VEFNET: AN EVENT AND RGB MODALITY FUSION NETWORK FOR VISUAL PLACE RECOGNITION

Ze Huang[1], Rui Huang[1], Li Sun[2], Cheng Zhao[3], Min Huang[4], and Songzhi Su[1,]*
[1]School of Informatics, Xiamen University, China
[2]Department of Computer Science, University of Sheffield, UK
[3]Department of Engineering Science, University of Oxford, UK
[4]College of Computer Engineering, Jimei University, China

ABSTRACT

Visual Place Recognition (VPR) on conventional RGB images is challenging due to the complex variety of illumination and seasonal changes. In terms of long-term localization, the emerging event stream cameras are naturally resilient to appearance changes. Thus, this paper proposes a novel multi-modal network (VEFNet) for VPR that simultaneously learns location-specific feature representations and leverages the attention mechanism to merge the location representations from color and event cameras. Specifically, we use a shared CNN backbone to extract dense features from RGB and event frames separately. Subsequently, features from two branches will be fed to the cross-modality attention module to establish correspondences between features from dual-modality. Moreover, a self-attention module is further used to learn the contextual integration within densely encoded features. Finally, the feature vector obtained after the global pooling layer will be regarded as the place representation of the dual-modality inputs. Experimental results show that the proposed method effectively combines information from both modalities and achieves SOTA performance on public datasets.

Index Terms— Visual Place Recognition, Event-based Vision, Multi-modality Fusion, Attention Mechanism

1. INTRODUCTION

Visual place recognition(VPR) solves the significant localization problem in a way of query-and-retrieval of the appearance images without need of Global Pointing System. Conventional VPR methods use RGB or gray-scale images. There are two main stream methods: either learning place-specific global descriptors or devising local descriptors and solving it in a feature matching way. Though deep learning achieves significant robustness in terms of dealing with motion blur and illumination changes, the weakness of traditional RGB cameras cannot be mitigated. Researchers investigate the use of range sensors (LiDAR) to improve the accuracy of VPR [1] [2]. However, it brings additional problems that the high power consumption and high data volume feedback. Aware of the lower power consumption, lower data volume and robustness to illumination, several previous works have begun to challenge VPR with event cameras, which can be divided into two parts according to the type of data they use. Part of the works are based on raw event-based data. [3] accumulates events under 10ms temporal resolution into a single frame, and then, those event frames will be fed into the SeqSLAM system [4]. With a consideration of the asynchronous and discrete characters of the event data, [5] proposes an MLP-based to represent event bins into EST voxel grids, then the classic NetVLAD method is used to obtain the global descriptor. Another part of the works pre-processes event stream to new data domains. [6] transforms the original event data into gray-scale frames through E2VID, and, the experimental result shows that the grey-scale event image representation is more advanced for the VPR task, compared to color images. Latter, an edge-based representation on event data is proposed [7].

The existing VPR research either uses color-image-based or event-based VPR representation, however, multi-modal representation has not been studied. Our intuition is to leverage the enriched semantics provided by traditional RGB images and the illumination invariance of event data to learn robust place-specific representation for VPR.

This paper proposes to effectively fuse data from color and event streams, thereby leveraging the strengths of multi-modal representation for VPR. Inspired by the remarkable achievements of attention mechanism in the field of modality fusion [8] [9], we design a CNN-Attention-CNN-liked network, called VEFNet. The architecture of VEFNet is shown in Fig.1. The first convolution module is mainly used to extract dense features from the dual-modality input. The two attention modules are used to fuse features from different modalities and achieve the purpose of adaptive learning respectively.

Our contributions are three-fold: (1) The cross-attention module empowers the network with capability of cross-modal feature selection to learn stable local features in the dense feature map. (2) The self-attention module assemblies the neighboring local features to build more distinctive contex-

Corresponding author: Songzhi Su, ssz@xmu.edu.cn.This research was funded by the National Natural Science Foundation of China (No. 61902330)

