Look longer to see better: Audio-visual event localization by exploiting long-term correlation

Longyin Guo[1], Qijun Zhao[1,2,3,*], Hongmei Gao[3]
[1]National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, China
[2]College of Computer Science, Sichuan University, Chengdu, China
[3]School of Information Science and Technology, Tibet University, Lhasa, China
Email:guolongyin@stu.scu.edu.cn, qjzhao@scu.edu.cn, ghm@utibet.edu.cn

***Abstract*—Visual and auditory modalities both contain a large amount of rich information about audio-visual events. While the human perception system can effectively fuse the information of the dual modalities in recognizing events, it is still an open issue how to effectively integrate dual-modal information for the task of automatic localization of audio-visual events in videos. In this paper, we propose an audio-visual long-term correlation network to capture the longer correlation of audio and visual features, which is underused by existing methods. To this end, we first propose the time-spatial guided attention (TSGA) module, which locates the spatial region of the audio-visual events in the video and focuses on continuous changes in that location. We then propose the positive time residual fusion (PTRF) module, which encodes the temporal correlation matrix of video and audio, and uses residual fusion to combine audio and visual features. We finally evaluate our method for the fully supervised and weakly supervised tasks on the AVE dataset. The results prove the superiority of our method over its counterparts.**

***Index Terms*—audio-visual event, time-spatial guide, temporal correlation, residual fusion.**

I. Introduction

Hearing and vision, as the two types of perceptions that contain the most abundant information, play an important role in the perception system. With the help of audio and visual information, human perception system can effectively combine information to complete complex tasks [2]. Imitating the human perception system, audio-visual learning [24] uses two modalities of audio and visual to complete a variety of practical tasks that cannot be solved by a single modality, such as audio-visual speaker verification [14], audio-visual emotion recognition [1], talking face generation [22]. However, visual and audio do not always assist each other. In unconstrained video, the subjects of visual and audio events might do not correspond, which makes event recognition challenging. Being aware of such challenge, Tian et al. [16] proposed a task of audio-visual event (AVE) localization, which refers to the use of audio and visual features to locate events that are both visible and audible in temporal sequence. As shown in Figure 1, the task of audio-visual event localization not only requires distinguishing events and backgrounds in complex environments, but also requires that the identified events must simultaneously appear in visual and audio modalities. Therefore, how to detect the region where the event is located in the visual modality and to distinguish whether the dual-modal features are consistent in the audio-visual event are crucial to this task.

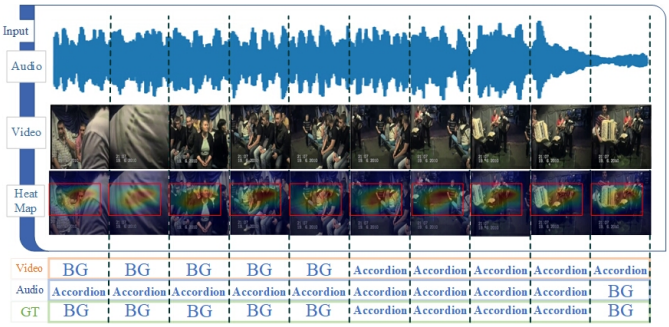


Fig. 1. Illustration of audio-visual event localization. The event is playing accordion which must be both visible (the subject of accordion is present) and audible (the sound of accordion is present), only 6-9 segments are annotated as ground truth (GT). The heat maps illustrate the attention of our proposed method.

Previous AVE localization methods [11], [18] mostly focus on combining the intra-modal and inter-modal features of visual and audio to improve spatial region localization. Yet, localization of the visual space region is also affected by the dual-modal fusion module. Therefore, for improving the performance of fusion modules, some methods [3], [19]–[21] exploit cross-modal co-attention to explore temporal correlation and synchronous pairing of audio and visual. However, they do not consider eliminating the interference of uncorrelated paired audio-visual segments in the fusion of features. Zhou et al. [22] proposed positive sample propagation (PSP) to cut low correlation connections. Despite its promising performance, PSP requires additional supervision signals, and does not consider optimizing the location of visual space region. Moreover, all the above methods fail to consider the continuous changes of the region where the event is located on successive time segments.

