BI-DIRECTIONAL MODALITY FUSION NETWORK FOR AUDIO-VISUAL EVENT LOCALIZATION

Shuo Liu[1,2], *Weize Quan*[1,2*], *Yuan Liu*[3], *Dong-Ming Yan*[1,2]

[1]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]Speech Lab, Alibaba Group, China

ABSTRACT

Audio and visual signals stimulate many audio-visual sensory neurons of persons to generate audio-visual contents, helping humans perceive the world. Most of the existing audio-visual event localization approaches focus on generating audio-visual features by fusing the audio and visual modalities for final predictions. However, an audio-visual adjustment mechanism exists in a complicated multi-modal perception system. Inspired by this observation, we propose a novel bi-directional modality fusion network (BMFN), which not only simply fuses audio and visual features, but also adjusts the fused features to increase their representativeness with the help of the original audio and visual contents. The high-level audio-visual features achieved from two directions with two forward-backward fusion modules and a mean operation are summarized for the final event localization. Experimental results demonstrate that our method outperforms state-of-the-art works in both fully- and weakly- supervised learning settings. The code is available at `https://github.com/weizequan/BMFN.git`.

Index Terms— Event Localization, Bi-Directional, Audio-Visual Modality Fusion, Multi-Modal Perception System

1. INTRODUCTION

Humans' multi-modal perception system [1, 2] is really helpful for scene understanding, where audio-visual sensory neurons play an important role in generating and adjusting the audio-visual content. Consequently, audio-visual event (AVE) localization, which aims to know what and when audio-visual event occurs in machines (cf. Fig. 1), should be studied.

Many recent learning-based approaches have been proposed to solve the AVE localization tasks. Tian *et al.* [3] proposed a CNN model, which consists of an audio-guided visual attention block and a dual multi-modal residual block, to fuse the audio and visual features trained on the shared AVE dataset. Lin *et al.* [4] proposed a sequence-to-sequence dual network, which learns global and local audio-visual features to improve the model's capacity. Ramaswamy [5] successively explored the inter- and intra-modality interactions and then concatenated them as the final audio-visual fusion features for AVE localization. Similarly, Xu *et al.* [6] learned the audio-visual relationship via the encoder of transformer [7, 8]. These works intended to utilize the audio and visual signals by following only the forward direction to generate the final fused features. However, according to mammals' multi-modal perception system, an adjustment mechanism [9] for enhancing the audio-visual fusion signals also exists to obtain more information and thus perceive scenes accurately.

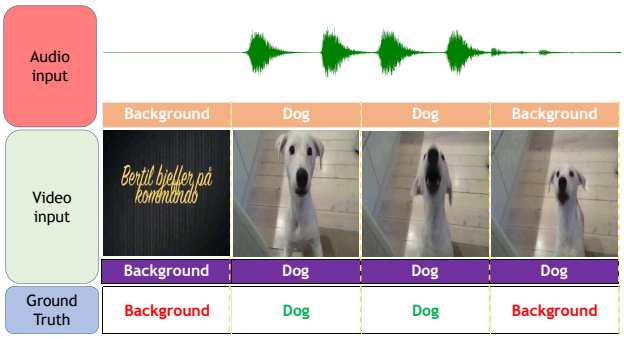

Fig. 1. An example of the audio-visual event localization: an audio-visual event is both audible and visible event, otherwise, it will be regarded as background. The goal of the task is to predict the segment-level event.

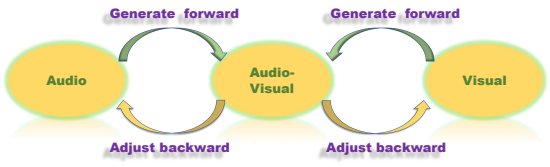

Fig. 2. Mechanism of bi-directional modality fusion: generate forward and adjust backward the audio-visual features.

This observation motivates us to efficiently fuse audio and visual features in a bi-directional way (cf. Fig. 2). On the one hand, the audio and visual features are forward close to the audio-visual features. On the other hand, the audio-visual features are adjusted backward with the help of audio and vi-

*Corresponding author: qweizework@gmail.com.

