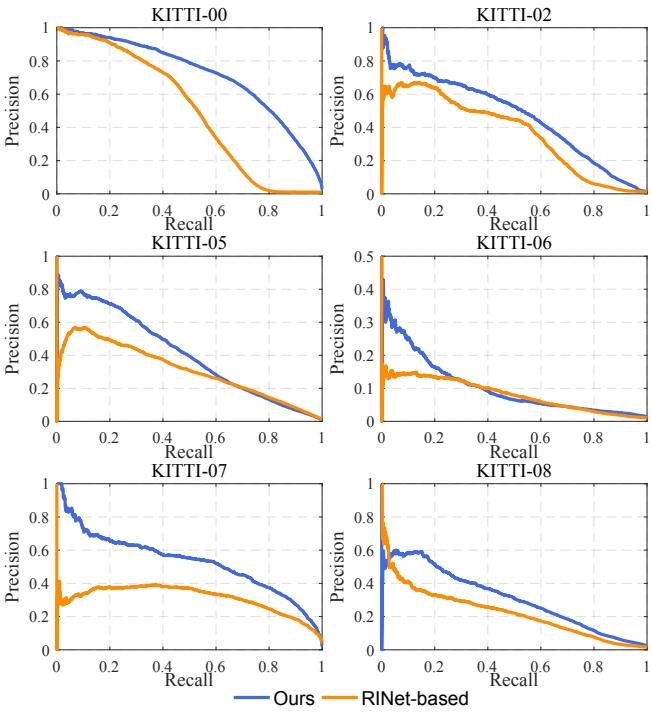

Fig. 4. Precision-recall curves of our method and the RINet-based method on the KITTI dataset. It shows that using image sequences for cross-modal place recognition has better PR curves than using single images.

Then, we train the entire network with 50 epochs to empower the network with the ability of cross-modal place recognition. Both training processes adopt the cosine decay [26] as the learning rate schedule.

B. Evaluation of Place Recognition

Given a set of place pairs, we evaluate the recognition performance of algorithms according to whether they can identify that the input places come from the same or different places. For the test data, we take all positive place-pair samples and a part of the negative samples from the dataset. Following [12], the total number of selected negatives is 100 times that of positives. For comparison, we choose image to point-cloud place recognition methods. However, existing methods have no open-source implementation. Therefore, we improve the classic vision-based place recognition algorithm NetVLAD [27] and one of the latest LiDAR-based place recognition methods RINet [12], so that they can adapt to our cross-modal place recognition task.

To quantitatively and qualitatively analyze the NetVLAD-based method, RINet-based method and our method, we adopt the maximum F1 score and precision-recall curve

TABLE III

COMPARISON OF RECALL@N PERFORMANCE (%) IN LOOP CLOSURE DETECTION ON THE KITTI DATASET.

Method	KITTI 00			KITTI 02			KITTI 05			KITTI 06			KITTI 08			Mean		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
RINet-based	59.7	71.0	77.6	20.6	38.1	49.8	39.5	58.3	66.7	31.1	**59.4**	66.9	30.4	49.6	58.6	36.3	55.3	63.9
Ours	**71.6**	**83.2**	**87.3**	**28.3**	**55.6**	**63.5**	**41.7**	**63.4**	**71.9**	**31.5**	53.4	**67.3**	**36.8**	**62.9**	**71.0**	**42.0**	**63.7**	**72.2**