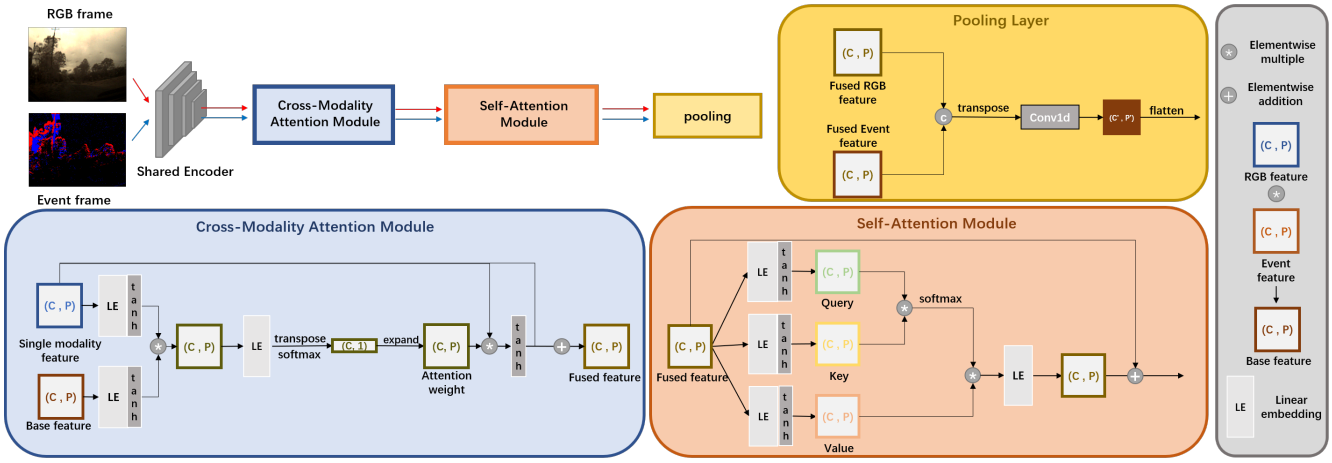

Fig. 1. Overall asymmetric architecture of the proposed VEFNet. The entire network adopts a special CNN-Attention-CNN architecture, see Section2 for more details.

tual patterns. (3) The comparison experiments on public dataset demonstrate that, our approach outperforms existing single-modality-based SOTA methods. And the ablation studies show the effectiveness of each module of VEFNet.

2. PROPOSED METHOD

2.1. Input Modality

Suppose there is an RGB image I_v taken at time t_s. A stream of events will be used to generate event frame I_e.

$$\begin{aligned}[x, y, t, +1] &\mapsto F_e[x, y] = [255, 0, 0], \\ [x, y, t, -1] &\mapsto F_e[x, y] = [0, 0, 255]\end{aligned} \tag{1}$$

Among them, x, y represent the events' location; t represents timestamp; $+1, -1$ represent the events' binary polarity. The number of events used to generate event frames is related to the event camera's resolution.

$$N = \tilde{n} * H * W \tag{2}$$

$\tilde{n}$ is a constant coefficient (0.35 is used), and the timestamps of the continuous events be used should be exactly greater than or equal to t_s. Fig.2 shows the generated data.

2.2. Shared Encoder

We crop a part of vgg16bn [10] as the shared encoder for feature extraction. Due to the same original size of the dual-modality input, the outputs of the encoder are both shaped as (C, H_e, W_e). Subsequently, the two feature maps are reshaped and fed into a shared linear embedding.

$$\begin{aligned}I_e \in \mathbb{R}^{3HW} &\xrightarrow{Encoder} Out_e \in \mathbb{R}^{CH_eW_e} \\ Out_e \in \mathbb{R}^{CH_eW_e} &\xrightarrow{reshape} Out_e \in \mathbb{R}^{C(H_e*W_e)} \\ Out_e \in \mathbb{R}^{C(H_e*W_e)} &\xrightarrow{Linear} F_e \in \mathbb{R}^{CP}\end{aligned} \tag{3}$$

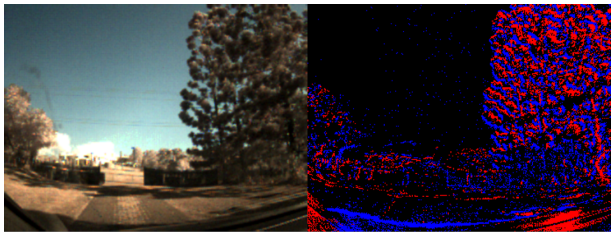

Fig. 2. The inputs of the network are two frames sized $346 * 260$:(a) RGB frame; (b) Event frame generated based on Eq.(1) and Eq.(2).