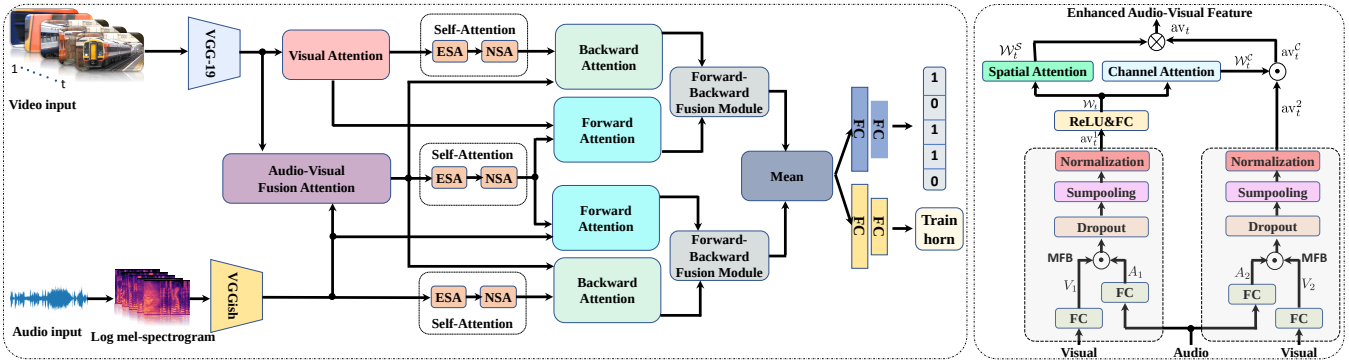

Fig. 3. Left: the architecture of BMFN. Audio-Visual Fusion Attention module produces the audio-visual features by fusing the original audio features and visual features. Self-attention operations consisting of expansion self-attention (ESA) and normal self-attention (NSA) are applied to the corresponding features. Then, forward and backward attention modules are utilized to generate and adjust the audio-visual features. Forward-Backward Fusion modules (FBFM) can integrate features from two directions. The final comprehensively fused features are the mean of results from FBFM. The predictions are obtained by event-relevant results and event-category results as [10]. Right: the architecture of AVFA. Audio-visual features obtained by MFB are enhanced in channel and spatial dimensions. $\odot$ and $\otimes$ separately stand for the element-wise and matrix multiplications.

sual features to improve the feature representation capacity. The information from all audio-visual fusion cells in these two branches is comprehensively and effectively integrated for final event localization. Our key contributions are as follows: i) we propose an audio-visual fusion attention (AVFA) module to obtain high-level semantic audio-visual features; ii) we propose a novel bi-directional modality fusion network (BMFN) to generate and adjust audio-visual features; iii) our extensive experimental results on the AVE dataset [3] illustrate the superiority of our BMFN.

2. PROPOSED METHOD

2.1. Problem Statement

Each segment of the videos from the AVE dataset [3] is represented as $S = (v_t, a_t)_{t=1}^{T}$, where v_t and a_t respectively stand for the visual and audio features of the t-th segment. The aim is to localize the segment-level event. The detailed segment-level event label $y_t = \{y_t^c \mid y_t^c \in \{0,1\}, c = 1, ..., C, \sum_{c=1}^{C} y_t^c = 1\}$ can be obtained in a fully-supervised setting, where C is the total number of event categories plus one background label. Otherwise, only the video-level label is provided for training in the weakly-supervised setting.

2.2. Audio-Visual Bi-directional Fusion Network

The overall architecture of the BMFN is shown in Fig. 3(left). The details of all the modules are described below, where σ, δ, $\tanh$, and SM respectively stand for the ReLU, sigmoid, tanh, and softmax functions, $\odot$ and $\otimes$ respectively denote the element-wise and matrix multiplications, and GVP denotes the global average pooling in the spatial dimension.

Feature Extraction: According to previous works [3, 10, 4, 11], we apply the VGG-19 model [12] (pre-trained on ImageNet [13]) to extract the visual features $v_t \in \mathbb{R}^{d_v \times (HW)}$ (where H and W are the height and width of feature maps, respectively). In addition, we utilize VGGish [14] (pre-trained on AudioSet [15]) to extract the audio features $a_t \in \mathbb{R}^{d_a}$.