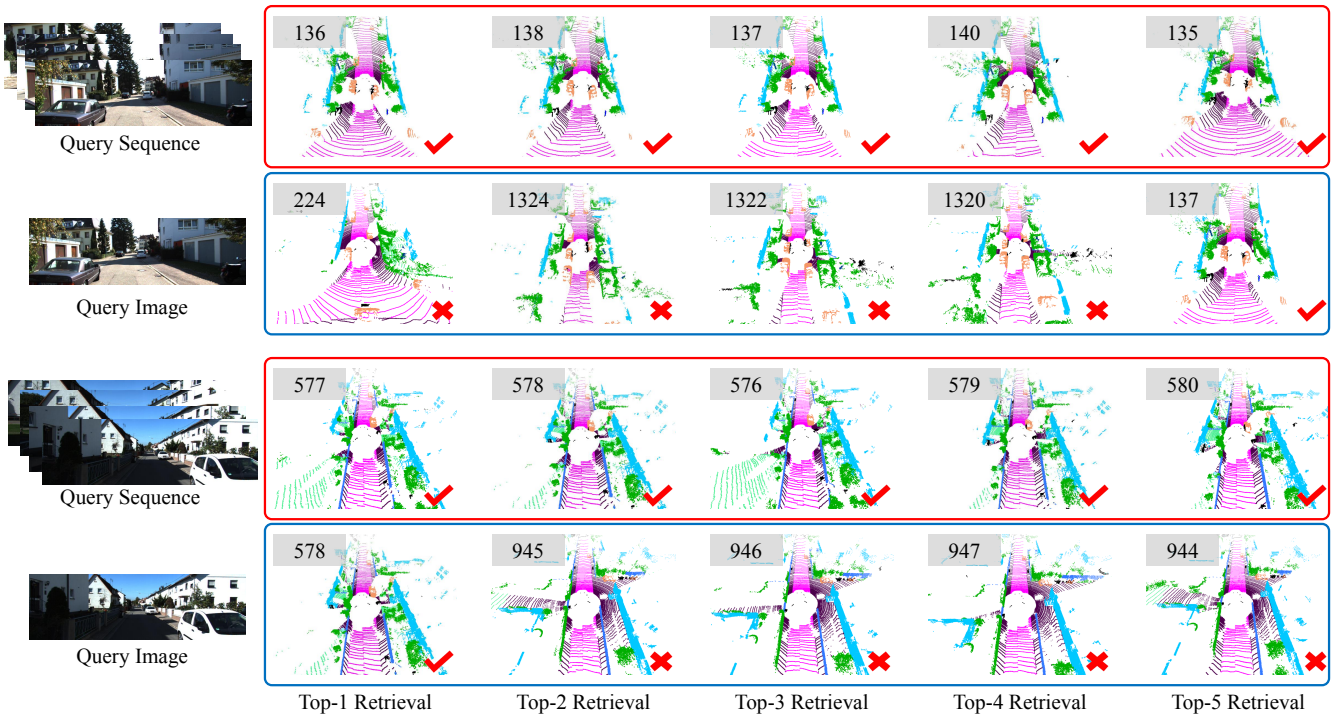

Fig. 5. Visualization of the top 5 retrievals. The upper query sequence is from 00 trajectory, and the lower one is from 05 trajectory. The annotated numbers represent the indices of places. The red box presents our retrieval results, and the blue one shows the results of the RINet-based method. ✔ indicates a correct retrieval whose spatial distance from the query place is within 5 meters, and ✖ means an incorrect retrieval.

metrics, respectively. The results are presented in Tab. II and Fig. 4. It can be seen that the per-trajectory and mean maximum F1 scores of our method are higher than those of the NetVLAD-based method and RINet-based method on all the 6 trajectories, indicating the more accurate performance of our method in cross-modal place recognition. In other words, our method achieves a better balance between precision and recall. Fig. 4 visualizes the precision-recall curves of the best two methods, that is, our method and the RINet-based method. It can be seen that the curves of our method are better than the RINet-based method. In other words, our method can have higher precision and recall at the same time. Combining quantitative and qualitative results, we can also conclude that using sequences (our method) for cross-modal place recognition leads to better performance than using images (RINet-based method and NetVLAD-based method).

C. Evaluation of Loop Closure Detection

Loop closure detection requires the algorithm to identify the revisited place which is consistent with the current place in space, thereby forming a closed loop on the trajectory. We can use loop closure detection to evaluate the place recognition performance of different algorithms. Specifically, given a query place, the algorithm compares it with all previous but not nearby places, ranks them according to similarity scores, and checks whether any matched places among the top K places can form a closed loop. We employ the recall@N metric and set $N = 1, 5, 10$. Recall@5 and recall@10 are necessary because they can perform geometrical verification to re-rank the candidate places and find the correct one in the top 5 and top 10 retrievals. We do not test algorithms on the 07 trajectory due to the small number of loops. In addition, we do not evaluate the NetVLAD-based method, because it belongs to image to point-cloud algorithms like the RINet-based method but has worse performance.

The quantitative comparison is shown in Tab. III. The average performance of our method on five trajectories is significantly better than the RINet-based method, with (recall@1, recall@5, recall@10) higher by $(5.7\%, 8.4\%, 8.3\%)$. It is worth noting that although the loops of trajectory 08 are all reverse loops (that is, the orientation of the robot arriving at the same place twice is opposite), our method still achieves good performance, whose recall@10 value reaches 71%. This indicates that the detection is robust to orientation

changes. Fig. 5 shows the qualitative results, from which it can be seen that the method using sequences to describe the query place can accurately retrieve matching places to make up loop closures, whereas the method using images would mistakenly identify other similar places as matches. This is because sequences can provide more environment details than images, which contributes to more accurate place recognition.

V. Conclusions

In this paper, we investigated cross-modal place recognition. To the best of our knowledge, this is the first work studying deep learning-based cross-modal method using online image sequences and off-line point clouds. We proposed a novel cross-modal network that recognizes places by extracting descriptors of visual sequences and LiDAR point clouds, and then comparing them. To extract image-sequence descriptors, we integrated multi-image information to generate sequence structural features using Transformers with a mask, and then learned the basic descriptors and their complements. The experiments demonstrate the effectiveness of the proposed network, which can even successfully detect challenging loop closures such as reverse loops. The evaluations of place recognition and loop closure detection both illustrate that using image sequences leads to better performance in cross-modal place recognition compared to using single images.

