FE-Fusion-VPR: Attention-Based Multi-Scale Network Architecture for Visual Place Recognition by Fusing Frames and Events

Kuanxu Hou, Delei Kong, Junjie Jiang, Hao Zhuang, Xinjie Huang, and Zheng Fang, Member, IEEE

Abstract—Traditional visual place recognition (VPR), usually using standard cameras, is easy to fail due to glare or high-speed motion. By contrast, event cameras have the advantages of low latency, high temporal resolution, and high dynamic range, which can deal with the above issues. Nevertheless, event cameras are prone to failure in motionless scenes, while standard cameras can still provide appearance information in this case. Thus, exploiting the complementarity of standard cameras and event cameras can effectively improve the performance of VPR algorithms. In the paper, we propose FE-Fusion-VPR, an attention-based multi-scale network architecture for VPR by fusing frames and events. First, the intensity frame and event volume are fed into the two-stream feature extraction network for shallow feature fusion. Next, the three-scale features are obtained through the multi-scale fusion network and aggregated into three sub-descriptors using the VLAD layer. Finally, the weight of each sub-descriptor is learned through the descriptor re-weighting network to obtain the final refined descriptor. Experimental results show that our FE-Fusion-VPR outperforms existing frame-based, event-based and fusion-based VPR methods in most cases on Brisbane-Event-VPR and DDD20 datasets. In a word, compared to the previous works, our FE-Fusion-VPR achieves new state-of-the-art (SOTA) VPR performance on Brisbane-Event-VPR and DDD20 datasets by fusing frames and events.

Index Terms—Visual place recognition, event camera, attention mechanism, multi-scale network, visual sensor fusion.

I. INTRODUCTION

VISUAL place recognition (VPR) [1], [2], [3], [4] is a vital sub-problem in the autonomous navigation of mobile robots, which has attracted the attention of many researchers in recent years. VPR aims to help a robot determine whether it locates in a previously visited place. Specifically, there is an existing database about the environment, which stores visual data (such as frames) for various places in the environment. Now given a query data, we expect to obtain its location information by finding database data captured at the same (or close) location as the query data. In a word, VPR can assist mobile robots or autonomous unmanned systems in localization and loop closure detection (LCD) in GPS-denied environments.

Currently, since standard frame cameras could provide rich appearance information of everyday scenes, existing frame-based VPR methods could achieve good performance in those scenarios. However, standard cameras suffer from low frame rate, motion blur, and sensitivity to illumination changes, which makes traditional VPR methods difficult to handle challenging scenes (e.g., high speed and high dynamic range). Event cameras [5], [6], [7], [8], which are neuromorphic vision sensors, record microsecond-level pixel-wise brightness changes and offer significant advantages (e.g., low latency, rich motion information and high dynamic range). Nevertheless, event cameras lack appearance (texture) information in some cases (e.g., still or low-speed scenes). Therefore, the above two kinds of vision sensors are complementary. As shown in Fig. 1, we present two representative examples (glare and motionless) respectively, and these challenging situations can be solved by visual sensor fusion.

Inspired by their complementarity, we consider combining standard frame cameras and event cameras to lift the limitations faced by single vision sensors in large-scale place recognition problems, thereby improving the performance of the VPR pipeline in challenging scenarios. Recently, there have been some works trying to combine frames and events for some machine vision tasks and achieve excellent results [9], [10].

Fig. 1. Illustration of advantages of the proposed FE-Fusion-VPR. As can be seen, both frames and events have challenging scenarios that are difficult to deal with alone, so combining frames and events can effectively deal with complex scenarios and improve the performance of VPR pipeline.
However, combining frames and events for VPR is not very straightforward. We need to deal with several challenges: (1) How to fuse raw frame with event data? (2) How to extract multi-scale features? (3) How to fuse multiple features into a unified descriptor? Addressing the above challenges, in this letter, we propose a robust VPR method by fusing frames and events (FE-Fusion-VPR) for large-scale place recognition problems. FE-Fusion-VPR is an attention-based multi-scale deep network architecture for the mixed frame/event-based VPR. It can effectively combine the advantages of standard frame cameras and event cameras, and suppress their disadvantages, thus achieving better VPR performance than using frames or events alone. In summary, the main contributions of this letter are as follows:

- **Aiming at the VPR task**, we propose a novel two-stream feature extraction network (TSFE-Net) with convolutional attention mechanism, which can process frame and event data simultaneously and extract the specific features of each modality for hybrid feature fusion.
- We propose an attention-based multi-scale network (MSF-Net) and design a re-weighting network (DRW-Net) that can assign weights to different sub-descriptors to obtain the best descriptor representation, achieving the SOTA VPR performance.
- In addition to extensive comparison experiments and ablation studies, we also evaluate our FE-Fusion-VPR under motionless and glare conditions, which demonstrates that the two kinds of vision sensors are complementary to a certain extent and that our sensor fusion scheme can effectively improve VPR performance in the above challenging cases.