In this paper, we argue that the variation of the event region during the duration of the event can help distinguishing between background and event, and such variation can be exploited by considering long-term correlation in the video. We thus propose an audio-visual long-term correlation network for audio-visual event localization, which includes two key novel modules, time-spatial guided attention (TSGA) module

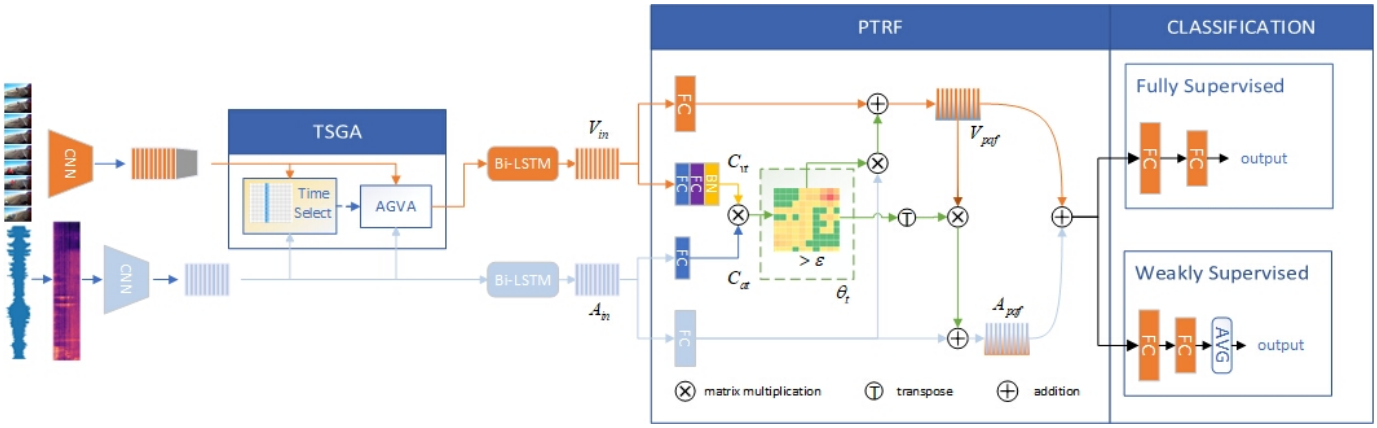


Fig. 2. Overview of our proposed audio-visual long-term correlation network. We extract audio and visual features from two CNN backbones. The TSGA utilizes the spatial and temporal relation information of audio and visual to focus on visual spatial area. Then the temporal feature on the single modality is encoded by LSTM. The PTRF consists of a few fully connected layers that encode features to find which audio and visual pairs are relevant to the event via setting a threshold filters out the weights of low correlation pairs, and then fuse audio and visual features by cross-modal residual structure. Finally, the fusion feature is passed to classification modules to implement fully-supervised or weakly-supervised AVE localization.

and positive time residual fusion (PTRF) module. TSGA combines audio and visual features to focus on the spatial region of the visual. Before generating the visual spatial region attention map, TSGA generates a temporal weight matrix for the spatial attention map of the temporal segments by considering the temporal correlation between visual and audio. This weight matrix updates the spatial attention maps of all time segments so that all attention maps focus on the same region where the event may occur. In this way, the subsequent network will learn the changes of specific regions of the visual in consecutive time segments and improve the ability to discriminate between background and event. PTRF encodes the correlation of the dual-modality in the time dimension, filters the low correlation weights through a threshold, and uses the correlation weights to gradually update the visual and audio features via cross-modal residual fusion, and finally merges the dual-modal features.

