Dense Modality Interaction Network for Audio-Visual Event Localization

Shuo Liu, Weize Quan, Chaoqun Wang, Yuan Liu, Bin Liu, and Dong-Ming Yan

***Abstract*—Human perception systems can integrate audio and visual information automatically to obtain a profound understanding of real-world events. Accordingly, fusing audio and visual contents is important to solve the audio-visual event (AVE) localization problem. Although most existing works have fused audio and visual modalities to explore their relationship with attention-based networks, we can delve into their relationship more deeply to improve the fusion capability of the two modalities. In this paper, we propose a dense modality interaction network (DMIN) to elegantly leverage audio and visual information by integrating two novel modules, namely, the audio-guided triplet attention (AGTA) module and the dense inter-modality attention (DIMA) module. The AGTA module enables audio information to guide the network to pay more attention to event-relevant visual regions. This guidance is conducted in the channel, temporal, and spatial dimensions, which emphasize informative features, temporal relationships and spatial regions, to boost the capacity of representations. Furthermore, the DIMA module establishes the dense-relationship between audio and visual modalities. Specifically, the DIMA module leverages the information of all channel pairs of audio and visual features to formulate the cross-modality attention weight, which is superior to the multi-head attention module that uses limited information. Moreover, a novel unimodal discrimination loss (UDL) is introduced to exploit the unimodal and fused features together for more exact AVE localization. The experimental results show that our method is remarkably superior to the state-of-the-art methods in fully- and weakly-supervised AVE settings. To further evaluate the model's ability to build audio-visual connections, we design a dense cross modality relation network (DCMR) to solve the cross-modality localization task. DCMR is a simple deformation of a DMIN, and the experimental results further illustrate that DIMA can explore denser relationships between the two modalities. Code is available at https://github.com/weizequan/DMIN.git.**

***Index Terms*—Multi-modality; Audio-visual event localization; Dense modality interaction; Attention**

I. Introduction

INSPIRED by the multi-modality perception property of human beings [1], [2], machine perception is remarkably improved by transferring from single-modality learning to multi-modality learning, with the significant advances in vision, speech, and language processing [3], [4]. The fusion of the two most important and prevalent modalities, namely, the audio and visual modalities, has attracted extensive attention from many communities. Audio-visual learning has various applications, such as audio-visual separation [5], [6], [7], audio-visual matching [8], [9], [10], speech recognition [11], [12], [13], [14], [15], [16], audio-visual generation [17], [18], [19], [20], and audio-visual event (AVE) localization [21], [22], [23], [24], [25].

AVE localization means localizing the events temporally while identifying the corresponding category, which can facilitate video understanding. An AVE is defined as the audible and visible event in a video segment [21] (see Fig. 1).

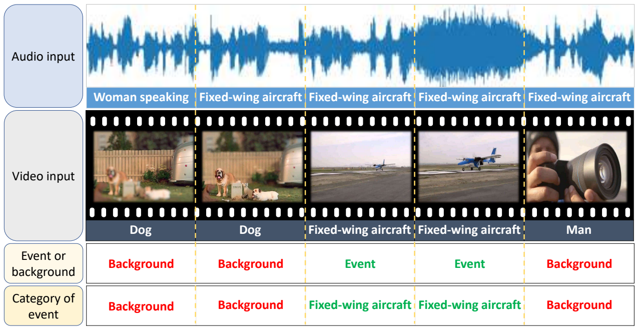


Fig. 1. An example of the audio-visual event localization. The goal of AVE localization is to predict the event label for each segment of a given video sequence. An event that is both audible and visible is regarded as an audio-visual event ("Fixed-wing aircraft" in the third and fourth segments). On the contrary, a segment is not both audible and visible, which is predicted as background. In addition, the asynchronous situation makes the AVE localization more challenging, for example, the sound of the fixed-wing aircraft is heard in the second segment, while the appearance of fixed-wing aircraft is seen in the third and fourth segments.