## II. RELATED WORK

### A. Frame-Based VPR Methods

Conventional VPR algorithms mainly consist of two steps: feature extraction and feature matching. In feature extraction, key low-level features need to be extracted from high-dimensional visual data for storage. These key features are generally called descriptors or representations and can be local/global or sparse/dense. Earlier frame-based VPR works focused on hand-crafted algorithms, including local feature extractors (SIFT [17], SURF [18], ORB [19]) and global feature extractors (HoG [20], GIST [21]), feature clusters (BoW [22], [23], FV [24], [25], VLAD [26], [27], [28], [29]). However, hand-crafted methods usually need to be elaborately designed and are not robust to changes in illumination, seasons, viewpoint, etc. In contrast, learning-based (especially deep learning) feature extraction algorithms can learn general features and achieve better performance than hand-crafted algorithms in large-scale image retrieval tasks [3]. For example, Arandjelovic et al. improved VLAD as a trainable pooling layer (called NetVLAD [30]) for direct integration into the CNN-based VPR framework. On this basis, Patch-NetVLAD [31] first computes the global NetVLAD descriptors to filter out some reference candidates, then uses the local patch NetVLAD descriptors for further fine-tuning matching, which obtains a higher performance than NetVLAD. MR-NetVLAD [32] augments NetVLAD with multi-resolution image pyramid encoding, resulting in rich place representations to overcome challenging scenarios (such as illumination and viewpoint changes) better. Although learning-based VPR methods have achieved good performance, they still have difficulties in glare and high-speed scenes due to the inherent shortcomings of standard cameras.

### B. Event-Based VPR Methods

Recently, using event cameras to solve VPR problems has drawn more and more attention from researchers. First, Tobias et al. proposed an event-based VPR method (Ensemble-Event-VPR) with ensembles of temporal windows [33], which reconstructs intensity frames from event streams of different temporal windows, extracts visual descriptors using NetVLAD [30], and then integrates the distance matrix of multiple descriptors for VPR. Lee et al. proposed a VPR method (EventVLAD) to recover edge images from event streams and then use NetVLAD for descriptor generation [34]. No matter whether using intensity frames or edge images, the above two methods need to transform event streams into frames. Therefore, they are still frame-based VPR methods in essence. To utilize event streams directly, we previously proposed an event-based end-to-end deep network architecture for VPR (Event-VPR) [35]. Event-VPR uses EST voxel grid representation, combines deep residual network and VLAD layer to extract visual descriptors, and adopts weakly supervised loss for training, which achieves excellent performance in multiple challenging driving datasets using events directly. However, since event cameras cannot output the appearance information of the scene directly, it is still tricky for VPR to work properly in motionless scenes. The latest work, VEFNet [36], uses a cross-modality attention module and a self-attention module for frame and events fusion based on the VGG network feature extractor. And it achieves good VPR performance on Brisbane-Event-VPR [33] dataset. Different from their work, our FE-Fusion-VPR uses two-stream visual data fusion and multi-scale feature fusion, and outperforms both frame-based and event-based SOTA methods in recognition performance.

## III. METHODOLOGY

### A. The Overall Architecture

An overview of the proposed FE-Fusion-VPR pipeline is shown in Fig. 2. Our FE-Fusion-VPR comprises a two-stream feature extraction network (TSFE-Net), a multi-scale fusion network (MSF-Net), and a descriptor re-weighting network (DRW-Net). Its backbone consists of residual blocks of ResNet34 [37]. The intensity frame and events are first fed into TSFE-Net to obtain their shallow fusion feature. Then, MSF-Net extracts the multi-scale features which are aggregated to three sub-descriptors. Next, DRW-Net can learn the weights of the sub-descriptors for getting the final refined descriptor. Finally, the final descriptor is used to match the query data with reference data. The whole network is trained in an end-to-end manner with weakly supervision. In addition, the attention layer is commonly used in our FE-Fusion-VPR network architecture, which consists of a channel attention mechanism and a spatial attention mechanism [38]. We pass the feature maps of different scales through the attention layer to obtain more sophisticated feature maps, which can effectively improve VPR performance.
B. Two-Stream Feature Extraction Network