II. Related Works

Many attention mechanisms have been introduced for audio-visual event localization. Tian et al. [16] first proposed audio-guided visual attention (AGVA), which expands audio features and allows audio and visual features to jointly participate in the generation of spatial attention maps to improve the performance of event region selection. Xu et al. [18] then proposed an audio-guided spatial-channel attention (AGSCA) that introduced an attention mechanism on channel positions before AGVA, and proposed a cross-modality attention mechanism (CMRA) to explore the event correlation of audio and visual features. The spatial location of events is not only determined by the features of the current segment but also related to the features of its adjacent segments. Therefore, Lin et al. [11] proposed to use the transformer mechanism to combine the audio and visual features of the current segment and adjacent segments. Xue et al. [21] focused on modeling spatial and semantic correlations. Unlike AGVA, their co-spatial attention not only selects visual-spatial regions but also updates audio features.

For learning the consistency of dual-modal features, previous methods mainly fuse audio and visual features in pairs. Lin et al. [10] first proposed to combine audio and visual pairs through the Bi-LSTM network, which generates global representations of audio and visual features and fuses them with local representations from CNNs. Wu et al. [17] proposed a dual-attention matching (DAM) module, which aims to obtain more event-related information by looking into longer temporality in the process of fusing paired audio-visual features. But DAM needs supervision signals for event-related prediction. Xuan et al. [19], [20] proposed a cross-modal attention framework, by using adaptive attention to fuse audio and visual features, in which adaptive attention not only effectively matches paired audio and visual features, but also decides to rely on audio or visual according to hidden correlations.

Since the correlation of audio-visual feature pairs is more beneficial to AVE localization, Yang et al. [23] proposed a novel positive sample propagation (PSP) module to filter the features with low audio-visual correlation to ensure that the fusion feature is more relevant to AVE. Hu et al. [7] took a different approach and proposed a Deep Multi-Modal Attention Network, which converts dual-modal features into a multi-modal separator and uses the separator to fuse consistent audio and visual features as the input of the subsequent multi-modal matching classifier. To summerize, although previous methods fully consider the spatial and channel dimensions, they can not use the entire audio and visual sequence to obtain long-term changes in the region where the event is located. Our method is able to capture the region where the event is located in the spatial region localization, and to learn the long-term variation of this region and meanwhile filter low-correlation features in the dual-modal fusion module.

III. METHOD

The overall structure of our proposed method is shown in Figure 2. The visual frames extracted from the video clips are passed into the visual's convolutional neural networks (CNNs) to obtain the feature of the visual segments. The audio clips are converted into Mel spectrograms and then passed into the audio's CNNs to obtain the feature of audio segments. TSGA is used to fuse the features of two modalities to select the spatial region of visual feature which is more related to the event. Then audio feature and new visual feature in time sequence are encoded by Bi-LSTM. Afterwards, the aligned audio and visual features will be passed into PTRF. PTRF encodes features through a dual-branch structure, and then uses them to generate a temporal correlation matrix. The correlation matrix will be filtered by a set threshold to retain the weight of high correlation. The final audio and visual features are combined with the correlation matrix and updated through the residual structure. The fusion feature is obtained by adding the final audio and visual features, and passed to the classification module to predict whether there is an event in each time segment.

A. Time-Spatial Guided Attention

Through CNNs backbone of visual, we get the visual segment features as $V \in \mathbf{R}^{T\times S\times D_v}$, where T is the number of segments (the duration of a segment is one second), D_v is the feature dimension, and $S = W \times H$. Here, W and H are the width and height of feature map. Through CNNs backbone of audio, we get the audio segment features as $A \in \mathbf{R}^{(T\times D)}$. TSGA consists of AGVA and time selection module, see figure 3 for details. The AGVA, proposed by Tian et al. [16], aims to combine audio and visual features to obtain the feature that is more related to event. The feature is encoded into the spatial attention maps of visual to focus on the more critical spatial areas of the visual. However, when the two modal features in same segment are inconsistent, audio feature can not assist in encoding the spatial attention map of the corresponding visual feature. In order to obtain more effective visual feature through spatial attention, we design a spatial temporal guided attention module, which aims to focus on event-related region in all segments and learn the difference between background and event. TSGA adds a temporal selection module to select the attention maps which generated by AGVA and determine the final spatial region of all segments. Because it considers the audio and visual features over the whole time sequence, the final focused spatial region is more effective for AVE localization. Specifically, we average the spatial dimensions of visual features are first averaged as $V_{avg} \in \mathbf{R}^{T\times D_v}$, and the temporal selection module then generates a temporal weight matrix T_m for selecting visual spatial attention maps:

$$V_{vd} = W_{vd}V_{avg} + b \tag{1}$$

$$V_{ft} = W_{vt}V_{fd} + b \tag{2}$$

$$A_{fd} = W_{ad}A + b \tag{3}$$

$$A_{ft} = W_{at}A_{fd} + b \tag{4}$$

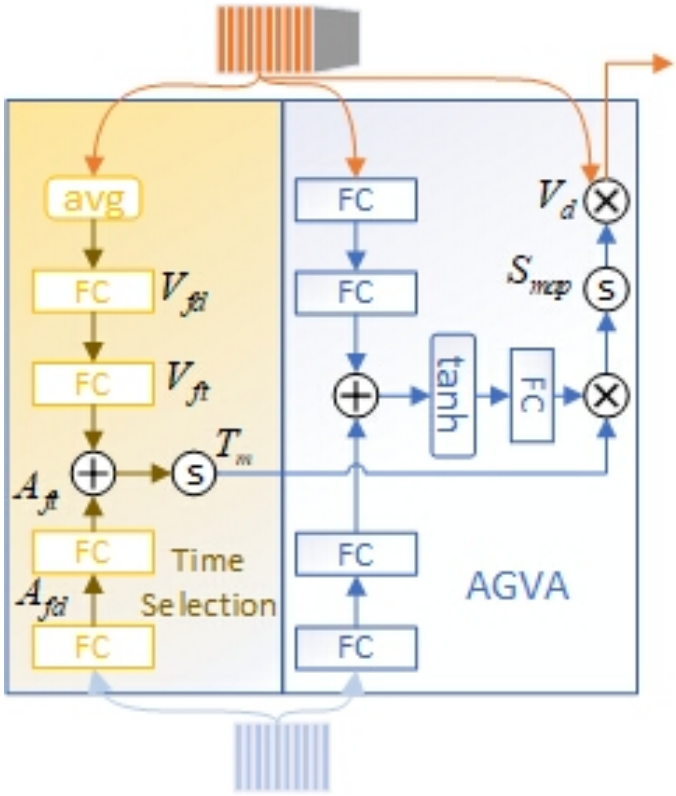


Fig. 3. Illustration of the TSGA module. The time selection module which leverages temporal information and the AGVA module which leverages spatial information together constitute the selection of visual spatial features.

$$T_m = softmax(V_{ft} + A_{ft}) \tag{5}$$

$$V_d = softmax(T_m S_{map})V \tag{6}$$

Where $W_{vd} \in \mathbf{R}^{D_v\times d}$, $W_{ad} \in \mathbf{R}^{D_a\times d}$, W_{vt} and $W_{at} \in \mathbf{R}^{d\times T}$ are learnable parameters of linear transformations, implemented by fully-connected layers. V_{fd}, A_{fd} both remain in the same hidden dimension, where b is the bias of each layer and the activation function of each full-connected layer is ReLu. The temporal weight T_m is obtained through the softmax activation function. The initial values of T_m are random, and only one column has a high value after training. By matrix multiplication with S_{map}, T_m selects the spatial attention map on a specific segment, which focuses on the region where the event is located. The output V_d is obtained by equation (6).

B. Positive Time Residual Fusion

The audio and visual features are paired in time sequence, and each segment has a visual feature and a corresponding audio feature. While audio features can be combined with visual features into pairs, we can obtain the features corresponding to the events in the entire time sequence by modeling the correlation of all pairs. In the dual-modal fusion module, such correlation of two modalities plays a huge role in distinguishing whether audio and visual features are consistent in audio-visual events. Inspired by PSP [23], PTRF also adopts positive

connections and filters the interference of low-correlation audio and visual pairs in the temporal correlation matrix through a pre-specified threshold. The temporal correlation matrix θ_t is obtained by multiplying the features encoded by the two branches of audio and visual. Due to the features between segments on each branch being temporally continuous, this correlation matrix captures the long-term correlation of audio and visual features.