Visual Attention: To reduce the spatial dimension and enhance the representation capability of visual features elegantly, the visual features are enhanced by successively applying the channel attention $\mathcal{M}_t^{\mathcal{C}} \in \mathbb{R}^{d_v \times 1}$ and spatial attention $\mathcal{M}_t^{\mathcal{S}} \in \mathbb{R}^{1 \times (HW)}$. The entire computation process of the enhanced visual features $v_t^{\mathcal{E}} \in \mathbb{R}^{d_v}$ is written as:

$$\begin{aligned} &\mathcal{M}_t^{\mathcal{C}} = \delta(\mathbf{W}_2\sigma(\mathbf{W}_1(\mathrm{GVP}(\Phi_v^c v_t)))), v_t^{\mathcal{C}} = \mathcal{M}_t^{\mathcal{C}} \odot v_t, \\ &\mathcal{M}_t^{\mathcal{S}} = \mathrm{SM}\left(\tanh\left(\mathbf{W}_3 v_t^{\mathcal{C}}\right)\right), v_t^{\mathcal{E}} = v_t^{\mathcal{C}} \otimes \left(\mathcal{M}_t^{\mathcal{S}}\right)^T, \end{aligned} \tag{1}$$

where $\Phi_v^c \in \mathbb{R}^{d_v \times d_v}$ is a fully-connected layer with ReLU activation, $\mathbf{W}_1 \in \mathbb{R}^{d \times d_v}$, $\mathbf{W}_2 \in \mathbb{R}^{d_v \times d}$ and $\mathbf{W}_3 \in \mathbb{R}^{1 \times d_v}$ are three linear transformations with $d = 256$.

Fusion Feature Attention: As shown in Fig. 3(right), we design an AVFA block to initially fuse the audio and visual features and extract high-level semantic features by applying the multi-modal factorized bilinear (MFB) method [16] and the spatial and channel attention mechanisms. $\mathrm{av}_t^1, \mathrm{av}_t^2 \in \mathbb{R}^{d_v \times (HW)}$ are separately achieved by applying the MFB method on $V_1 = \Phi_{V_1} v_t^s$, $A_1 = \Phi_{A_1} a_t$ and $V_2 = \Phi_{V_2} v_t^s$, $A_2 = \Phi_{A_2} a_t$ in each spatial location $s \in [1, ..., HW]$, where $\Phi_{V_1}, \Phi_{V_2} \in \mathbb{R}^{d_m \times d_v}$ and $\Phi_{A_1}, \Phi_{A_2} \in \mathbb{R}^{d_m \times d_a}$ are rise-dimension transformations with $d_m = 2560$. av_t^1 can be computed as follows:

$$\begin{aligned} &\tilde{\mathrm{av}}_t^1 = D(\mathrm{SP}(V_1 \odot A_1, p)), \\ &\hat{\mathrm{av}}_t^1 = \mathrm{sign}(\tilde{\mathrm{av}}_t^1)\sqrt{\left|\tilde{\mathrm{av}}_t^1\right|}, \mathrm{av}_t^1 = \hat{\mathrm{av}}_t^1 / \left\|\hat{\mathrm{av}}_t^1\right\|, \end{aligned} \tag{2}$$

where $\mathrm{SP}(f, p)$ represents the sum pooling operation with latent parameter $p = 5$, and $D(\cdot)$ is a dropout layer. The power and L_2 normalizations are applied to stabilize the model training. The channel-attentive weight $\mathcal{W}_t^{\mathcal{C}}$ and spatial attentive

weight $\mathcal{W}_t^{\mathcal{S}}$ are formulated from $\mathcal{W}_t$ ($\mathcal{W}_t = \sigma(\mathbf{W}_4(\mathrm{av}_t^1))$), which are then applied to the features av_t^2 yielding final attentive fusion features $\mathrm{av}_t \in \mathbb{R}^{d_v}$. The details are as follows:

$$\begin{aligned} &\mathcal{W}_t^{\mathcal{C}} = \delta(\mathbf{W}_5(\mathrm{GVP}(\mathcal{W}_t))), \mathrm{av}_t^{\mathcal{C}} = \mathcal{W}_t^{\mathcal{C}} \odot \mathrm{av}_t^2, \\ &\mathcal{W}_t^{\mathcal{S}} = \mathrm{SM}\left(\tanh\left(\mathbf{W}_6\left(\mathcal{W}_t\right)\right)\right), \mathrm{av}_t = \mathrm{av}_t^{\mathcal{C}} \otimes \left(\mathcal{W}_t^{\mathcal{S}}\right)^T, \end{aligned} \tag{3}$$

where $\mathbf{W}_4 \in \mathbb{R}^{d \times d_v}$, $\mathbf{W}_5 \in \mathbb{R}^{d_v \times d}$, and $\mathbf{W}_6 \in \mathbb{R}^{1 \times d}$ are three linear transformations with $d = 256$.

Self-Attention: To explore the temporal relationship for each modality, we separately feed the $v_t^{\mathcal{E}}, a_t, \mathrm{av}_t$ into self-attention blocks obtaining self-attentive features $v_t^{\mathrm{self}}, a_t^{\mathrm{self}}, \mathrm{av}_t^{\mathrm{self}} \in \mathbb{R}^d$. A self-attention block is composed of an expansion self-attention (ESA) and a normal self-attention (NSA). ESA is an encoder of transformer embedded with linear expansion and linear reduction according to [17]. The details are as follows, where BN and LN respectively refer to the batch and layer normalizations and FFN is the feed forward network [7]:

$$\begin{aligned} &\hat{Q} = Q\mathbf{W}_{\hat{Q}}, \hat{K} = K\mathbf{W}_{\hat{K}}, \hat{V} = V\mathbf{W}_{\hat{V}}, \\ &\mathrm{Att} = \mathrm{SM}\left(\frac{\hat{Q}\hat{K}^T}{\sqrt{d}}\right), A = \mathrm{FFN}(\mathrm{LN}(B + \hat{Q})), \\ &\hat{\mathrm{Att}} = \mathrm{BN}(\mathbf{W}_8(\delta(\mathrm{BN}(\mathbf{W}_7\mathrm{Att})))), B = (\hat{\mathrm{Att}} + \mathrm{Att})\hat{V}, \\ &\mathrm{Encoder}_{\mathrm{ESA}}(K, Q, V) = \mathrm{LN}(A + B), \end{aligned} \tag{4}$$

where $Q, K, V \in \mathbb{R}^{T \times d}$ are input reshaped features with $d = 256$; $\mathbf{W}_{\hat{Q}}, \mathbf{W}_{\hat{K}}, \mathbf{W}_{\hat{V}} \in \mathbb{R}^{d \times d}$ are projection matrices; Att and $\hat{\mathrm{Att}}$ are attention maps with the dimension of $n_h \times T \times T$ (the head n_h equals to 4 and the temporal dimension T equals to 10); $\mathbf{W}_7 \in \mathbb{R}^{(n_h * r) \times n_h}$(the expansion ratio r is 2) is the linear attention expansion matrix; and $\mathbf{W}_8 \in \mathbb{R}^{n_h \times (n_h * r)}$ is the linear reduction matrix. NSA is an encoder of transformer block [7]. The calculation is as follows:

$$\begin{aligned} &\tilde{Q} = Q\mathbf{W}_{\tilde{Q}}, \tilde{K} = K\mathbf{W}_{\tilde{K}}, \tilde{V} = V\mathbf{W}_{\tilde{V}}, \\ &X = \mathrm{MHA}(\tilde{Q}, \tilde{K}, \tilde{V}), Y = \mathrm{FFN}(\mathrm{LN}(X + \tilde{Q})), \\ &\mathrm{Encoder}(K, Q, V) = \mathrm{LN}(X + Y), \end{aligned} \tag{5}$$

where Q, K, $V \in \mathbb{R}^{T \times d}$ are output features from ESA; $\mathbf{W}_{\tilde{Q}}, \mathbf{W}_{\tilde{K}}, \mathbf{W}_{\tilde{V}} \in \mathbb{R}^{d \times d}$ are projection matrices; and MHA is the multi-head attention [7] with 4 heads.