Due to the asynchronous characteristics of events, combining them with intensity frame remains challenging, especially for learning-based methods. In Event-VPR [35], our experiments have shown that different event representations have little effect on VPR tasks. Therefore in this letter, event volumes are processed to event frames [39] directly, which are fed into our TSFE-Net together with intensity frames. As shown in Fig. 2, in order to learn multi-modal shared feature, our two-stream feature extraction network (TSFE-Net) \( f_{\text{TSFE-Net}}(\cdot) \) extracts two kinds of shallow features and fuses them. Inspired by Event-VPR [35], we use convolutional layer conv1 and residual blocks conv2_x cut from ResNet34, and attach an attention layer to each to improve the quality of shallow features. For better recognition performance, we concatenate the shallow features along the channel dimension [40]. Finally, we use a max-pooling operation and an attention layer to obtain rich and effective scene information. Our two-stream feature extractor \( f_{\text{TSFE-Net}}(\cdot) \) can be summarized as follows:

\[
X^H = f_{\text{TSFE-Net}}(F, E) = f_{EC}^F(F) \oplus f_{EC}^E(E) = X^F \oplus X^E,
\]

where \( f_{EC}^F(\cdot) \) and \( f_{EC}^E(\cdot) \) are the feature encoders processing frames and events respectively. The structure of both feature encoders we designed is Conv(\(7 \times 7, 64, /2\))-Attn(\(5 \times 5\))-MaxPool2d(\(/2\))-ResBlock0(\(3 \times 3, 64\))-ResBlock1(\(3 \times 3, 64\))-ResBlock2(\(3 \times 3, 64\))-Attn(\(5 \times 5\))-BatchNorm-ReLU. Here, the Conv is convolutional layer (\(7 \times 7\) is the kernel size, \(64\) is the number of output channels, \(/2\) is the stride, the same below), the Attn is attention layer described in [38], the MaxPool2d is max-pooling layer, the ResBlock0/1/2 is the residual block of ResNet34, the BatchNorm is batch normalization layer, and the ReLU is rectified linear unit. \( F \) and \( E \) are intensity frame and event volume, \( X^F \) and \( X^E \) are the primary features of intensity frame and event volume respectively, \( \oplus \) is the concatenation operation, and \( X^H \) is the hybrid feature after fusing.

C. Multi-Scale Fusion Network

Many works [41], [42] have demonstrated that mid-level visual features exhibit robustness to appearance changes, while high-level visual features are robust to viewpoint changes in VPR tasks. Therefore, the accuracy of VPR can be improved theoretically by using multi-scale network architecture. Here, our idea is inspired by feature pyramid network (FPN) [43], which is a typical multi-scale network architecture. However, the performance of FPN degrades in some cases (such as the detection of large objects [44]). To achieve more efficient communication between different levels, our proposed multi-scale fusion network (MSF-Net) \( f_{\text{MSF-Net}}(\cdot) \) fuses different-scale features in the following way (as shown in Fig. 2). First, our backbone network performs bottom-up feature extraction, which contains three stages of residual structure. The output features of each stage are \( S_1, S_2, S_3 \) respectively. In particular, the residual structure of the first stage is contained in \( f_{\text{TSFE-Net}}(\cdot) \). Thus, we directly adopt the max-pooling layer \( f_{MP}(\cdot) \) and the attention layer \( f_{Attn}(\cdot) \) for the output fusion feature \( X^H \) of the TSFE-Net to obtain the feature \( S_1 \) of the first stage. Then, features \( S_2 \) and \( S_3 \) are extracted through the remaining two stages of residual structure (residual blocks conv4_x and conv5_x), which are expressed as follows:

\[
S_1 = f_{Attn}(f_{MP}(X^H)) ,
S_2 = f_{Attn}(f_{Res,2}(S_1)) ,
S_3 = f_{Res,3}(S_2).
\]

where the channels of features \( S_1, S_2, S_3 \) are 128, 256, 512 respectively. Next, different from FPN, to make more efficient use of the multi-scale information of each stage, we adopt concatenation to perform stage-wise fusion based on the backbone network. Specifically, for each stage of the backbone network, we add the branch network as a lateral connection (passway) to fuse the features of the two adjacent stages. The branch network generally includes a convolutional layer (convolution kernel is \(1 \times 1\)), upsampling layer (\(x2\) Up), batch-normalization layer (BN), and ReLU activation layer. By adjusting the channels and spatial resolution of features, we can obtain features \( M_1, M_2, M_3 \) with 256 channels and \(8 \times 8, 16 \times 16\) and \(32 \times 32\) spatial resolutions respectively:

\[
M_1, M_2, M_3 = f_{\text{MSF-Net}}(X^H).
\]
TABLE I
SCENARIOS, SEQUENCES, TRAINING AND TESTING SETS OF BRISBANE-EVENT-VPR [33] AND DDD20 [11] DATASETS USED IN OUR EXPERIMENTS