$$C_{at} = W_{at}A_{in} + b \tag{7}$$

$$C_{vt} = (W_{vt}V_{in} + b) \tag{8}$$

$$C_{vh} = BN(tanh(W_{ft}C_{vh} + b)) \tag{9}$$

$$\theta_t = softmax(C_{at}C_{vh}) \tag{10}$$

Where W_{vt} and $W_{at} \in \mathbf{R}^{(d_{in}\times c)}$ are the learnable parameter full-connected layers, d_{in} denotes input feature dimension, the size of c is twice the size of the event categories, representing all events and their corresponding backgrounds to enhance the semantics of features, b is the bias of each layer, and their activation function is ReLu. $W_{ft} \in \mathbf{R}^{c\times c}$ is also a fully-connected layer, which prepares for the subsequent calculation of θ_t, and its activation function is tanh. BN is a batch normalization layer. θ_t needs to be filtered through a threshold ε to ensure that the subsequent fusion feature has better positive propagation, i.e., $\theta_t = \theta_t\delta(\theta_t)$. Here, $\delta(\cdot)$ is an indicator function whose output is 1 when the input is greater than or equal to ε, and 0 otherwise .

Unlike the single-layer structure of the audio branch, the visual branch has three layers. This is because the visual branch preserves event-related features and keeps background-related features at low values, to ensure that the correlation matrix multiplied by the audio branch is more sensitive to events.

After that, the audio and visual features are updated by residual structure. First, we multiply the audio feature with θ_t to get a more event-relevant positive feature, which is added to the visual feature to get the final visual feature V_{paf}. Then we choose V_{paf} which is more consistent with audio feature to fuse with audio feature in the same way to get the final audio feature A_{paf}. This residual method can ensure that each updated feature has more positive information and the final audio and visual features have higher correlation.

$$V_{paf} = (W_{vp}V_{in} + b) + \theta_t(W_{ap}A_{in} + b) \tag{11}$$

$$A_{paf} = (W_{ap}A_{in} + b) + \theta_t^T V_{paf} \tag{12}$$

$$F_{paf} = V_{paf} + A_{paf} \tag{13}$$

Where W_{vp} and $W_{ap} \in \mathbf{R}^{d_{in}\times d_{in}}$ are the parameters of the full-connected layers. Finally, the fusion feature F_{paf} is obtained by adding the dual-modal features. Due to the high correlation of the dual-modal features, the fusion feature has higher values on the segments with high audio-visual consistency.

C. Localization Classifier

AVEL can be done in either fully-supervised or weakly-supervised manner. The fully supervised AVEL gives the category labels of all time segments of the training sample, while the weakly supervised AVEL only provides the entire time sequence of event categories [16]. In our method, different localization classifiers are employed for these two types of AVEL tasks.

Fully Supervised AVEL. Our final AVE localization is achieved through classification. After two fully connected layers, the output is $Y_{fully}^{out} \in \mathbf{R}^{T\times C_{cate}}$, where C_{cate} is the number of all categories. Note that all backgrounds are divided into one category. The loss function of the final classifier is the cross entropy (CE) loss.

Weakly Supervised AVEL. Compared with the fully-supervised classifier, the weakly-supervised classifier has one more layer of average pooling, because the weakly-supervised ground truth lacks the label of the time segment. Hence the final output of the weakly-supervised classifier is $Y_{weak}^{out} \in \mathbf{R}^{C_{cate}}$. Following [18], [23], we adopt the binary cross entropy (BCE) loss.

