2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

# Enhancing Visual Place Recognition with Multi-modal Features and Time-constrained Graph Attention Aggregation

Zhuo Wang, Yunzhou Zhang*, Xinge Zhao, Jian Ning, Dehao Zou, Meiqi Pei

***Abstract*— Visual place recognition(VPR) is a crucial technology for autonomous driving and robotic navigation. However, severe appearance and perspective changes often lead to degradation of algorithm performance. Current methods mainly utilize single-modality RGB images, which are sensitive to environmental changes. To address this challenge, we propose a novel multi-modal visual place recognition method by incorporating depth information as auxiliary data to enhance the robustness of the VPR algorithm. The pipeline involves dual-branch feature extraction and shared multi-modal feature fusion based on transformer(SFFM) to enable full interaction between semantic and structural information. Furthermore, we introduces a time-constrained graph attention aggregation(TC-GAT) that propagates node information across time and space to deal with perceptual aliasing. Extensive experiments on the Oxford Robotcar and MSLS datasets demonstrate that the proposed algorithm is not only effective in appearance changes but also competitive in opposing viewpoints.**

## I. Introduction

Visual place recognition(VPR) refers to the task of recognizing whether a location has been previously visited based on visual clues captured in the image. It plays a significant role of place perception and comprehension and is widely used in simultaneous localization and mapping(SLAM)[1], augmented reality(AR)[2], and autonomous driving[3]. However, VPR remains challenging due to dynamic environmental changes[4][5][6] and annoying perceptual aliasing[7]. Therefore, obtaining a robust and discriminative place representation has become one of the current research frontiers.

Current place recognition methods primarily rely on single-modality RGB images to construct place descriptors. Some[8][9][10][11][12] extract information from the whole image to build global and compact descriptors, but they usually fail to capture discriminative local regions, which limits their ability to handle perspective changes. Others[15][13][14][19] extract local keypoints and perform cross-matching and spatial verification, but such methods suffer from a drawback in slower matching speed. To achieve a good trade-off between the recognition accuracy and computational complexity, researchers[16][17][18] have proposed a two-stage place matching pipeline where top-N candidates are generated using global descriptors firstly, and then these candidates are re-ranked using local descriptors.

*The corresponding author of this paper
All authors are with College of Information Science and Engineering, Northeastern University, Shenyang 110819, China. zhangyunzhou@mail.neu.edu.cn

This work was supported by National Natural Science Foundation of China (No. 61973066, 61471110) and Major Science and Technology Projects of Liaoning Province(No. 2021JH1/10400049).

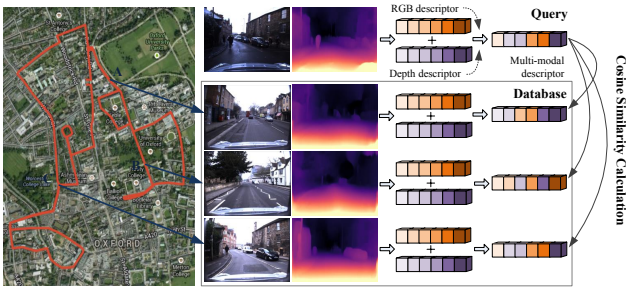



Fig. 1. **Depth-based multi-modal visual place recognition.** Global descriptors are computed by RGB images from camera and depth images from monocular depth estimation. Cosine similarity is used to calculate distance between query and database descriptors.

However, there is an inherent issue with image-based methods that RGB images are susceptible to interference factors and can only provide limited information.

Depth images have stable 3D structure and immune to factors like lighting and color which can be easily acquired through depth sensors[20] or monocular depth estimation[21]. Some researchers[22][23] introduce depth modality for VPR. [24] embedded depth information into feature space through a multi-task architecture of depth prediction and semantic segmentation. [23] utilized late-fusion strategy to encoder RGB and depth information to get robust place representations. Inspired by [23], we also introduce depth information as auxiliary information to solve place recognition problem. RGB images contain rich content information, but are susceptible to appearance and viewpoint variations. Depth images have stable geometric structure but contain relatively less semantic information. Therefore, we propose a multi-modal network architecture that fuses RGB and depth information to maximize the advantages of both and improve the accuracy and stability of place recognition, as shown in Fig.1.

Perceptual aliasing is another tricky problem in VPR. When perceptual aliasing occurs, the descriptors of two images may be very similar, resulting in a small spatial distance between them. However, this does not necessarily mean that they are taken at the same location. To address this issue, some methods[7][25][26][27][28] integrated sequential information for scene modeling and performed sequence-to-sequence or sequence-to-image matching. However, directly aggregating sequential frames without selection poses a significant challenge to the reliability of the model due to potential noise in image acquisition. To overcome this problem, the emerging graph attention networks(GAT) have drawn the attention of researchers. [29] achieved better recognition

results by propagating influence messages at space level using the GAT. Motivated by [26][29], we consider using a graph attention network with time constraints to dynamically explore potential relationships between sequential samples at time and space levels through attention mechanism to deal with perceptual aliasing.