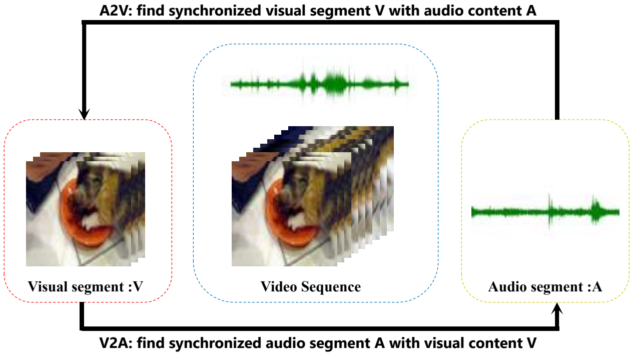

A2V: find synchronized visual segment V with audio content A
Visual segment :V
Video Sequence
Audio segment :A
V2A: find synchronized audio segment A with visual content V

Fig. 2. Illustration of cross-modality localization. For A2V, we want to find its synchronized K-second visual segment by giving a K-second audio segment. For V2A, we want to find its synchronized K-second audio segment by giving a K-second visual segment.

S. Liu, W. Quan, C. Wang, B. Liu, and D.-M. Yan are with NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: yandongming@gmail.com). *(Shuo Liu and Weize Quan are co-first authors.) (Corresponding author: Dong-Ming Yan.)*

Y. Liu is with Speech Lab, Alibaba Group.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes a Supplementary Material, which provides more comparisons and analysis of our method. This material is 414KB in size.

Several difficulties and challenges are experienced in the AVE localization problem. First, noise sound, such as ambient sound and target event sound, exists in video. Second, sound sources are out-of-screen (*i.e.*, asynchrony), thereby causing difficulty in establishing the exact connections between objects and sounds. Third, learned features among similar categories, *e.g.*, guitars, ukuleles, and mandolins, tend to be very close in the feature space, thereby causing difficulty in distinguishing the corresponding event objects. The core of solving these issues is to address how to fuse audio and visual information effectively to obtain the correlation of the two modalities. The correlation of two modalities is more elegant, the higher localization performance will be obtained.

Most existing methods use deep-learning-based models with various attention schemes [21], [26], [27], [22], [23], [24], [25]. These methods often contain some or all of the following components: audio-guided visual attention (AGVA), intra-modality attention, cross-modality attention, and audio-visual feature fusion. However, several limitations affect their localization accuracy. First, although many works have applied audio-guided attention to obtain better visual features, their fusion computation is usually coarse-grained. That is, they only consider spatial attention [21], [27] or spatial-channel attention [28]. The representation capability of visual features could be enhanced by conducting fine-grained attention in the channel, temporal, and spatial dimensions appropriately. Second, intra- and inter-modality relationships are modeled with the multi-head attention module [29]. Unfortunately, the computation of the correlation weight mainly follows the traditional sparse dot-product operation, which cannot fully exploit the dense relationships between audio and visual representations. Third, for network training, previous methods overemphasize the fused features of audio and visual modalities (only adding loss functions in the final classifiers), but ignore the classification capability of the unimodal features. Intuitively, one modality feature may have larger differences when another modality feature is very similar to the close categories.

To enhance the accuracy of AVE localization, in this paper, we propose a dense modality interaction network (DMIN) to elegantly integrate audio features with visual features by introducing two novel modules, namely, the audio-guided triplet attention (AGTA) module and the dense inter-modality attention (DIMA) module. When the audio features are related to visual features, with the exact guidance of audio signals, the AGTA module aims to improve the representation capability of visual features by highlighting event-relevant visual regions while reducing the interference from background regions or irrelevant objects. When the audio features are not perfectly in sync with the visual features, the final audio-guided attention can avoid the event-relevant visual regions and focus on event-irrelevant regions. Specifically, the guidance of the AGTA module is applied in the channel, temporal, and spatial dimensions.

Cross-modality relation attention (CMRA) [28] uses the multi-head attention module, whose query originated from one modality while the key and value originated by temporally concatenating two modalities. The motivation is to exploit the cross-modality relations, while not neglecting the intra-modality relation information. However, the two parts of the key use the same classical sparse correlation weight computation. As a result, the query prefers to leverage the information of their own modality, thereby hindering the fusion of two modalities. To solve this problem, we propose a DIMA module to model the correlation between audio and visual representations in a dense modality interaction manner. A novel dense attention block with a full channel pair product replaces the multi-head attention to obtain the dense inter-modality correlation weight. This block formulates all channel pairs of audio and visual features to explore more fine-grained attention computation. Intuitively, all channel pairs of audio and visual features can provide information that is richer than the limited channel pairs, which will be more helpful for exploring inter-modality relationships.