Bi-directional Modality Fusion: Inspired by the enhancement and adjustment mechanisms in the audio-visual cells [9], we propose a bi-directional modality fusion module for obtaining more graceful fusion features. The module consists of the forward and backward attention blocks, where the forward attention module enables the unimodal features to step close to the fusion features, and the backward attention module adjusts the fusion features with the help of unimodal features for more representative features. Two forward attention blocks (FA) exist, which are represented as FA_1 and FA_2, whose queries are $v_t^{\mathcal{E}}$ and a_t, respectively, and keys (same as values) separately are $v_t^{\mathcal{E}}$ and a_t concatenated with $\mathrm{av}_t^{\mathrm{self}}$. Two backward attention blocks (BA) exist in the network, including BA_1 and BA_2, whose queries are attentive fusion feature av_t, and keys (same as values) separately are v_t^{self} and a_t^{self} concatenated with av_t. FA_1, FA_2, BA_1, and BA_2 follow the encoder architecture of transformer like Eqn.(5) and output the audio-visual features fa_t^1, fa_t^2, ba_t^1, $\mathrm{ba}_t^2 \in \mathbb{R}^d$, respectively. The fusion process is written as:

$$\begin{aligned} &\mathrm{ba}_t^1 = \mathrm{Encoder}(\mathrm{av}_t, \mathrm{cat}(\mathrm{av}_t, v_t^{\mathrm{self}}), \mathrm{cat}(\mathrm{av}_t, v_t^{\mathrm{self}})), \\ &\mathrm{fa}_t^1 = \mathrm{Encoder}(v_t^{\mathcal{E}}, \mathrm{cat}(v_t^{\mathcal{E}}, \mathrm{av}_t^{\mathrm{self}}), \mathrm{cat}(v_t^{\mathcal{E}}, \mathrm{av}_t^{\mathrm{self}})), \\ &\mathrm{fa}_t^2 = \mathrm{Encoder}(a_t, \mathrm{cat}(a_t, \mathrm{av}_t^{\mathrm{self}}), \mathrm{cat}(a_t, \mathrm{av}_t^{\mathrm{self}})), \\ &\mathrm{ba}_t^2 = \mathrm{Encoder}(\mathrm{av}_t, \mathrm{cat}(\mathrm{av}_t, a_t^{\mathrm{self}}), \mathrm{cat}(\mathrm{av}_t, a_t^{\mathrm{self}})). \end{aligned} \tag{6}$$

Forward-Backward Fusion Module (FBFM): The FBFM is also an encoder of transformer and designed for further integrating the two directional fusion features. For example, fa_t^1 and ba_t^1 are fed into FBFM_1 obtaining $F_1 \in \mathbb{R}^d$, where the query is the element-wise multiplication of fa_t^1 and ba_t^1, and key same as value is the temporal concatenation of fa_t^1 and ba_t^1. fa_t^2 and ba_t^2 are also processed in the same manner via FBFM_2 obtaining $F_2 \in \mathbb{R}^d$. The detailed formulation is summarized below:

$$\begin{aligned} &F_1 = \mathrm{Encoder}(\mathrm{fa}_t^1 \odot \mathrm{ba}_t^1, \mathrm{cat}(\mathrm{fa}_t^1, \mathrm{ba}_t^1), \mathrm{cat}(\mathrm{fa}_t^1, \mathrm{ba}_t^1)), \\ &F_2 = \mathrm{Encoder}(\mathrm{fa}_t^2 \odot \mathrm{ba}_t^2, \mathrm{cat}(\mathrm{fa}_t^2, \mathrm{ba}_t^2), \mathrm{cat}(\mathrm{fa}_t^2, \mathrm{ba}_t^2)). \end{aligned} \tag{7}$$

Fully-Supervised Event Localization: The mean of F_1 and F_2 is formulated as the final fusion feature F, which is similar to integrating information from all audio-visual cells. Eqn.(8) shows that F is used to compute the event-relevant score $s \in \mathbb{R}^T$ and event category score $s_c \in \mathbb{R}^C$, where T and C are the number of temporal dimension and foreground categories.

$$s = \mathrm{Sigmoid}(\mathrm{FC}(F)), s_c = \mathrm{SM}(\mathrm{FC}(\mathrm{MP}(F))), \tag{8}$$

where FC is the classifier and MP denotes the max-pooling operation. For fully-supervised training, the segment-level event label is available. The corresponding objective function is the summation of the binary cross-entropy loss for s and the cross-entropy loss for s_c. In the inference stage, if $s_t \geqslant 0.5$, the t-th segment is predicted as an event, and its category is depended on s_c, otherwise, it is predicted as background.