In summary, our contributions are as follows:

- We propose a novel and effective baseline for place recognition, which integrates content and geometry information to built robust multi-modal global descriptors.
- A shared multi-modal feature fusion module based on transformer(SFFM) is designed to fuse RGB and depth features, and integrate multi-scale information.
- A time-constrained graph attention aggregation(TC-GAT) is designed to interact with descriptors information at both temporal and spatial levels to address perceptual aliasing and improve recognition accuracy.
- We conduct a series of experiments on the Oxford RobotCar[30] and MSLS[31] datasets, and the results demonstrate that our method achieves superior performance respond to drastic appearance and viewpoints variations compared to state-of-the-art methods.

## II. Related Works

### A. Visual Place Recognition

The main challenge of VPR lies in uncontrollable environmental changes. Deep learning-based methods have greatly advanced the development of VPR technology. NetVLAD[8] used triplet loss to learn the optimal VLAD encoding to construct global place representations. PatchNetVLAD[13] divided feature maps into patches of different sizes and performed feature matching from both global and local perspectives. MR-NetVLAD[32] was designed to use a low-dimensional image pyramid to capture more detailed information. SeqNet[26] exported temporal information to avoid perceptual aliasing. However, the aforementioned methods only apply the single-modality RGB images, which are sensitive for environmental changes.

On the other hand, depth maps reflect the geometric structure and distance relationships, which are highly invariant to appearance change. [22] utilized the disparity of the depth map to filter keypoints and constructed a local topological representation of the reference image sequences. However, this method relies on the accuracy of keypoints extraction and is not effective in low-light conditions. Piasco[23] designed a multi-task network and simply added depth features with RGB features in last layer to generate a global descriptor. However, this method does not fully integrate the content information of RGB features with the geometric information of depth images. In this paper, we design a shared feature fusion module based on transformer, which allows for comprehensive feature interactions, and achieves better results.

### B. Graph Neural Networks

Graph Neural Networks(GNN) was first proposed by Goio et al.[33] and widely used in natural language processing and social network fields. A typical GNN consists of nodes and edges, where each node is influenced by all its adjacent edges. [34] introduced attention mechanism and proposed graph attention networks(GAT), which enables network to focus on the nodes and edges relevant to the task.

With the development of GNN, this technology has gradually applied to other tasks. [35] proposed spatio-temporal graph convolutional networks for human action recognition, meanwhile [36] introduced similarity-guided graph neural networks to learn the relationship between reference and query images for pedestrian re-identification. [29] applied GAT to place recognition and used GAT to propagate influence messages to refine feature maps. Inspired by [29], we also adopt the GAT to propagate node information. Distinct from [29], we directly construct a graph model with the generated place descriptors as nodes to explore the potential relationships in space. Furthermore, we impose soft time constraints on the propagation mechanism of nodes which helps to suppress similar nodes which belong to different locations and avoids perceptual aliasing.

## III. Method

We present a multi-modal visual place recognition pipeline that combines RGB and depth information. First, depth images are obtained through a self-supervised monocular depth estimation module. Then, RGB and depth images are fed into a dual-branch feature extraction network with a shared feature fusion module based on transformer(SFFM) to capture robust scene features. A time-constrained graph attention aggregation(TC-GAT) module is employed to integrate temporal and spatial information of the descriptors. Finally, we utilize the cosine similarity to calculate the distance between the reference descriptor and the query descriptor to yield the most similar candidate image. The framework of our proposed approach is illustrated in Fig.2(a).

### A. Monocular Depth Estimation

To obtain the depth image, the self-supervised monocular depth estimation network Monodepth2[39] is adopted to estimate the pixel-level depth map of input image. Monodepth2 uses ResNet50 as backbone and employs skip connections to directly connect features of encoder to decoder. This approach generates disparity through the network and then converts it to a depth map. In monodepth2, depth ground truth is not required, instead utilizing inter-frame geometric constraints as supervision signal for training on the KITTI datasets. We choose this method because it has good generalization and can be used on datasets without depth ground truth. In our experiment, we use the model only trained on the KITTI datasets without performing any fine-tuning.

### B. Dual Branch Feature Extraction

Our dual branch feature extraction consists of RGB feature extraction and depth feature extraction. For RGB feature extraction branch, we adopt VGG16[40] which removes the last pooling layer and fully connected layer removed as the RGB feature encoder to capture rich content information. Given the input RGB image $I_R$, the output of the encoder is a