| Datasets                              | All Scenarios and Sequences                                                                 | Experiments | Training Sets (Database / Query) | Testing Sets (Database / Query) |
|---------------------------------------|---------------------------------------------------------------------------------------------|-------------|---------------------------------|---------------------------------|
| Brisbane-Event-VPR [33]               | sunrise (ss) {2020-04-29-06-20-23}, morning (mn) {2020-04-28-09-14-11},                 | 1           | (dt & mn) [4712] / ss [2620]    | ss1 [1768] / ss2 [1768]         |
|                                       | daytime (dt) {2020-04-24-15-12-03}, sunset (ss1, ss2) {2020-04-21-17-03-03, 2020-04-22-17-24-21} | 2           | (ss2 & dt) [4002] / ss [2478]   | ss1 [2492] / ss [2492]          |
| DDD20 [11]                           | street {rec1501983083, rec1502648048, rec1502355857},                                  | 3           | (ss2 & dt) [4002] / ss [2478]   | ss [2478] / mn [2478]           |
|                                       | freeways {rec1500922481, rec1501191354, rec1501268968}                                 | 4           | (ss2 & mn) [4246] / ss [2478]   | ss1 [2181] / dt [2181]          |
|                                       |                                                                                            | 5           | **83 [1146] / **57 [1001]       | **83 [1146] / **48 [1146]       |
|                                       |                                                                                            | 6           | **81 [5177] / **54 [5583]       | **81 [5177] / **68 [5177]       |

The text in "([)" indicates the sequence name, and the text in "([])" indicates the number of samples. "&" means to merge sequences of two scenes. "+" means to distinguish between database and query. "*" means omitted.

concatenate them to obtain the primary multi-scale descriptor $D$ of which dimension is $3 \times N$:

$$D = \| (D_1) \| = \| (f_{VLAD}(M_i)) \|,$$ (4)

where $\| (\cdot) \|$ represents concatenation operation. Since the multi-scale fusion features contain different levels of visual information, they can provide robust feature information for the DRW-Net to improve our network’s overall performance.

D. Descriptor Re-Weighting Network

In MSF-Net, we have obtained the primary multi-scale descriptor aggregated by three different-scale features. In order to represent the scenes better, we need to redesign the multi-scale descriptor to obtain a compact global descriptor. Therefore, we propose a descriptor re-weighting network (DRW-Net) $f_{DRW-Net}^{\prime}$, as shown in Fig. 2, to obtain a robust global descriptor:

$$D^{\prime} = f_{DRW-Net}^{\prime}(D).$$ (5)

For the multi-scale descriptor $D$, we calculate the average $f_{avg}^{\prime}$ and maximum $f_{max}^{\prime}$ of each sub-descriptor respectively:

$$G_{avg} = f_{avg}^{\prime}(D) = \| (\frac{1}{N} \sum_{m=1}^{N} D_i(m)) \|,$$

$$G_{max} = f_{max}^{\prime}(D) = \| (\max_m (D_i(m))) \|,$$ (6)

where $G_{avg}$ and $G_{max}$ are the channel-wise global representations of sub-descriptors, $N$ denotes the index of the sub-descriptor, $m$ is the spatial coordinates of the sub-descriptors. Then, we append two fully connected (FC) layers respectively to learn two kinds of weights of the sub-descriptors $w_{avg}$ and $w_{max}$, and add the above two weights to obtain the final weights $w$ of sub-descriptors through a soft-max layer $f_{SM}^{\prime}$:

$$w_{avg} = f_{FC,2}^{\prime}(f_{ReLU}(f_{FC,1}^{\prime}(G_{max}))),$$

$$w_{max} = f_{FC,2}^{\prime}(f_{ReLU}(f_{FC,1}^{\prime}(G_{max}))),$$

$$w = f_{SM}^{\prime}(w_{avg} + w_{max}),$$ (7)

where $f_{FC,1}^{\prime}(\cdot), f_{FC,1}^{\prime}(\cdot) : \mathbb{R}^{3 \times 1} \rightarrow \mathbb{R}^{M \times 1}$ and $f_{FC,2}^{\prime}(\cdot), f_{FC,2}^{\prime}(\cdot) : \mathbb{R}^{M \times 1} \rightarrow \mathbb{R}^{1 \times 1}$ are two kinds of fully connected (FC) layers respectively. $M = 12$ is the transformation length of global representations in hidden layers. The soft-max operation $f_{SM}^{\prime}(\cdot)$ makes the weights of these three descriptors mutually balanced, and the sum of their weights is 1. Finally, we multiply the multi-scale descriptors with weights and sum multi-scale descriptors channel by channel, which is expressed as follows:

$$D^{\prime} = D \otimes w = D_1 \otimes w_1 + D_2 \otimes w_2 + D_3 \times w_3$$

$$= D_1^{\prime} + D_2^{\prime} + D_3^{\prime},$$ (8)

Thus, we obtain the final multi-scale weighted aggregate descriptor $D^{\prime}$, which makes our FE-Fusion-VPR more robust than methods using the primary descriptor $D$. (See our ablation studies on DRW-Net in Table III.)