IV. Experiments

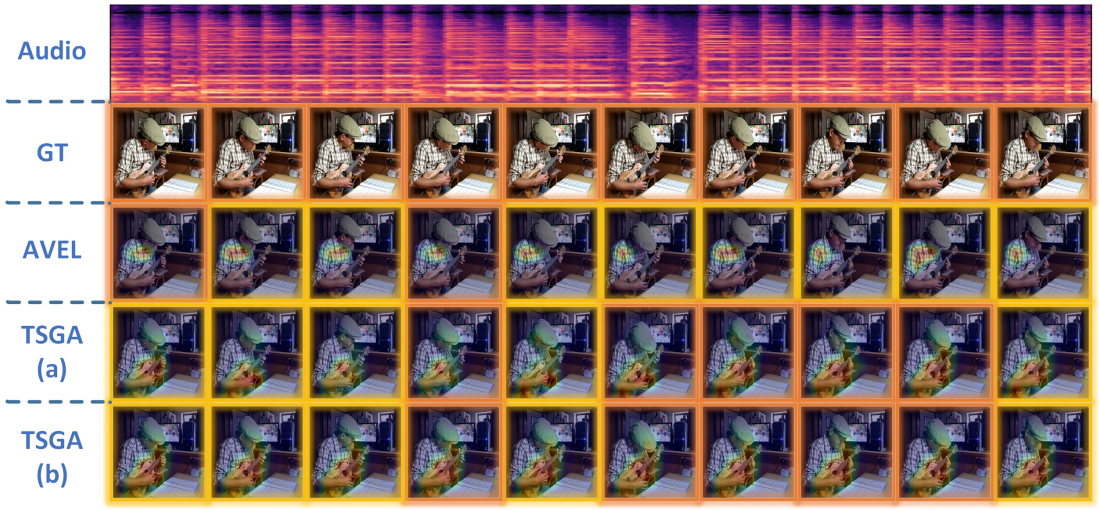
A. Dataset

We evaluate our method on AVE dataset [16] that is publicly available. It provides 4143 videos with complex scenes and noisy training conditions. These video contents come from daily life, including 28 event categories (e.g., motorcycle, banjo, goat, male speech, etc.) and a background category contains the background of all events. Each video is 10 seconds in length and is equally divided into 10 segments. 80% of the data is used for training, 10% for validation, and 10% for testing.

B. Implementation Details

Feature Extractor. Before extracting the features of the video, we sample each one-second segment with 16 frames. These frames are extracted by VGG-19 [15] pre-trained on Imagenet [9] to extract visual features. Before extracting the features of the video, we sample each one-second segment with 16 frames. The frame-level features are the outputs of the $pool5$ layer in VGG-19 with dimensions of $7 \times 7 \times 512$. We average the features of the 16 frames to get the segment-level features. For audio, we first transform the raw audios into logmel spectrograms and then use the VGG-like [6] network pretrained on AudioSet [4] to extract their acoustic features with dimensions of 128 for each segment.

Training Settings. We use dropout technique in the linear layers of PTRF. The probability of an element to be zeroed in dropout technique is 0.2. In the fully-supervised task, we set the threshold ε to 0.0599, and in the weakly-supervised task, we set the threshold ε to be the mean of the result of multiplying C_{vt} and C_{at}, i.e., $mean(C_{vt}C_{at})$. The batch size is 128. We apply Adam [8] as an optimizer. We set the initial learning rate as 0.001.



(a) Example attention results of spatial localization

(b) Example attention results of Audio-visual Fusion

Fig. 4. Visualization analysis. Orange boxes indicate that the segment is an audio-visual event; green boxes indicate that the segment is the background; yellow boxes indicate that the segment prediction does not match the ground truth.

TABLE I
EVENT LOCALIZATION ACCURACY (%) OF DIFFERENT METHODS IN BOTH FULLY-SUPERVISED AND WEAKLY-SUPERVISED SETTINGS.

Method	Fully-supervised	Weakly-supervised
AVEL [16]	72.7	66.7
AVRB [13]	74.8	68.9
AVIN [12]	75.2	69.4
AVSDN [10]	75.4	74.2
AVT [11]	76.8	70.2
CMRAN [18]	77.4	72.9
CMAN [20]	77.1	75.7
MTNLI [5]	77.9	**76.2**
LSSC [21]	76.5	70.2
PSP [23]	77.8	73.5
Ours	**78.3**	72.1

C. Quantitative Analysis

1) Comparison with state-of-the-arts: We apply our audio-visual long-term correlation network in supervised and weakly-supervised AVE localization. For fair comparisons, we compare our method with existing methods that fulfill both fully-supervised and weakly-supervised AVE localization tasks. The results of these state-of-the-art methods come from their published papers. The comparison results are shown in Table I. As can be seen, our method achieves the highest accuracy in the fully-supervised task, and is higher than the baseline model in [16] by 5.6%. Our method is also effective in weakly-supervised tasks, achieving an accuracy 5.4% higher than that of the baseline model. In the fully-supervised task, our network better handles the associated features distributed in the time sequence, but in weakly-supervised tasks, the ability of the model to capture long-term related features is weakened due to the inability of supervised signals to give segment-level positive feedback.