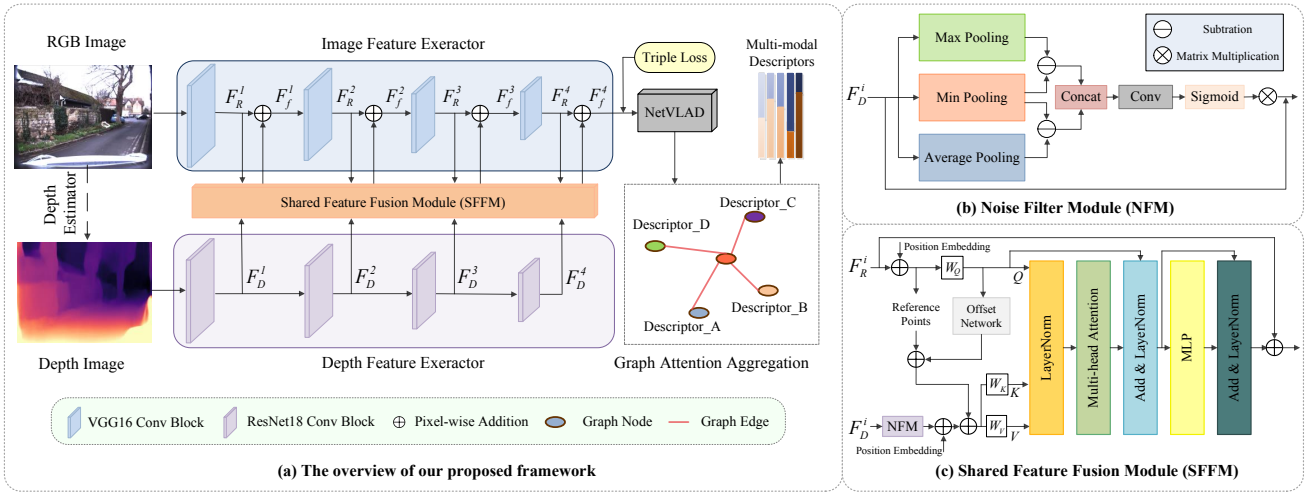



Fig. 2. **The overview of our proposed method.** (a) Pipline. The dual-branch feature extraction network and shared feature fusion based transformer(SFFM) is intorduced to construct multi-modal descriptors. Then, graph attention aggregation is employed to refine descriptors. (b) The structure of the noise filter module(NFM). (c) The structure of shared feature fusion module based on transformer(SFFM).

set of feature maps $F_R^i(i=1,2,3,4)$, where represents the i-th layer RGB feature map. For depth feature extraction branch, we adopt truncated ResNet18[23] with residual connections as the depth feature encoder to avoid gradients vanishing caused by relatively less information in depth images. Given the input depth image $I_D$, the output of the depth encoder is a set of depth feature maps $F_D^i(i=1,2,3,4)$.

### C. Shared Feature Fusion Module

Inspired by the potent multi-head attention in the transformer[43], adept at capturing global context, we employ the transformer to fuse RGB and depth features to fully interact content and geometric information. Unlike [43], our approach introduces three changes:

(1) We propose a noise filter module (NFM) to filter noise and guide the network to focus on geometric structure and distance relationship of the depth map. We take RGB and filtered depth features by NFM as inputs of shared feature fusion module.

(2) We replace the standard multi-head attention with deformable attention[44], assigning RGB sequences to queries and depth sequences to keys and values which allows the network to min supplementary distance information from regions of interest in the depth features.

(3) In order to integrate multi-scale information and reduce the model parameter, a shared transformer block is proposed, in which the patch embedding is obtained from $F_R^i$ and $F_D^i$, while the remaining parts are shared across the network. The strategy is grounded in the inherent flexibility of transformer, which can adapt to input sequences of varying lengths.

**Noise filter module(NFM)** utilizes the minimum pooling to suppress noise to achieve more pure depth features, which is inspired by the spatial attention module in CBAM[42].

As is illustrated in Fig.2(b), the depth feature map $F_D$ is subjected to maximum pooling, minimum pooling, and average pooling along the channel, and the corresponding results are denoted as $F_{DMax}$, $F_{DMin}$ and $F_{DAvg}$. Then, the results of maximum pooling and average pooling are subtracted by the result of minimum pooling. We concat the corresponding results along channel dimension and then normalized to obtain the weight factors. The filtered feature map $\widetilde{F_D}$ is obtained by multiplying the weight factor with the original feature map. The NFM module is formulated as follows,

$$\widetilde{F_D} = \sigma(Conv([F_{DMax} - F_{DMin}; F_{DAvg} - F_{DMin}])) \times F_D \quad (1)$$

where $\sigma$ represents sigmoid function and $Conv$ represents the $7\times 7$ convolution operation.