IV. EXPERIMENTS

A. Experimental Setup

1) Dataset Selection: To evaluate the performance of the proposed method, we conduct experiments on Brisbane-Event-VPR [33] and DDD20 [11] datasets. Brisbane-Event-VPR [33] consists of data recorded using a DAVIS camera together with GPS. It includes six traverses of the same path at different time of the day, including sunrise, morning, daytime, sunset, and night. We discard the night sequence since the frame rate of intensity frames is too low for frame-to-event visual fusion. DDD20 [11] is the event camera end-to-end driving dataset under various lighting conditions. We selected six sequences of two urban scenes, of which two sequences have glare illumination, and three sequences consist of highways. We use the timestamps of the intensity frames to get the event volumes. The intensity frame’s interval of Brisbane-Event-VPR and DDD20 datasets that we select are approximately 0.25 s and 0.2 s, respectively. We use the same selection for corresponding event volumes. The time window of the event volume is 25 ms and 20 ms for Brisbane-Event-VPR and DDD20 datasets, respectively.

2) Parameters in Training: We train our FE-Fusion-VPR network with weakly supervision using a triplet ranking loss [35]. Except for the optimizer and learning rate, we use the same parameters in all experiments for a fair comparison. Where, the number of cluster centers (vocabulary size) $K = 128$, margin $\varepsilon = 0.1$. When training, for a query data, 1 positive (within potential positive distance threshold $\lambda = 25$ m) and 10 negatives (far away than randomly negative distance threshold $\delta = 75$ m) are selected for triplet metric learning. When testing, the query data is deemed correctly localized if at least one of the top $N$ retrieved database data is within the true positive geographical distance threshold $\phi = 75$ m from the ground truth position of the query.
TABLE II
COMPARISON OF OUR FE-FUSION-VPR AGAINST FRAME-BASED [30], [31], [32] AND EVENT-BASED [33], [35], [36] SOTA VPR METHODS ON BRISBANE-EVENT-VPR [33] AND DDD20 [11] DATASETS WITH THE BEST RESULT BOLDED

| Modal | VPR Methods | Brisbane-Event-VPR [33] Dataset | DDD20 [11] Dataset |
|-------|-------------|---------------------------------|--------------------|
| Frame | NetVLAD [30] | Recall@1 (%) | F1-max  |
|       | 94.34, 97.29 | 90.61, 96.35 | 86.97, 94.11 | 77.40, 88.49 | 87.33 |
|       | 0.9709 | 0.9516 | 0.9330 | 0.8762 | | |
|       | Patch-NetVLAD [31] | 96.95, 99.15 | 92.68, 97.79 | 85.27, 91.25 | 81.96, 90.60 | 89.41 |
|       | 0.9845 | 0.9638 | 0.9434 | 0.8944 | | |
|       | MR-NetVLAD [32] | 94.23, 97.06 | 91.46, 95.43 | 87.20, 93.06 | 79.64, 88.72 | 88.14 |
|       | 0.9733 | 0.9582 | 0.9343 | 0.8929 | | |
| Event | Ensemble-Event-VPR [33] | 87.33, 89.70 | 88.21, 84.21 | 84.42, 78.33 | 74.62, 72.71 | 62.05 |
|       | 0.9345 | 0.7246 | 0.7550 | 0.6319 | | |
|       | Event-VPR [35] (Ours) | 84.79, 93.83 | 65.65, 82.52 | 66.67, 84.20 | 44.56, 66.10 | 65.41 |
|       | 0.9236 | 0.7984 | 0.7946 | 0.6703 | | |
| Fusion | VEFNet [36] | 94.86, 97.34 | 90.98, 96.99 | 90.38, 89.23 | 76.65, 87.51 | 95.52 |
|       | 0.9733 | 0.9486 | 0.8930 | 0.8278 | | |
|       | FE-Fusion-VPR (Ours) | 95.64, 98.36 | 93.58, 97.19 | 87.41, 93.87 | 86.15, 93.03 | 90.70 |
|       | 0.9780 | 0.9671 | 0.9310 | 0.9247 | | |

The specific meaning of training / testing set abbreviations is shown in Table I.

TABLE III
ABLATION STUDIES ON THE IMPACT OF TSFE-NET, MSF-NET, AND DRW-NET ON THE PERFORMANCE OF FE-FUSION-VPR ON BRISBANE-EVENT-VPR [33] AND DDD20 [11] DATASET WITH THE BEST RESULTS BOLDED