Audio-visual fused features may destroy some originally useful information in unimodal features. Several existing works have followed similar ideas that add constraints to the middle-level features. Hou *et al.* [30] propose an audio-visual deep CNN with a multi-task learning framework to enhance the speech. During training, the auxiliary visual information at the output layer serves as part of the constraints. Li *et al.* [31] design a two-stream network with additional emotion constraints on video and audio branches for cross-modal music retrieval. Vielzeuf *et al.* [32] and Fayek *et al.* [33] add loss functions to unimodal and multimodal predictions and then propose specific strategies for combining multiple predictions to obtain the final prediction. However, this is nontrivial as stated in [34]. Based on the above insights, we propose a unimodal discrimination loss (UDL) to emphasize the localization capability of unimodal features without introducing additional tasks or designing a complex combination strategy for multiple predictions. Specifically, we simply add a category classification loss in the middle-level visual features before the cross-modality fusion, and this loss is combined with the common event localization losses to train the model in an end-to-end manner.

To further validate the superiority of DIMA at exploring the cross-modality correlation, we design a dense cross modality relation network (DCMR) to solve the cross-modality localization (CML) task. As shown in Fig. 2, CML aims to search for the synchronized segment of one modality by giving the segment of the other modality, in other words, to build the bridge from one modality (audio/visual features) to the other modality (visual/audio features). Different from the AVE localization task, no semantic (or event category) label is provided in CML. Tian *et al.* [21] design an audio-visual distance learning network based on the two modality features, and Wu *et al.* [27] apply a global representation of one modality to check every segment of the other modality for localization. However, these two methods do not exploit the dense inter-modality relationships and thus have limited cross-modality localization performance.

The main contributions of our work are summarized as follows:

- We develop an *audio-guided triplet attention* module to enhance the representation capability of visual features with fine-grained audio guidance in the channel, temporal, and spatial dimensions.

- We propose a *dense inter-modality attention* module to elegantly fuse the audio and visual modalities via a dense fusion attention block with a full channel pair product.
- We introduce a *unimodal discrimination loss* to emphasize the ability of unimodal features. This loss is combined with the common event localization losses to allow the network to simultaneously explore the localization capabilities of integrated features and unimodal features.
- We devise a *dense modality interaction network* for AVE localization that combines the three aforementioned modules. The experimental results on the publicly available AVE dataset show that our method achieves state-of-the-art performance in fully- and weakly-supervised settings.
- We design a *dense cross modality relation network* for the cross-modality localization task. DCMR is a simple deformation of a DMIN. In DCMR, the AGTA module and the audio-visual fusion module are removed from the DMIN, and a cross-matching mechanism is introduced. Our network achieves superior performance.

II. Related Work

A. Attention Mechanisms

An attention mechanism imitates the human perception system to capture long-range dependencies automatically and highlight the critical part of input signals selectively. Hu *et al.* [35] propose a channel-wise attention mechanism that considers the global information of each channel to select the meaningful feature maps and suppress the others. The mechanism is intended to model the inter-dependencies between the channels of its spatial features. Woo *et al.* [36] further combine channel-wise and spatial attention and verify that the use of both is superior to only utilizing channel-wise attention. Chen *et al.* [37] introduce a convolutional neural network that embeds spatial and channel-wise attentions for image captioning. Vaswani *et al.* [29] present a self-attention mechanism to acquire the global dependencies between the input and the output, which greatly improves machine translation performance. Devlin *et al.* [38] extend the self-attention mechanism for pre-training word embedding and achieve good performance. Wang *et al.* [39] utilize a self-attention mechanism in the vision domain to attempt to capture the pixel-level long-range dependencies in spatial and time dimensions via a non-local (NL) operation. Although previous works have paid attention to capturing long-range dependencies in channel, temporal, and spatial dimensions, there is no module simultaneously combining attention in all these dimensions. Consequently, we propose an AGTA module that conducts attention operations in channel, temporal, and spatial dimensions. Moreover, the encoder of the transformer architecture is applied to explore the intra-modality relationship in our work.