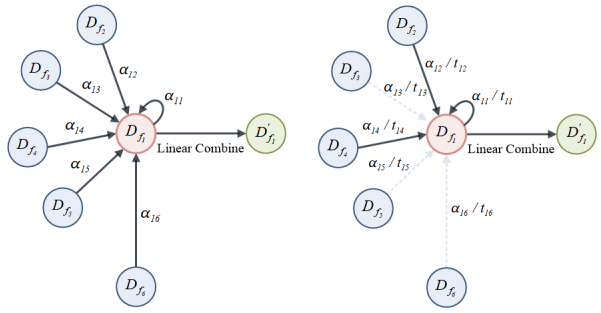
**The shared feature fusion module (SFFM)** takes RGB features $F_R^i \in \mathbb{R}^{H\times W\times C}$ and depth features $\widetilde{F_D^i} \in \mathbb{R}^{H\times W\times C}$ filtered through the NFM as input, as illustrated in Fig.2(c). We split inputs into nonoverlapping patches, linearly project each of them to $C^{'}$ dimensions and add position embeddings to obtain sequence vectors $Z_R^i \in \mathbb{R}^{N\times C^{'}}$, $Z_D^i \in \mathbb{R}^{N\times C^{'}}$, where $C^{'} = P^2 \times C$, $N = HW/P$ is the number of patches and $P$ is the size of patch. Given a flattened RGB feature map $Z_R^i$, a uniform grid of points $p \in \mathbb{R}^{N_G\times 1}$ are generated as the references, where $N_G = N/r$, $r$ indicates downsampling factor. The values of reference points are linearly spaced 1D coordinates and normalize them into $(-1,+1)$, which $-1$ indicates the leftmost and $+1$ indicates the rightmost. To obtain offset for each reference points, the flattened RGB feature map $Z_R$ are linearly projected to query $Q_R = Z_R \times W_Q$, where $W_Q \in \mathbb{R}^{C^{'}\times C^{'}}$ is projection matrix. Then $Q_R$ is fed into offset network which consists of depthwise convolution, GELU activation and a $1\times 1$ convolution to get 1D offset, $\triangle p = Offset(Q_R)$. The deformed keys and values are calculated as:

$$K_D = \overline{Z_D} \times W_K, V_D = \overline{Z_D} \times W_V \quad (2)$$

where $\overline{Z_D} = \phi(Z_D : p + \triangle p)$, $\phi()$ is interpolation operation and $W_k, W_V \in \mathbb{R}^{C^{'}\times C^{'}}$. Multi-head attention with $m$ heads is formulated as:

$$z = sigmoid(Q_R K_D/\sqrt{d})V_D \quad (3)$$

where $d = C^{'}/m$ is the dimensions of each head. The transformer block is formulated as:




(a) Node updating of standard GAT (b) Node updating of TC-GAT

Fig. 3. The diagram of the time-constrained graph attention aggregation.

$$z^{'} = MHSA(LN(z)) + z, \overline{F_D} = MLP(LN(z^{'})) + z^{'} \quad (4)$$

After that, the multi-modal features are acquired by pixel-wise addition along the channel dimension, $F_{fusion} = F_R + \overline{F_D}$. Finally, the NetVLAD layer[8] is used to obtain the multi-modal place descriptors $D_f$. The triplet loss is adopted to encourage the network to generate robust global descriptors.

### D. Time-constrained Graph Attention Aggregation

In this section, we propose a time-constrained graph attention aggregation(TC-GAT) to explore the spatial information of descriptors and ulitize soft time constraints to effectively suppress perceptual aliasing.

We construct an undirected complete graph $G(V,E,T)$:

$$\begin{cases} V = \{v_1, v_2, ...v_n\} \\ E = \{e_{ij}\}, T = \{t_{ij}\} \end{cases} \quad (5)$$

where $V$ denotes the set of nodes and $v_i \in R^{d^{'}}$ , $d^{'}$ denotes the node vector dimension. In this work, each node denotes a multi-modal global descriptor. The edges $E$ between nodes are expressed by Euclidean distance, that is $e_{ij} = \left\| v_i - v_j \right\|_2$ that denotes the spatial relationship between node $v_i$ and $v_j$.

To compare the time differences between different nodes, we normalize the calculation to get soft time label $t_{ij} = t_{ij} / \sum_{k \in N} t_{ik}$. At the same time, in order to represent the dependencies between different descriptors, that is, the relative importance between nodes, we define an attention matrix which introduces soft time label and uses the L1 norm for normalization. The attention matrix can be expressed as:

$$\alpha_{ij} = \frac{exp(-e_{ij}^2/\eta)/t_{ij}}{\sum_{k \in N} exp(-e_{ik}^2/\eta)/t_{ik}} \quad (6)$$

where $\eta$ is hyperparameters. The introduction of soft time label effectively filters out nodes with a large time gap, reducing the number of nodes in the graph and increasing computation speed. Refined description nodes are obtained by linearly computing the original features and attention matrix as illustrated in Fig.3. The formula is expressed as:

$$\overline{v_i} = (1-\theta)v_i + \theta \sum_{k \in N} \alpha_{ik} v_i \quad (7)$$

where $\theta$ is balance parameter. To reduce computational complexity, we use PCA to reduce the dimension of descriptors from 32768 to 4096 before building the graph.

## IV. Experiment

In this section, we introduce the experimental preparation and then give the comparison results with the state-of-the-art place recognition methods. In addition, we conduct comprehensive ablation and qualitative analysis to demonstrate the effectiveness of the proposed algorithm.