2) Ablation studies: To investigate the effectiveness of different components of our model, we conduct ablation studies

TABLE II
ABLATION STUDY RESULTS (%) OF OUR METHOD.

AGVA	TSGA	PTRF	Fully	Weakly
✓			75.0	70.5
✓		✓	77.8	71.4
	✓		76.1	71.2
	✓	✓	**78.3**	**72.1**

of main components in both fully supervised task and weakly supervised task. In the experiments, we keep the same settings for each model. The role of AGVA and TSGA in the model is to select the feature of the video in the spatial area, so we verify the effectiveness of TSGA by comparing them. As can be seen from the results in Table II, regardless of whether PTRF is added or not, the performance of TSGA is better than that of AVGA. We add V_{in} and A_{in} directly in those network models without PTRF. As shown in Table II, the fusion strategy of PTRF can significantly improve the accuracy of the aduio-visual event localization.

In PTRF, the threshold ε determines which weights in the time correlation matrix are filtered. Table III shows the influence of different ε on the model. We range from 0 to 0.2 for ε. We found that when ε is 0.0599, the accuracy of the fully-supervised model is the highest. For the weakly-supervised model, due to the lack of time sequence segment tags, the time sequence correlation information obtained by the model is not sufficient, and the relation between each segment is not stable enough in the iterative process. Therefore the externally fixed threshold is not as effective as the threshold learned by the model itself.

TABLE III
RESULTS (%) OF OUR METHOD UNDER DIFFERENT THRESHOLD ε

	threshold ε					
	0.000	***0.0599***	***0.0099***	***0.1599***	***0.1999***	$mean$
fully	76.3	**78.3**	77.6	75.2	76.3	76.3
weakly	71.1	71.8	71.0	69.7	71.3	**72.1**

D. Qualitative Analysis

We visualize the results of our model and the baseline AVEL model in Figure 4. The attention map used inTSGA(a) is S_{map} in TSGA. The attention map in TSGA(b) is obtained by multiplying S_{map} and T_m.

Spatial Localization. For the audio-visual event in Figure 4(a), the ukulele playing takes place on the hands and on the ukulele. However, the spatial position where AVEL focuses on is inaccurate and too small (only on the player's shoulder which is above the location of the audio-visual event). In contrast, our method is more accurate in spatial localization and the region of interest almost exactly corresponds to the location of the entire event. Comparing AVEL and TSGA(a) in Figure 4(a), We find that AGVA performs better in TSGA than in AVEL. As stated above, S_{map} is able to capture more event-related information in our method. This suggests that our learned long-term correlations can tune the S_{map} to improve the accuracy of spatial region localization. Not only that, TSGA(b) in Figure 4(a) which is the final attention map of TSGA aligns the attention regions of each segment of the video. Besides not only ensures that the model can capture the region related to audio-visual events but also allows the model to learn continuous changes in the region.

Audio-visual Fusion. Singe modality cannot accurately identify audio-visual events. As shown in Figure 4(b), the audio events of the helicopter occur all the time but the visual events of the helicopter only appear in the last three segments. Therefore, the fusion stage not only needs to obtain event-related features but also needs to identify whether the audio and video match. Figure 4(b) shows that AVEL captures the event-related features of the audio, but fails to recognize that the video and audio mismatch. However, our method performs better than AVEL. In the segments of the audio-visual event, our method locates the area of the helicopter. And in the background segments, our method can effectively fuse the audio and visual features and correctly recognize the mismatch of the audio and visual. In addition, our method expects to detect the distinction between background and event by observing continuous changes in specific regions where audio-visual events occur. As can be seen in Figure 4(b), in these segments which are background, we still pay attention to the special area where the helicopter occur in the event segments.