Weakly-Supervised Event Localization: Only video-level labels can be used in this setting, and joint scores $s_f \in \mathbb{R}^{T \times C}$ are calculated by element-wise multiplying s copied C times with s_c duplicated T times. Video-level predictions can be obtained by aggregating segment-level predictions s_f into the MIL pooling [18]. The inference method is the same as that in the fully-supervised setting.

3. EXPERIMENTS

Dataset: The *Audio-Visual Event* dataset [3] is used to evaluate our model, which is a subset of AudioSet [15]. The dataset has 4,143 videos and 28 audio-visual event categories. Each video is 10s, where the audio-visual event lasts from 2s to

Table 1. Comparisons of accuracy (%) on the AVE dataset for the fully- and weakly-supervised settings. * and ** separately mean that the results are reproduced by [6] and [11].

Methods	Fully-super.	Weakly-super.
AVEL(Only Video) [3]	55.3	52.9
AVEL(Only Audio) [3]	59.5	53.4
AVSDN* [4]	72.6	66.8
AVEL [3]	72.7	66.7
CMAN** [19]	73.3	70.4
DAM [10]	74.5	-
AVRB [20]	74.8	68.9
AVIN [5]	75.2	69.4
AVT [21]	76.8	70.2
CMRAN [6]	77.4	72.9
PSP [11]	77.8	73.5
Ours	**78.7**	**74.0**

10s. The video- and segment-level event labels are provided. The used train/test split is same with all the existing methods.

Training Details: Our experiments are implemented using PyTorch 1.2.1 on a TITAN RTX GPU of NVIDIA®. The Adam optimizer with a mini-batch of 32 is used to train the models. We set the initial learning rate to 5e-4, which we divide by 2 every 10 epochs, and freeze after 30 epochs.

Comparison Results: Table 1 shows the performance comparisons between our method and state-of-the-art methods in terms of both fully- and weakly-supervised AVE localization tasks under the fair experimental setting. Our method follows the bi-directional modality fusion and consistently achieves the highest accuracies, *i.e.*, 78.7% and 74.0%, outperforming all other methods, which only adopt one direction. In addition, Fig. 4 shows a comparison example with CMRAN [6]. The audio signal from rat is weak, and the cover of the hand in vision makes this example very challenging. The rich high level semantic audio-visual features extracted from our bi-directional method provide exact predictions.

Ablation Studies: Table 2 reports the ablation study on the bi-directional modality fusion method. BMFN* is a variant of BMFN, whose FBFM_1^* is used to process features fa_t^1 and fa_t^2 instead of ba_t^1, and FBFM_2^* fuses features ba_t^2 and ba_t^1 rather than fa_t^2. In other words, FBFM_1^* only joins the audio-visual features obtained through the forward direction, and the audio-visual features from the backward direction are only fed into FBFM_2^*. The first and second rows indicate that only one direction is adopted for localization. The corresponding results show the limitation of obtaining audio-visual features via one direction. The efficiency of fusing forward and backward features via the FBFM module can be seen by comparing BMFN* with BMFN. Furthermore, compared with the sixth row, the performance drop of the fourth and fifth rows (separately removing FBFM_1 and FBFM_2) demonstrates that richer and more comprehensive audio-visual features are helpful in improving the localization performance. Table 3 shows the ablation study of the final joint method for all the audio-visual cells, indicating that a simple average obtains the best performance.

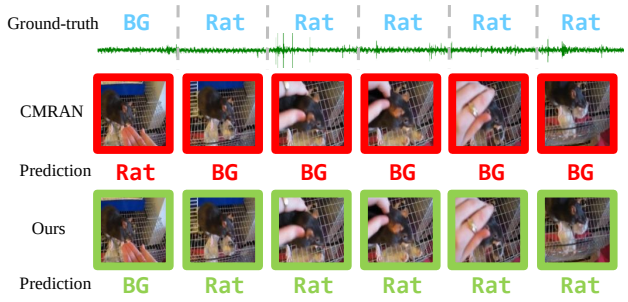

Fig. 4. Qualitative comparisons with CMRAN [6]. Green and red color separately refer to correct and incorrect results.