### A. Experimental Preparation

**Datasets.** We evaluate the effectiveness of our algorithm on three benchmark datasets: Oxford RobotCar[30], Mapillary Street Level Sequences (MSLS)[31] and Synthia[45].

Oxford RobotCar datasets traverse the city center of Oxford, which include various appearance changes. We train and test our algorithm on six sequences under different appearance conditions. The details are shown in Tab. I.

MSLS datasets are the largest and most diverse datasets in the field of VPR, collected from different cities and including various environment conditions. In our experiments, we follow the same experimental settings as [26].

Synthia dataset is rendered from virtual city scenes, including high-resolution RGB images, high-quality semantic segmentation map, and depth ground truth. We use Synthia dataset to verify quality of depth estimation.

**Compared Methods.** We compare our method with state-of-the-art approaches to validate effectiveness. AP-GeM[11] utilizes generalized mean pooling(GeM) and a listwise ranking loss to directly optimize mean Average Precision(mAP). NetVLAD[8] utilizes differentiable VLAD-pooling features. SeqNet[26] aggregates temporal information to re-rank candidate images. PatchNetVLAD[13] extracts patch-level features for cross-matching. MR-NetVLAD[32] uses low-resolution image pyramid to capture more richer representations. Look no deeper[22] uses the depth disparity information to filter local keypoints. For appearance changes, we compare with AP-GeM, NetVLAD, SeqNet, PatchNetVLAD, and MR-NetVLAD. For viewpoint changes, we compare with NetVLAD and look no deeper.

**Evaluation Metrics.** We use Recall(Recall@N) as the evaluation metric, which is defined as the proportion of correctly retrieved positive samples(the top N) results in a specified radius range to the total number of positive samples.

**Implementation Details.** The experiments for VPR are performed using Pytorch framework on an NVIDIA Titan XP GPU with 12 GB memory. We transform input RGB and depth images to $224 \times 224$ and employ the SGD optimizer with a momentum of 0.9 and a weight decay of 0.001 to optimize network. Furthermore, the total number of iterations

TABLE I

Sequences selection

| Environment conditions | Sequences selection |
|---|---|
| Overcast - Night | 2015-03-17-11-08-44 & 2014-12-16-18-44-24 |
| Snow - Night | 2015-02-03-08-45-10 & 2014-12-16-18-44-24 |
| Summer - Night | 2015-08-18-09-50-22 & 2014-12-16-18-44-24 |
| Overcast - Overcast | 2015-03-17-11-08-44 & 2014-12-09-13-21-02 |
| Snow - Overcast | 2015-02-03-08-45-10 & 2015-03-17-11-08-44 |
| Summer - Overcast | 2015-08-18-09-50-22 & 2015-03-17-11-08-44 |

TABLE II

PERFORMANCE COMPARISON OF RECALL@N WITH STATE-OF-THE-ART METHODS ON OXFORD ROBOTCAR DATASETS. THE BEST RESULTS AND SUB-OPTIMAL RESULTS ARE HIGHLIGHTED IN BOLD UNDERLINED AND BOLD, RESPECTIVELY.

| Method | Night | | | Day | | |
|---|---|---|---|---|---|---|
| | Overcast-night | Winter-night | Summer-night | Overcast-overcast | Winter-overcast | Summer-overcast |
| | Recall@1/5/20 | Recall@1/5/20 | Recall@1/5/20 | Recall@1/5/20 | Recall@1/5/20 | Recall@1/5/20 |
| AP-GeM[11] | 0.286/0.498/0.709 | 0.255/0.457/0.667 | 0.176/0.370/0.580 | 0.605/0.707/0.796 | 0.563/0.695/0.799 | 0.621/0.770/0.868 |
| NetVLAD[8] | 0.819/0.940/0.974 | 0.681/0.864/0.908 | 0.358/0.566/0.747 | 0.657/**0.763**/0.823 | 0.640/**0.769**/**0.832** | 0.843/0.925/0.963 |
| PatchNetVLAD[13] | 0.739/0.890/0.965 | 0.658/0.880/0.929 | 0.446/0.676/0.827 | 0.583/0.729/0.810 | 0.592/0.752/0.825 | 0.718/0.882/0.953 |
| SeqNet[26] | 0.834/0.936/0.976 | **0.877**/0.949/0.959 | 0.537/0.688/0.780 | **0.658**/0.726/0.832 | **0.657**/0.732/0.828 | **0.930**/**0.967**/**0.982** |
| MR-NetVLD[32] | **0.911**/**0.977**/**0.986** | 0.865/**0.971**/**0.987** | **0.556**/**0.749**/**0.880** | 0.657/0.760/**0.850** | **0.652**/**0.775** /**0.832** | 0.889/0.940/**0.986** |
| Ours | **0.936**/**0.980**/**0.996** | **0.886**/**0.974**/**0.989** | **0.656**/**0.832**/**0.921** | **0.664**/**0.764**/**0.834** | 0.634/0.766/**0.835** | **0.930**/**0.968**/**0.982** |