The final output of both modalities will be fed into the Cross-modality attention module introduced in the next subsection.

2.3. Cross-modality Attention Module

The Cross-modality attention module is used to fuse dense features from two different modalities extracted by the shared CNN encoder. The blue part in Fig. 1 shows the structure of the proposed Cross-modality attention module. The module computes the Cross-modality feature attention as

$$\begin{aligned}\tilde{F}_e &= F_e + F_e * Softmax(\varrho_e(\theta_e(F_e * F_v) * \varphi_e(F_e))) \\ \tilde{F}_v &= F_v + F_v * Softmax(\varrho_v(\theta_v(F_v * F_e) * \varphi_v(F_v)))\end{aligned} \tag{4}$$

where ϱ, θ and φ are all linear embeddings. Essentially, we first use the element-wise multiple method to roughly combine the two features to a base feature. Then through a series of element-wise multiplication and softmax operations, we obtain the attention score and re-weighting the single modality feature. The final fused feature will be obtained by element-wise adding the re-weighted feature and the origin single modality feature.

2.4. Self-attention Module

To boost the internal connections and focus on the important contextual information in the fused features, a self-attention module is further used based on the output of the Cross-modality attention module:

$$\begin{aligned}\tilde{\tilde{F}}_v &= \tilde{F}_v + o_v(Softmax(q_v(\tilde{F}_v) * k_v(\tilde{F}_v)) * v_v(\tilde{F}_v)) \\ \tilde{\tilde{F}}_e &= \tilde{F}_e + o_e(Softmax(q_e(\tilde{F}_e) * k_e(\tilde{F}_e)) * v_e(\tilde{F}_e))\end{aligned} \tag{5}$$

where q, k, v are all linear embeddings. This module acts on features of a single modality separately.

2.5. Pooling Module

In the pooling module, we concatenate the two features, and a $Conv1D$ layer with large kernal size of 8 is used to aggregate dimensional information:

$$\begin{aligned}\tilde{\tilde{F}}_e, \tilde{\tilde{F}}_v &\xrightarrow{concat} F \in \mathbb{R}^{C(2*P)} \\ F &\xrightarrow{Conv1D} F^{'} \in \mathbb{R}^{C^{'}P^{'}}\end{aligned} \tag{6}$$

Previously, this practice was often treated as a separate component called $whitening$ [11] [12]. We extend it to the features after the attention module and make it end-to-end trained. Finally, the feature is flattened to a vector. The closer the cosine distance of the feature vectors from two scenes, the more similar the two scenes are.

Table 1. Classification Error Rate

	Method	Error@ 1	Error@ 5
Cifar100	$vgg13bn$ [10]	28.00	9.71
	$vgg16bn$ [10]	<u>27.07</u>	<u>8.84</u>
	$vgg19bn$ [10]	27.77	<u>8.84</u>
	$attention59$ [16]	33.75	12.90
	$attention92$ [16]	36.52	11.47
	ViT [17]	43.89	17.73
	$VEFNet(ours)$	**25.04**	**6.79**

Table 2. Results of Ablation Studies

	Method	Recall@ 1	Recall@ 5	Recall@ 10
Hybrid Dataset	$p-d(ve)$	79.21	89.36	92.06
	$C-S(ve)$	65.71	75.39	86.19
	$C-S-d(ve)$	78.25	85.39	89.68
	$p-C-S(ov)$	78.41	89.52	<u>92.22</u>
	$p-C-S(ve)$	79.04	**92.38**	91.42
	$p-C-d(ve)$	<u>82.06</u>	90.79	<u>92.22</u>
	$p-S-d(ve)$	79.36	89.20	91.73
	$p-C-S-d(oe)$	57.14	78.89	86.03
	$p-C-S-d(ov)$	78.57	90.95	91.74
	$p-C-S-d(ve)$	**83.01**	<u>92.06</u>	**93.97**