| Ablation Studies | Settings | Recall@1 (%) | Brisbane-Event-VPR [33] Dataset | DDD20 [11] Dataset |
|------------------|----------|--------------|---------------------------------|--------------------|
|                  |          | ss1 / ss2 | ss1 / sr | ss1 / mn | ss1 / dt | **83 / **84 | **81 / **85 | **83 / **84 | **85 / **86 |
| TSFE-Net          | Only Frame Encoder | 85.63 | 78.38 | 76.23 | 67.98 | 58.02 | 40.59 |
|                  | Only Event Encoder | 93.16 | 63.74 | 82.16 | 33.90 | 49.11 | 37.14 |
|                  | Frame & Event Encoder | 93.64 | 94.38 | 87.41 | 86.15 | 72.77 | 54.05 |
| MSF-Net           | Only 8×8 (X(*) | 86.82 | 91.33 | 76.33 | 83.59 | 73.84 | 68.03 | 60.32 | 54.39 | 42.20 | 37.34 |
|                  | Only 16×16 (X(*) | 86.03 | 95.53 | 73.25 | 78.15 | 82.01 | 72.22 | 65.56 | 65.08 | 42.15 | 48.45 |
|                  | Only 32×32 (X(*) | 90.03 | 93.53 | 84.16 | 86.08 | 84.93 | 83.23 | 82.02 | 76.01 | 64.34 | 69.21 | 46.09 | 48.33 |
|                  | Multi-Scale (X(*) | 94.06 | 95.64 | 90.89 | 93.58 | 85.85 | 87.41 | 80.09 | 86.15 | 65.87 | 72.77 | 69.98 | 54.05 |
| DRW-Net           | Concatenation in Length | 91.97 | 68.51 | 75.30 | 73.21 | 59.22 | 40.67 |
|                  | With Self-Attention Module | 87.23 | 89.77 | 81.89 | 79.45 | 59.17 | 44.50 |
|                  | With Re-Weighting Layer | 95.64 | 93.58 | 87.41 | 86.15 | 72.77 | 54.05 |

3) Evaluation Metrics: We use PR curve and Recall@N to evaluate the experimental results, which you can refer to Event-VPR [35] for the specific description of the metrics. For a more comprehensive comparison, we calculate F1-max for all the VPR methods, which is as follows:

$$F_{\text{1}, \text{max}} = \max_k \left( \frac{2 \times P_k \times R_k}{P_k + R_k} \right), \quad (9)$$

where $P_k$ and $R_k$ are the $k$-th precision and recall in PR curves respectively. In addition, we also present the retrieval success-rate maps to show the performance of our algorithm more intuitively.

B. Comparison Against SOTA Methods

In this section, we present the evaluation details, including frame-based and event-based SOTA VPR methods. And then, we analyze the reasons why our FE-Fusion-VPR has the best performance.

1) Comparison Against Frame-Based VPR Methods: We compare our FE-Fusion-VPR against frame-based SOTA algorithms (Patch-NetVLAD [31] and MR-NetVLAD [32]), and the experimental results are shown in Table II, Figs. 4, 5, and 6. The results show that our method outperforms the above two algorithms in most cases. Table II shows that the Recall@1 of FE-Fusion-VPR is 1.29% and 2.56% higher than Patch-NetVLAD and MR-NetVLAD on average on the Brisbane-Event-VPR dataset. On the DDD20 dataset, the advantages of our FE-Fusion-VPR are more obvious. The Recall@1 of FE-Fusion-VPR is 23.10% and 15.89% higher than that of the above two algorithms on average. The reason is that there are sequences with obvious differences in intensity appearance in the Brisbane-Event-VPR dataset, and some glare scenarios in...
Fig. 4. PR curves of NetVLAD [30], Patch-NetVLAD [31], MR-NetVLAD [32], Ensemble-Event-VPR [33], Event-VPR (ours) [35], VEFNet [36] and FE-Fusion-VPR (ours) on Brisbane-Event-VPR [33] and DDD20 [11] datasets. The first four subgraphs are on Brisbane-Event-VPR [33], and the last two subgraphs are on DDD20 [11]. Our FE-Fusion-VPR (red) performs better than SOTA methods under most scenes.

Fig. 5. Recall@N curves of NetVLAD [30], Patch-NetVLAD [31], MR-NetVLAD [32], Ensemble-Event-VPR [33], Event-VPR (ours) [35], VEFNet [36] and FE-Fusion-VPR (ours) on Brisbane-Event-VPR [33] and DDD20 [11] datasets. The first four subgraphs are on Brisbane-Event-VPR [33], and the last two subgraphs are on DDD20 [11]. Our FE-Fusion-VPR (red) is superior in most scenes. The “N” in the horizontal axis means “number of top database candidates.”

Fig. 6. Retrieval success-rate maps of MR-NetVLAD [32] and FE-Fusion-VPR (ours) on Brisbane-Event-VPR [33] and DDD20 [11] datasets. The first four subgraphs are on Brisbane-Event-VPR [33], and the last two subgraphs are on DDD20 [11]. Top: MR-NetVLAD [32], bottom: FE-Fusion-VPR (ours).

the DDD20 dataset, which will limit the performance of frame-based VPR methods. However, event cameras hardly affected by illumination changes can significantly improve the performance of the VPR algorithms. Besides, our MSF-Net and DRW-Net can improve the VPR performance due to the multi-scale feature fusion, especially in the DDD20 dataset.