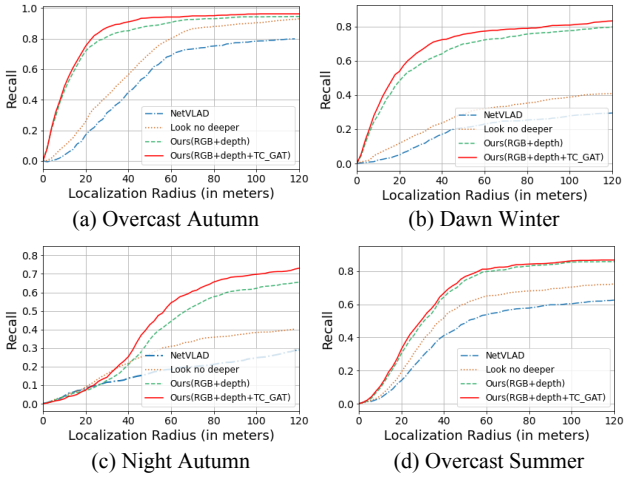
is 60 and initial learning rate is 0.00001 with a decay factor of 0.5 every 50 epochs. For other hyperparameters, the number of cluster centers is set to 64, balance parameter $\theta$ is 0.8, $\eta$ in attention matrix is 20, the head of multi-head attention $m$ is 4, and the localization radius is set as to be 10m for the Oxford datasets and 20m for the MSLS datasets.

### B. Comparison to State-of-the-art Methods

**Results on the Oxford RobotCar dataset.** As is illustrated in Tab. II, we conduct experiments on six sequences which three sets for daytime conditions and three sets for nighttime conditions of the RobotCar datasets to verify the effectiveness of our method. It is evident that our method achieves the best or second-best performance. Notably, compared to NetVLAD that solely relies on RGB modality, our approach yields an average increase on Recall@1 of 20% during the evening and an average increase of approximately 3% during the day. The results demonstrate that our method is very effective and can make up for the disadvantage of insufficient semantic information to a large extent in the night and other feature degradation conditions.

**Results from the opposing viewpoint.** We conduct experiments under opposing viewpoints following the same experimental settings as in [22]. Four sets of experiments are performed using four forward-view sequences for different environmental settings. Furthermore, we calculate the matching ground truth for the front-view and rear-view sequences using GPS and use recall as the evaluation metric. The experimental results are presented in Fig.4, which clearly show that our method achieves the best performance. Particularly, our method outperforms [22] by almost twice in the Dawn Winter and Night Autumn scenarios, where the environment is poorly illuminated and images contain less content information. Our method encodes both RGB and depth images, enabling the fusion of details and 3D information, resulting in a more robust scene representation that can effectively handle feature degradation scenarios.

**Results on the MSLS dataset.** To validate the generalizability of our method, we test on four other cities using the pre-trained model on Melbourne. The generalization results are shown in Tab. III. It can be seen that our method achieves compeletive performance. Compared with NetVLAD, our method achieves an average improvement of about 11% on Recall@1, and compared with SeqNet,




(a) Overcast Autumn (b) Dawn Winter

(c) Night Autumn (d) Overcast Summer

Fig. 4. Performance comparison of Recall with state-of-the-art methods for opposing viewpoints and different environmental conditions.

an average improvement of about 6% on Recall@1. It indicates that our method has good generalization and is more robust to appearance variations. In addition, SeqNet and PatchNetVLAD also achieve good results. SeqNet aggregates temporal information using temporal convolution and uses hierarchical matching to address appearance changes and perceptual aliasing. PatchNetVLAD can also perform well due to the large number of test sets in this dataset (an average of around 10,000) and discriminative objects in images.

### C. Ablation Studies

**Ablation study on the proposed components.** To verify the validity of each module in our proposed approach, we perform ablation analysis on two sequences under various environment of the Oxford RobotCar datasets, and the experimental results are shown in Tab. IV. From the first and second rows of the table, it shows that our proposed multi-modal approach improves Recall by an average of around 3% in daytime scenes, compared to the single RGB modality. This confirms that depth modality is fully effective in scenes with poor lighting and shadows. The second and third rows of the table indicate that our proposed SFFM module effectively fuse RGB and depth features to capture discriminative representation. The third and forth rows of the table demonstrate that the approach with the TC-GAT module outperforms the approach with the regular GAT,

TABLE III

GENERALIZABILITY ABILITY OF RECALL@N WITH STATE-OF-THE-ART METHODS ON MSLS DATASETS. THE BEST RESULTS AND SUB-OPTIMAL RESULTS ARE HIGHLIGHTED IN BOLD UNDERLINED AND BOLD, RESPECTIVELY.