V. CONCLUSIONS

In this paper, we propose an audio-visual long-term correlation network that incorporates TSGA and PTRF to learn long-term changes in the region where the event is located to improve accuracy of AVE localization. In TSGA, we exploit long-term correlations across time sequence to select region of visual feature. In PTRF, we generate a temporal correlation matrix to assist in updating audio and visual features. Our network performs well on both fully supervised and weakly supervised AVE localization tasks on the AVE dataset, especially on fully supervised tasks.

ACKNOWLEDGMENT

This work is partially supported by the National Natural Science Foundation of China (No. 62176170, 62166038), and Tibetan information processing and Machine Translation Key Laboratory of Qinghai province and Key Laboratory of Tibetan information, Ministry of Education (No. 2020Z001).

REFERENCES

[1] Egils Avots, Tomasz Sapinski, Maie Bachmann, and Dorota Kaminska. Audiovisual emotion recognition in wild. *Mach. Vis. Appl.*, 30(5):975–985, 2019.

[2] David A Bulkin and Jennifer M Groh. Seeing sounds: visual and auditory interactions in the brain. *Current opinion in neurobiology*, 16(4):415–419, 2006.

[3] Bin Duan, Hao Tang, Wei Wang, Ziliang Zong, Guowei Yang, and Yan Yan. Audio-visual event localization via recursive fusion by joint co-attention. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 4012–4021. IEEE, 2021.

[4] Jort F. Gemmeke, Daniel P. W. 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 776–780. IEEE, 2017.

[5] Yixuan He, Xing Xu, Xin Liu, Weihua Ou, and Huimin Lu. Multimodal transformer networks with latent interaction for audio-visual event localization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[6] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 131–135. IEEE, 2017.

[7] Ruihan Hu, Songbin Zhou, Zhi-Ri Tang, Sheng Chang, Qijun Huang, Yisen Liu, Wei Han, and Edmond Qi Wu. DMMAN: A two-stage audio-visual fusion framework for sound separation and event localization. *Neural Networks*, 133:229–239, 2021.

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012.

[10] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 2002–2006. IEEE, 2019.

[11] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In Hiroshi Ishikawa, Cheng-Lin Liu, Tomás Pajdla, and Jianbo Shi, editors, *Computer Vision - ACCV 2020 - 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 - December 4, 2020, Revised Selected Papers, Part VI*, volume 12627 of *Lecture Notes in Computer Science*, pages 274–290. Springer, 2020.

[12] Janani Ramaswamy. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 4372–4376. IEEE, 2020.

[13] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 2959–2968. IEEE, 2020.

[14] Leda Sari, Kritika Singh, Jiatong Zhou, Lorenzo Torresani, Nayan Singhal, and Yatharth Saraf. A multi-view approach to audio-visual speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 6194–6198. IEEE, 2021.

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[16] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, volume 11206 of *Lecture Notes in Computer Science*, pages 252–268. Springer, 2018.

[17] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6291–6299. IEEE, 2019.

[18] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 3893–3901. ACM, 2020.

[19] Hanyu Xuan, Lei Luo, Zhenyu Zhang, Jian Yang, and Yan Yan. Discriminative cross-modality attention network for temporal inconsistent audio-visual event localization. *IEEE Trans. Image Process.*, 30:7878–7888, 2021.

[20] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 279–286. AAAI Press, 2020.

[21] Cheng Xue, Xionghu Zhong, Minjie Cai, Hao Chen, and Wenwu Wang. Audio-visual event localization by learning spatial and semantic co-attention. *IEEE Transactions on Multimedia*, 2021.

[22] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9299–9306. AAAI Press, 2019.

[23] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8436–8444. Computer Vision Foundation / IEEE, 2021.

[24] Hao Zhu, Mandi Luo, Rui Wang, Aihua Zheng, and Ran He. Deep audio-visual learning: A survey. *Int. J. Autom. Comput.*, 18(3):351–376, 2021.