2) Comparison Against Event-Based VPR Methods: As shown in Table II, Figs. 4, and 5, our FE-Fusion-VPR is much more robust than the pure event-based SOTA methods (Ensemble-Event-VPR [33] and Event-VPR [35]). On the Brisbane-Event-VPR dataset, the Recall@1 of FE-Fusion-VPR is 25.29% and 28.65% higher than Event-VPR and Ensemble-Event-VPR on average. On the DDD20 dataset, Recall@1 of FE-Fusion-VPR increases by an average of 37.20% and 43.48% over Event-VPR and Ensemble-Event-VPR. The accuracy of our algorithm is much better than the above two methods, illustrating the importance of information from the intensity frame to improve the performance of VPR networks. As it is known to all, event cameras can hardly capture information at low-speed intersections and highways with sparse texture, while standard cameras can capture more background information when the illumination is suitable. Therefore, our FE-Fusion-VPR can achieve higher SOTA VPR performance by combining the advantages of both sensors. Our descriptors can remain rich and vital information by using MSF-Net and DRW-Net.

3) Comparison Against VEFNet: We train and test VEFNet [36] on the two datasets with our settings. Experimental results show that our FE-Fusion-VPR outperforms the VEFNet [36] in all testing sets. Table II shows that the Recall@1 of our FE-Fusion-VPR is 5.18% higher than that of VEFNet on the Brisbane-Event-VPR dataset on average. On the DDD20 dataset, the advantage of our pipeline is more obvious. Our algorithm is 21.20% higher than VEFNet on average. From the PR curves in Fig. 4 and the Recall@N curves in Fig. 5, it can be seen that our FE-Fusion-VPR is better than VEFNet in most cases. We also reproduce numbers on
Brisbane-Event-VPR using the four testing sets (ss2/ss1, ss2/ssr, ss2/mn, and ss2/dt) and geographical threshold $\phi = 15$ m from VEFNet [36]. Our Recall@1 are 81.00%, 72.04%, 68.36%, and 54.40% respectively, and are on average 5.15% higher than VEFNet.

C. Ablation Studies

In this section, we explore the impact of TSFE-Net, MSF-Net, and DRW-Net on the performance of FE-Fusion-VPR on Brisbane-Event-VPR [33] and DDD20 [11] datasets.

1) Impact of TSFE-Net: The experimental results in Table III demonstrate that using a single type of sensor data leads to severe performance degradation on the Brisbane-Event-VPR dataset. Moreover, the SOTA performance of our FE-Fusion-VPR is attributed to using the two vision sensors simultaneously rather than a single type of visual sensor.

2) Impact of MSF-Net: The experimental results in Table III show that, in most cases, our multi-scale FE-Fusion-VPR can achieve better performance than networks using single-scale features (whatever with/without adding attention layers). Since features at different scales focus on information in different regions, using mid-level or high-level features alone is unreliable. Therefore, multi-scale fusion can achieve higher VPR performance. In addition, by adding attention layers at appropriate locations throughout the whole network, the performance of FE-Fusion-VPR can be further improved.

3) Impact of DRW-Net: In this experiment, we remove the DRW-Net and use the original VLAD layer. Before the features are input into VLAD layer, we flatten the three different scale feature maps $\{M_1, M_2, M_3\}$ (the dimensions are $(D, W_i, H_i), i \in \{1, 2, 3\}$) respectively output by MSF-Net into three descriptors $\{\tilde{D}_1, \tilde{D}_2, \tilde{D}_3\}$ with dimensions of $(D, m_i), m_i = H_i \times W_i$, and then we concatenate them along the length dimension:

$$\tilde{D} = \bigoplus_{i=1}^{3} \left( \tilde{D}_i \right) = \bigoplus_{i=1}^{3} \left( f_{\text{flatten}}(M_i) \right),$$

where the dimension of the descriptor $\tilde{D}$ is $(D, m), m = m_1 + m_2 + m_3$. We also replace our DRW-Net with the self-attention module in VEFNet [36]. The results in Table III show that our DRW-Net outperforms methods directly using original VLAD layers or self-attention module for multi-feature fusion in all testing sets. DRW-Net assigns the weight of each sub-descriptor through auto-learning, which can fully use each descriptor’s significant information, so that the final multi-scale descriptor has robust representation ability.

D. Evaluation Under Motionless and Glare Conditions

In this section, we explore the VPR performance of our FE-Fusion-VPR under motionless and glare conditions. We select motionless and glare data samples from the queries in testing sets of 6 experiments. The experimental results are shown in Table IV. Under the motionless condition, due to the event camera only triggering rare events, our FE-Fusion-VPR performs similarly to the frame-only pipeline when the event-only pipeline fails. In this case, the intensity frames are complementary to the event volumes. Under the glare condition, there are two different cases: direct sunlight glare (in sunset2, sunrise, morning, daytime, and **48) and tunnel glare (in **68). Significantly, standard cameras and event cameras are affected by direct sunlight glare similarly. However, standard cameras are more affected by tunnel glare than event cameras. But our FE-Fusion-VPR achieves higher performance in both direct sunlight glare and tunnel glare conditions than the frame-only pipeline or event-only pipeline. In this case, the event volumes are complementary to the intensity frames. In summary, the two kinds of vision sensors are complementary to a certain extent. Fig. 7 shows specific examples of motionless and two kinds of glare.