Table 2. Ablation study on the effect of bi-directional mechanism and FBFM module, measured by the accuracy (%).

Methods	Fully-super.	Weakly-super.
BMFN* w/o FBFM_1^*	75.2	72.9
BMFN* w/o FBFM_2^*	76.2	71.8
BMFN*	76.0	72.8
BMFN w/o FBFM_1	77.7	73.0
BMFN w/o FBFM_2	75.9	71.9
BMFN	**78.7**	**74.0**

Table 3. Ablation study on the final joint method, measured by the accuracy (%). Concat operation is applied in the channel dimension.

Methods	Fully-super.	Weakly-super.
Add	78.6	71.9
Element-wise Multiply	76.6	70.5
Concat	76.5	72.0
Mean	**78.7**	**74.0**

4. CONCLUSION

In this paper, we propose a novel BMFN, which effectively obtains audio-visual fusion features via two directions, including "generate forward" and "adjust backward", and then these high-level joint features are summarized for final localization. Quantitative and qualitative comparisons in both fully- and weakly-supervised settings with the current state-of-the-art methods show the superiority of our method.

5. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (62102418 and 62172415), and the Alibaba Group through Alibaba Innovative Research Program.

6. REFERENCES

[1] Linda Smith and Michael Gasser, "The development of embodied cognition: Six lessons from babies," *Artificial Life.*, vol. 11, no. 1-2, pp. 13–29, 2005.

[2] David A Bulkin and Jennifer M Groh, "Seeing sounds: visual and auditory interactions in the brain," *Current Opinion in Neurobiology.*, vol. 16, no. 4, pp. 415–419, 2006.

[3] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu, "Audio-visual event localization in unconstrained videos," in *Proc. of the European Conference on Computer Vision.*, 2018, pp. 247–263.

[4] Yu-Chiang Frank Wang Yan-Bo Lin, Yu-Jhe Li, "Dual-modality seq2seq network for audio-visual event localization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 2002–2006.

[5] Janani Ramaswamy, "What makes the sound?: A dual-modality interacting network for audio-visual event localization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 4372–4376.

[6] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan, "Cross-modal relation-aware networks for audio-visual event localization," in *Proc. of the ACM International Conference on Multimedia*, 2020, pp. 3893–3901.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and et al., "Attention is all you need," in *Proc. of the International Conference on Neural Information Processing Systems.*, 2017, pp. 6000–6010.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.

[9] Xiaoyan Wang, Liping Yu, Xiangyao Li, Jiping Zhang, and Xinde Sun, "Auditory-visual multisensory neurons and auditory-visual information integration in newborn rat cortex," *Chinese Journal of Zoology*, 2006.

[10] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang, "Dual attention matching for audio-visual event localization," in *Proc. of the IEEE International Conference on Computer Vision.*, 2019, pp. 6292–6300.

[11] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang, "Positive sample propagation along the audio-visual event line," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition.*, 2021, pp. 8436–8444.

[12] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition.*, 2009, pp. 248–255.

[14] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, and et al., "Cnn architectures for large-scale audio classification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 131–135. 2017.

[15] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 776–780.

[16] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. of the IEEE International Conference on Computer Vision.*, 2017, pp. 1821–1830.

[17] Daquan Zhou, Yujun Shi, Bingyi Kang, Weihao Yu, Zihang Jiang, Yuan Li, Xiaojie Jin, Qibin Hou, and Jiashi Feng, "Refiner: Refining self-attention for vision transformers," *CoRR*, 2021.

[18] Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition.*, 2015, pp. 3460–3469.

[19] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan, "Cross-modal attention network for temporal inconsistent audio-visual event localization," in *Proc. of the AAAI Conference on Artificial Intelligence.*, 2020, pp. 279–286.

[20] Janani Ramaswamy and Sukhendu Das, "See the sound, hear the pixels," in *IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2959–2968.

[21] Yan-Bo Lin and Yu-Chiang Frank Wang, "Audiovisual transformer with instance attention for audio-visual event localization," in *Proceedings of the Asian Conference on Computer Vision.*, 2020.