| Method | Amman | Copenhagen | Boston | San Francisco |
|---|---|---|---|---|
| | Recall@1/5/20 | Recall@1/5/20 | Recall@1/5/20 | Recall@1/5/20 |
| AP-GeM[11] | 0.122/0.203/0.334 | 0.417/0.558/0.666 | 0.207/0.292/0.370 | 0.317/0.458/0.582 |
| NetVLAD[8] | 0.217/0.286/0.328 | 0.449/0.590/0.693 | 0.244/0.327/0.414 | 0.458/0.595/0.706 |
| PatchNetVLAD[13] | **0.254/0.339/0.396** | 0.460/0.611/0.712 | **0.265/0.362/0.442** | 0.463/0.620/0.728 |
| SeqNet[26] | 0.221/0.283/0.341 | **0.550/0.664/0.739** | 0.275/0.334/0.417 | **0.502**/0.604/0.719 |
| MR-NetVLD[32] | 0.229/0.310/0.366 | 0.479/0.618/0.714 | 0.249/0.346/0.440 | 0.488/**0.626/0.732** |
| Ours | **0.287/0.345/0.419** | **0.547/0.674/0.755** | **0.353/0.470/0.569** | **0.623/0.751/0.829** |

TABLE IV

ABLATION EXPERIMENTON ON OXFORD ROBOTCAR DATASETS.

| RGB | Depth | SFFM | GAT | TC-GAT | Winter-night | Summer-overcast |
|---|---|---|---|---|---|---|
| | | | | | Recall@1/5/20 | Recall@1/5/20 |
| ✓ | × | × | × | × | 0.681/0.864/0.944 | 0.843/0.925/0.963 |
| ✓ | ✓ | × | × | × | 0.701/0.883/0.953 | 0.920/0.965/0.980 |
| ✓ | ✓ | ✓ | × | × | 0.866/0.962/0.984 | 0.927/0.965/0.980 |
| ✓ | ✓ | ✓ | ✓ | × | 0.867/0.964/0.985 | 0.927/0.968/0.980 |
| ✓ | ✓ | ✓ | × | ✓ | **0.886/0.974/0.989** | **0.930/0.968/0.982** |

indicating the validity of temporal constraints. Overall, with the help of the depth and TC-GAT module, our algorithm achieves the best performance, demonstrating the effectiveness of each part of our proposed method.

**Quality of Depth Estimation.** In order to verify the effect of depth image quality on our proposed algorithm performance, we conduct experiments on the synthetic Synthia[45] dataset and select the front-view DAWN and NIGHT images from the Sequence-02 datasets. In the experiments, we test the network with both the depth ground truth provided by the Synthia and the estimated depth obtained by the monodepth2 algorithm. The quantitative and qualitative results are shown in Tab. V and Fig. 5, respectively. By analyzing the quantitative results, it is observed that the method using the ground truth depth has slightly better performance than the method using the estimated depth, with an average improvement of 1% in Recall@1. This suggests that the estimated depth map is acceptable. As can be seen from Fig. 5, depth ground truth and estimated depth is similar, our method can learn right distance information from estimated depth.

TABLE V

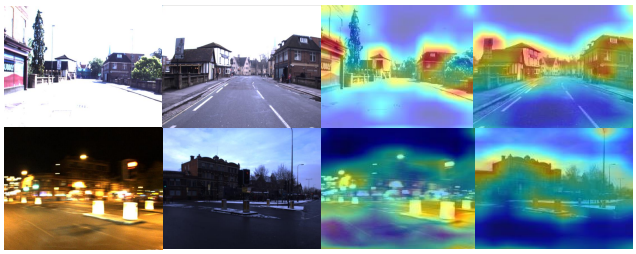
QUANTITATIVE COMPARISON OF RECALL@N WITH DEPTH GROUND TRUTH AND ESTIMATED DEPTH ON SYNTHIA DATASET.

| Method | Synthia |
|---|---|
| | Recall@1/5/10/20 |
| Ground truth | 0.171/0.705/0.884/0.934 |
| Estimated depth | 0.163/0.673/0.877/0.932 |

(a) Input image (b) Depth ground truth (c) Estimated depth

Fig. 5. Qualitative results with depth ground truth and estimated depth using Monodepth2 on Synthia dataset.

**Visualization of our proposed method.** The visualization in Fig. 6 provides a more intuitive understanding of the features learned by our proposed method. We take nighttime and summer images from Oxford RobotCar datasets with intense lighting changes as an example. The first and second columns present the query images and retrieved results. The third and fourth columns display the visual results using grad-cam++[46]. The super-imposed heat maps illustrate the divergent importance of visual elements to image representation. It can be observed that our proposed method can accurately capture the stable structural cues and pay more attention to them such as buildings and traffic lights which remains unchanged under appearance and viewpoint changes. The attention on invariant visual elements greatly ensures the robustness of our proposed algorithm. In addition, from the last two columns, we find that the query image and retrieved result have consistent focus, which is beneficial for matching at the descriptor level.