V. Conclusion

In this letter, we analyzed the limitation of VPR methods using a frame camera or event camera alone. On that basis, we proposed an attention-based multi-scale network architecture combining frames and events for VPR (named FE-Fusion-VPR) to achieve robust performance in challenging environments. The two key ideas of FE-Fusion-VPR are as follows: First, we achieve visual data fusion based on intensity frames and event volumes. Second, we complete feature fusion based on a multi-scale network and descriptor re-weighting network, which is validated to be effective in our ablation studies.
with existing frame-based and event-based SOTA methods, our FE-Fusion VPR achieves higher performance, especially in scenes with motionless and glare conditions. Our VPR pipeline is not limited to DAVIS cameras but can be naturally extended to sensor fusion devices. In the future, we will try to lightweight and accelerate our algorithm for deployment to autonomous vehicles or mini-UAVs. Furthermore, we will also try to realize a deep spiking VPR network architecture [45] for high energy efficiency inference.

REFERENCES

[1] S. Lowry et al., “Visual place recognition: A survey,” IEEE Trans. Robot., vol. 32, no. 1, pp. 1–19, Feb. 2016.
[2] X. Zhang, L. Wang, and Y. Su, “Visual place recognition: A survey from deep learning perspective,” Pattern Recognit., vol. 113, 2021, Art. no. 107760.
[3] C. Masone and B. Caputo, “A survey on deep visual place recognition,” IEEE Access, vol. 9, pp. 19516–19547, 2021.
[4] S. Garg, T. Fischer, and M. Milford, “Where is your place, visual place recognition?” in Proc. 13th Int. Joint Conf. Artif. Intell. 2021, pp. 4416–4425.
[5] T. Delbrück, B. Linares-Barranco, E. Culurciello, and C. Posch, “Activity-driven, event-based vision sensors,” in Proc. IEEE Int. Symp. Circuits Syst., 2010, pp. 2426–2429.
[6] G. Gallego et al., “Event-based vision: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 1, pp. 154–180, Jan. 2022.
[7] G. Chen, H. Cao, J. Conradt, H. Tang, F. Rohrbein, and A. Knoll, “Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception,” IEEE Signal Process. Mag., vol. 37, no. 4, pp. 34–49, Jul. 2020.
[8] T.-H. Wu, C. Gong, D. Kong, S. Xu, and Q. Liu, “A novel visual object detection and distance estimation method for hsr scenes based on event camera,” in Proc. Int. Conf. Comput. Commun., 2021, pp. 636–640.
[9] D. Gehrig, H. Rebecq, G. Gallego, and D. Scaramuzza, “Asynchronous, photometric feature tracking using events and frames,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 750–765.
[10] Z. Jiang et al., “Mixed frame/event-driven fast pedestrian detection,” in Proc. IEEE Conf. Robot. Automat., 2019, pp. 8332–8338.
[11] Y. Hu, J. Binas, D. Neil, S.-C. Liu, and T. Delbrück, “DD2D0 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction,” in Proc. Int. Conf. Intell. Transp. Syst., 2020, pp. 1–6.
[12] D. Gehrig, M. Ruegg, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, “Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction,” IEEE Robot. Automat. Lett., vol. 6, no. 2, pp. 2822–2829, Apr. 2021.
[13] L. Pan, M. Liu, and R. Hartley, “Single image optical flow estimation with an event camera,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 1669–1678.
[14] C. Lee, A. K. Kosta, and K. Roy, “Fusion-FlowNet: Energy-efficient optical flow estimation using sensor fusion and deep fused spiking-analog network architectures,” in Proc. Int. Conf. Robot. Automat., 2022, pp. 654–6510.
[15] A. R. Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, “Ultimate SLAM? combining events, images, and IMU for robust visual slam in HDR and high-speed scenarios,” IEEE Robot. Automat. Lett., vol. 3, no. 2, pp. 994–1001, Apr. 2018.
[16] J. H. Jung and C. G. Park, “Constrained filtering-based fusion of images, events, and inertial measurements for pose estimation,” in Proc. IEEE Int. Conf. Robot. Automat., 2020, pp. 644–650.
[17] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004.
[18] H. Bay, T. Tuytelaars, and L. V. Gool, “SURF: Speeded up robust features,” in Proc. Eur. Conf. Comput. Vis., 2006, pp. 404–417.
[19] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” in Proc. Int. Conf. Comput. Vis., 2011, pp. 2564–2571.
[20] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2005, vol. 1, pp. 886–893.
[21] A. Oliva and A. Torralba, “Building the gist of a scene: The role of global image features in recognition,” Proc. Brain Res., vol. 155, pp. 23–36, 2006.