3. EXPERIMENTS

3.1. Implementation Details

Four datasets are used during the whole experiment. MVSEC [13] and DDD17 [14] are employed for ablation studies. Brisbane-Event-VPR [6] is employed for comparative experiment. To ensure authenticity, simulated data are not used. In addition, Cifar100 [15] is used for network pre-training. We divide Brisbane-Event-VPR into the training set and the testing set. The training set contains triplets (anchor with positive and negative samples) from Sunset1 vs Morning, Daytime, and Sunrise sequences. Both MVSEC and DDD17 provide event stream, RGB images, and GPS data. According to these data, we manually select samples with fully considering the view and illumination change. Since there are too few samples in a single dataset, we combine the samples from MVSEC and DDD17 into one dataset, which is called hybrid dataset in the following text. Note that, we do not use hybrid dataset for comparative experiment since the low-quality sample generated for place recognization benchmarking and manual filtering of unsuitable data is necessary.

The entire network is implemented under PyTorch and trained on a single GTX 3090 Ti GPU. Take Brisbane-Event-VPR as an example. The encoder will increase the dimension of the input image to 512. After reshape and linear embedding, each feature is shaped to $(512, 256)$. The final pooling module reduces the feature dimension to 32. During training, the batch size is set to 16 (each batch contains anchor samples, positive samples, and 4 negative samples from dual-modality, for a total number of 12 samples). Triplet loss with a margin of 0.1 and the SGD solver with initial learning rate of 0.001 are used. All training-based comparison methods are also performed under the same configuration.

3.2. Network Pre-training

We initially train the entire network on Cifar100 [15] for the classification task to obtain pre-trained weights, and the pre-trained weights are further used to challenge VPR. This is done so that the network can gain prior knowledge, and also be able to allow a fair comparison with some comparative methods which use large-scale pre-trained models. Even though our network is designed for dual-modality input, our network still outperforms the VGG series and some attention-based networks on Cifar100 with only RGB input. Note that for fairness, we did not use extra data and training tricks except learning rate decay. Tab.1 shows the results.

Table 3. Results Compared With The SOTA Methods

Sequence	Method	Recall@		
		1	5	10
Sunset2 vs Sunset1	$NetVLAD$ [18]	63.42	**96.28**	97.70
	$SuperGlue$ [19] [20]	52.12	81.27	88.16
	$PatchNetVLAD$ [11]	64.07	90.62	90.80
	$Fischer\ et\ al.$ [6]	56.18	-	-
	$VEFNet(ve)$	**67.31**	95.40	**97.87**
Sunset2 vs Daytime	$NetVLAD$ [18]	52.12	85.68	92.57
	$SuperGlue$ [19] [20]	24.55	49.82	60.77
	$PatchNetVLAD$ [11]	46.55	66.73	67.43
	$Fischer\ et\ al.$ [6]	15.01	-	-
	$VEFNet(ve)$	**53.18**	**88.33**	**93.46**
Sunset2 vs Morning	$NetVLAD$ [18]	67.49	95.40	**97.97**
	$SuperGlue$ [19] [20]	51.59	80.74	88.51
	$PatchNetVLAD$ [11]	**72.21**	88.14	88.51
	$Fischer\ et\ al.$ [6]	34.98	-	-
	$VEFNet(ve)$	71.98	**95.93**	97.87
Sunset2 vs Sunrise	$NetVLAD$ [18]	60.42	**91.34**	**94.52**
	$SuperGlue$ [19] [20]	36.04	61.13	69.25
	$PatchNetVLAD$ [11]	**65.01**	82.30	82.83
	$Fischer\ et\ al.$ [6]	36.21	-	-
	$VEFNet(ve)$	62.72	90.45	94.16
Avg	$NetVLAD$ [18]	60.86	92.17	95.69
	$SuperGlue$ [19] [20]	41.08	68.24	76.67
	$PatchNetVLAD$ [11]	61.96	81.94	82.39
	$Fischer\ et\ al.$ [6]	35.60	-	-
	$VEFNet(ve)$	**63.78**	**92.53**	**95.84**

3.3. Ablation Studies

We conduct ablation studies on the hybrid dataset to verify the effectiveness of each module of VEFNet. Tab.2 shows the results. The meanings of the symbols in Tab.2 are as follows. p: use pre-trained model on Cifar100; C: use Cross-modality attention module; S: use self-attention module; d: use pooling module; ov: only RGB data used; oe: only event data used; ve: both RGB and event data are used.