(a) Query (b) Retrieved (c) Cam-Query (d) Cam-Retrieved

Fig. 6. The visualization of our approach. Our method pays more attention to the geometric structure information.

## V. CONCLUSION

In this paper, we propose a multi-modal visual place recognition algorithm aimed at addressing dynamic environmental changes. Specifically, the multi-modal feature extraction and shared feature fusion framework are designed to extract and integrate RGB and depth information. To avoid perceptual aliasing, we also propose to employ time-constrained graph attention aggregation to propagate node information across time and space levels. Experimental results demonstrate that the our method exhibits better performance in challenging scenarios. In the future, we will devote into exploring modality-specific features and modality-shared features in different modal spaces to reduce redundant features and obtain more precise place representations.

## REFERENCES


[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[2] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt, "Scalable 6-dof localization on mobile devices," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*. Springer, 2014, pp. 268–283.

[3] Y. Zhou, G. Wan, S. Hou, L. Yu, G. Wang, X. Rui, and S. Song, "Da4ad: End-to-end deep attention-based visual localization for autonomous driving," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*. Springer, 2020, pp. 271–289.

[4] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *2012 IEEE international conference on robotics and automation*. IEEE, 2012, pp. 1643–1649.

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

[6] S. Garg, T. Fischer, and M. Milford, "Where is your place, visual place recognition?" *arXiv preprint arXiv:2103.06443*, 2021.

[7] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, and E. Romera, "Towards life-long visual localization using an efficient matching of binary sequences from images," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 6328–6335.

[8] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[9] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 3223–3230.

[10] Y. Zhu, J. Wang, L. Xie, and L. Zheng, "Attention-based pyramid aggregation network for visual place recognition," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 99–107.

[11] J. Revaud, J. Almazán, R. S. Rezende, and C. R. d. Souza, "Learning with average precision: Training image retrieval with a listwise loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5107–5116.

[12] D. Liu, Y. Cui, L. Yan, C. Mousas, B. Yang, and Y. Chen, "Densernet: Weakly supervised visual localization using multi-scale feature aggregation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 7, 2021, pp. 6101–6109.

[13] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "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*, 2021, pp. 14 141–14 152.

[14] Y. Cai, J. Zhao, J. Cui, F. Zhang, T. Feng, and C. Ye, "Patch-netvlad+: Learned patch descriptor and weighted matching strategy for place recognition," in *2022 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 2022, pp. 1–8.sssssssssss

[15] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable cnn for joint description and detection of local features," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2019, pp. 8092–8101.

[16] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 716–12 725.

[17] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 726–743.

[18] Y. Shen, R. Wang, W. Zuo, and N. Zheng, "Tcl: Tightly coupled learning strategy for weakly supervised hierarchical place recognition," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2684–2691, 2022.

[19] J. Ning, Y. Zhang, X. Zhao, S. Coleman, K. Li, and D. Kerr, "Samloc: Structure-aware constraints with multi-task distillation for long-term visual localization," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 719–11 725.

[20] C. Pinard, L. Chevalley, A. Manzanera, and D. Filliat, "Learning structure-from-motion from motion," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[21] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2002–2011.

[22] S. Garg, M. Babu, T. Dharmasiri, S. Hausler, N. Suenderhauf, S. Kumar, T. Drummond, and M. Milford, "Look no deeper: Recognizing places from opposing viewpoints under varying scene appearance using single-view depth estimation," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4916–4923.

[23] N. Piasco, D. Sidibé, V. Gouet-Brunet, and C. Demonceaux, "Improving image description with auxiliary modality for visual localization in challenging conditions," *International Journal of Computer Vision*, vol. 129, pp. 185–202, 2021.

[24] H. Hu, Z. Qiao, M. Cheng, Z. Liu, and H. Wang, "Dasgil: Domain adaptation for semantic and geometric-aware image-based localization," *IEEE Transactions on Image Processing*, vol. 30, pp. 1342–1353, 2020.

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

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

[27] S. Garg, M. Vankadari, and M. Milford, "Seqmatchnet: Contrastive learning with sequence matching for place recognition & relocalization," in *Conference on Robot Learning*. PMLR, 2022, pp. 429–443.

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

[29] C. Qin, Y. Zhang, Y. Liu, S. Coleman, H. Du, and D. Kerr, "A visual place recognition approach using learnable feature map filtering and graph attention networks," *Neurocomputing*, vol. 457, pp. 277–292, 2021.

[30] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[31] F. Warburg, S. Hauberg, M. Lopez-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2626–2635.

[32] A. Khaliq, M. Milford, and S. Garg, "Multires-netvlad: Augmenting place recognition training with low-resolution imagery," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3882–3889, 2022.

[33] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 2. IEEE, 2005, pp. 729–734.

[34] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio *et al.*, "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48 550, 2017.

[35] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 486–504.

[36] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[38] T. Tieleman, G. Hinton *et al.*, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[39] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks

for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[44] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4794–4803.

[45] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3234–3243.

[46] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.