Acknowledgement

This work was supported in part by National Natural Science Foundation of China under Grant 62003286, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010116, in part by Zhejiang Lab under grant 2021NL0AB01, and in part by HK PolyU under Grants P0038980 and P0034801.

References

[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, 2015.

[2] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," *Pattern Recognit.*, vol. 113, p. 107760, 2021.

[3] M. U. M. Bhutta, Y. Sun, D. Lau, and M. Liu, "Why-so-deep: Towards boosting previously trained models for visual place recognition," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1824–1831, 2022.

[4] G. Kim, S. Choi, and A. Kim, "Scan context++: Structural place recognition robust to rotation and lateral variations in urban environments," *IEEE Trans. Robot.*, vol. 38, no. 3, pp. 1856–1874, 2021.

[5] F. Cao, F. Yan, S. Wang, Y. Zhuang, and W. Wang, "Season-invariant and viewpoint-tolerant lidar place recognition in gps-denied environments," *IEEE Trans. Ind. Electron.*, vol. 68, no. 1, pp. 563–574, 2020.

[6] D. Cattaneo, M. Vaghi, and A. Valada, "Lcdnet: Deep loop closure detection and point cloud registration for lidar slam," *IEEE Trans. Robot.*, 2022.

[7] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, "Transvpr: Transformer-based place recognition with multi-level attention aggregation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13 648–13 657.

[8] B. Ferrarini, M. J. Milford, K. D. McDonald-Maier, and S. Ehsan, "Binary neural networks for memory-efficient and effective visual place recognition in changing environments," *IEEE Trans. Robot.*, vol. 38, no. 4, pp. 2617–2631, 2022.

[9] R. Mereu, G. Trivigno, G. Berton, C. Masone, and B. Caputo, "Learning sequential descriptors for sequence-based visual place recognition," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 10 383–10 390, 2022.

[10] P. Yin, F. Wang, A. Egorov, J. Hou, Z. Jia, and J. Han, "Fast sequence-matching enhanced viewpoint-invariant 3-d place recognition," *IEEE Trans. Ind. Electron.*, vol. 69, no. 2, pp. 2127–2135, 2021.

[11] J. Ma, J. Zhang, J. Xu, R. Ai, W. Gu, and X. Chen, "Overlaptransformer: An efficient and yaw-angle-invariant transformer network for lidar-based place recognition," *IEEE Robot. Autom. Lett.*, 2022.

[12] L. Li, X. Kong, X. Zhao, T. Huang, W. Li, F. Wen, H. Zhang, and Y. Liu, "Rinet: Efficient 3d lidar-based place recognition using rotation invariant neural network," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 4321–4328, 2022.

[13] D. Cattaneo, M. Vaghi, S. Fontana, A. L. Ballardini, and D. G. Sorrenti, "Global visual localization in lidar-maps through shared 2d-3d embedding space," in *Proc. IEEE Int. Conf. Robot. Autom.* IEEE, 2020, pp. 4365–4371.

[14] N. C. Mithun, K. Sikka, H.-P. Chiu, S. Samarasekera, and R. Kumar, "Rgb2lidar: Towards solving large-scale cross-modal visual localization," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 934–954.

[15] Z. Zhao, H. Yu, C. Lyv, W. Yang, and S. Scherer, "Attention-enhanced cross-modal localization between 360 images and point clouds," *arXiv preprint arXiv:2212.02757*, 2022.

[16] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012.

[17] K. L. Ho and P. Newman, "Detecting loop closure with scene sequences," *Int. J. Comput. Vis.*, vol. 74, pp. 261–286, 2007.

[18] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. IEEE Int. Conf. Robot. Autom.* IEEE, 2012, pp. 1643–1649.

[19] J. M. Facil, D. Olid, L. Montesano, and J. Civera, "Condition-invariant multi-view place recognition," *arXiv preprint arXiv:1902.09516*, 2019.

[20] S. Garg, B. Harwood, G. Anand, and M. Milford, "Delta descriptors: Change-based place representation for robust visual localization," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 5120–5127, 2020.

[21] S. Garg and M. Milford, "Seqnet: Learning descriptors for sequence-based hierarchical place recognition," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 4305–4312, 2021.

[22] M. A. Uy and G. H. Lee, "Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4470–4479.

[23] K. Żywanowski, A. Banaszczyk, M. R. Nowicki, and J. Komorowski, "Minkloc3d-si: 3d lidar place recognition with sparse convolutions, spherical coordinates, and intensity," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 1079–1086, 2021.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[25] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008, vol. 39.

[26] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *Proceedings of the International Conference on Learning Representations*, 2017.

[27] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5297–5307.