As can be seen from the table, the results based on dual-modality input are significantly better than those based on single modality. The Cross-modality attention module can bring an average of 2-3 percentage points of gain for all metrics, and 1-2 from the self-attention module. Compared with single modality input, the pooling module is more helpful for the results of dual-modality input. Interestingly, even the single-modality-based pre-trained model can still bring great improvements to our fusion-based network.

3.4. Compare With State-of-the-art Methods

We conduct comparative experiment with open source state-of-the-art methods on Brisbane-Event-VPR. Tab.3 shows the results.

Fig. 3. The visual comparison on Brisbane-Event-VPR. The leftmost column is the image to be retrieved, and the right four columns are the best matching images obtained in the database by different methods. Our method makes correct matches even in environments with extreme illuminations and tiny markers.

Our method achieves the best results on most of the sequences and whole average evaluation metrics. On the metric $Recall@1$, our method stands out. Fig.3 gives the visualization results. Due to the auxiliary of event stream data, our method can deal with the localization under extreme illumination variance, so we can achieve better results on Brisbane-Event-VPR (see first two lines in Fig.3). The attention mechanism used in our approach can learn both intra-modal and cross-modal contextual patterns for robust place recognition. As shown in last two lines in Fig.3, our approach can retrieve the correct places with traffic lights, where fully-convolutional methods cannot works in such situations. Finally, the form of global feature representation enables our method work on data with low resolution and lack of significant geometric features, which is impossible for local-feature-based VPR methods such as SuperGlue.

4. CONCLUSION

This paper has proposed VEFNet, a modality fusion network that effectively fuses information from two different types of sensors, i.e. color camera and event camera for VPR. A high-quality backbone is utilized to extract dense features from the dual-modality input, and the Cross-modality attention module is proposed to fuse the two parts feature. Furthermore, we also use a channel self-attention module to enhance the correlation within individual features. The entire network is trained end-to-end. Ablation studies demonstrate the effectiveness of the proposed modules. The comparison results show our fusion method achieves SOTA performance on public dataset in terms of recall value, and prove that event camera is helpful for VPR.

5. REFERENCES

[1] Luca Di Giammarino, Irvin Aloise, Cyrill Stachniss, and Giorgio Grisetti, "Visual place recognition using lidar intensity information," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4382–4389.

[2] Zhicheng Zhou, Cheng Zhao, Daniel Adolfsson, Songzhi Su, Yang Gao, Tom Duckett, and Li Sun, "Ndt-transformer: Large-scale 3d point cloud localisation using the normal distribution transform representation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5654–5660.

[3] Michael Milford, Hanme Kim, Michael Mangan, Stefan Leutenegger, Tom Stone, Barbara Webb, and Andrew Davison, "Place recognition with event-based cameras and a neural implementation of seqslam," *arXiv preprint arXiv:1505.04548*, 2015.

[4] Michael J Milford and Gordon 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] Delei Kong, Zheng Fang, Haojia Li, Kuanxu Hou, Sonya Coleman, and Dermot Kerr, "Event-vpr: End-to-end weakly supervised network architecture for event-based visual place recognition," *arXiv preprint arXiv:2011.03290*, 2020.

[6] Tobias Fischer and Michael Milford, "Event-based visual place recognition with ensembles of temporal windows," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6924–6931, 2020.

[7] Alex Junho Lee and Ayoung Kim, "Eventvlad: Visual place recognition with reconstructed edges from event cameras," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 2247–2252.

[8] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun, "Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[9] Xinrui Song, Hengtao Guo, Xuanang Xu, Hanqing Chao, Sheng Xu, Baris Turkbey, Bradford J Wood, Ge Wang, and Pingkun Yan, "Cross-modal attention for mri and ultrasound volume registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 66–75.

[10] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias 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. 14141–14152.

[12] Hernán Badino, Daniel Huber, and Takeo Kanade, "Visual topometric localization," in *2011 IEEE Intelligent vehicles symposium (IV)*. IEEE, 2011, pp. 794–799.

[13] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis, "The multivehicle stereo event camera dataset: An event camera dataset for 3d perception," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2032–2039, 2018.

[14] Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck, "Ddd17: End-to-end davis driving dataset," *arXiv preprint arXiv:1711.01458*, 2017.

[15] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[16] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef 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.

[19] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[20] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.