B. Audio-Visual Event Localization

The goal of AVE localization is to localize a visible and audible event and identify its category in unconstrained videos. Tian *et al.* [21] first propose an audio-guided visual attention mechanism (AGVA) in the spatial dimension to guide a network for visual modeling. Subsequently, they fuse the temporally modeled audio and visual features via a dual multimodal residual network. Lin *et al.* [26] introduce a sequence-to-sequence dual network that first extracts the local and global features of videos and then feeds these features into a LSTM (long short-term memory) to solve the event localization task. Instead of fusing audio and visual representations at the local segment level, Wu *et al.* [27] propose a dual attention matching method to compute the relevance of events between the global features of one modality and the local features of another modality in a bi-directional manner. Ramaswamy [22] explores the inter- and intra-modality interactions simultaneously via an attention scheme and then concatenates these features for event localization. To improve the discrimination capacity of fused features, Ramaswamy [40] combines LSTM-based fusion and the multi-modal factorized bilinear (MFB) pooling method [41]. Xuan *et al.* [23] propose a three-stage attention-based framework that includes spatial, sequential, and cross-modality attention modules. Lin *et al.* [24] develop an audio-visual transformer and instance-level cross-modality attention to localize audio-visual event. Zhou *et al.* [25] propose a positive sample propagation (PSP) module to fuse cross-modality by emphasizing highly similar audio and visual features while filtering out irrelevant features. Different from previous methods [21], [27] that conducted audio-guided visual attention in the spatial dimension, Xu *et al.* [28] apply attention computations in the channel and spatial dimensions. They also introduced a relation-aware module to build connections between audio and visual modalities by exploiting both intra- and inter-modality information jointly.

Previous works have proposed some audio-guided visual attention mechanisms to capture sound sources in visual regions for better performance. However, these audio-guided visual attention modules are subject to limited dimensions. Instead, our proposed audio-guided triplet attention (AGTA) is processed in channel, temporal, and spatial dimensions with MFB-based fusion attention for fine-grained enhanced visual features guided by audio features. In addition, existing works have utilized collaborative attention, LSTM-based attention, and the encoder of a transformer to explore the inter-modality relationships for more representative fusion features. Unfortunately, these methods are somewhat coarse-grained by simply and directly using the features of the two modalities. Therefore, we design a dense inter-modality attention (DIMA) based on dense fusion attention (DFA) to fuse audio and visual features delicately for better localization performance. Moreover, previous works ignore maintaining the localization ability of unimodal features. However, our proposed unimodal discrimination loss can further enhance the localization capability of network.

C. Cross-Channel Correlations

In this work, we propose dense fusion attention based on a full channel pair product to model dense inter-modality relationships. In the following, we review several existing works that focus on the cross-channel correlations and highlight the differences with our method.

Yue *et al.* [42] extend the NL operation [39] to compute the correlations between any two positions across the channels that yield the generalized non-local (GNL) operation. In particular, they collapse the elements in all channels, spatial and temporal positions in one dimension and then use a general kernel function to formulate a pairwise weight matrix. Essentially, they update the feature with each element as the key, which may introduce potential confusion, whereas our DIMA keeps the original feature architecture with the feature vector as the key. Kuo *et al.* [43] propose a fully generalized non-local (FGNL) operation to solve the singer identification task. Specifically, they propose the FGNL operation extended among all of the elements across channels and layers to obtain richer features. Concatenating the feature matrices from different layers and subsequently rolling the matrices along the channel axis with the classical attention computation yields concatenated features. Moreover, a modified squeeze-and-excitation scheme is utilized to highlight the correlated feature channels. However, the processing of full channel dense correlations is apparently different: FGNL divides these correlations into multiple groups, and each group is independently processed to update the features many times, whereas our DFA sums up these correlations with different weights and then updates the features at once, which is simpler and more effective. In addition, the post-processing operation for restoring the original channel dimensions may cause confusion in FGNL. However, DIMA does not have a similar operation to keep the original feature architecture and enhance the semantic representation. Hsienh *et al.* [44] propose a so-called co-attention mechanism, where the query and key are essentially different compared with traditional self-attention. In addition, this co-attention is realized via a non-local operation. That is, the classical dot product is used to compute the correlation, which is obviously different from our DFA based on the full channel pair product.

D. Cross-Modality Localization

The aim of the cross-modality localization task is to evaluate a model's capability of utilizing the audio-visual relationship. The task aims to find the synchronized segment of one modality from the other modality. Tian *et al.* [21] propose the AVLN method, which measures the correlation of the extracted features from two modalities based on the simple Euclidean distance. The DAM [27] utilizes the global features by watching a long event sequence and then checks each segment of the other modality to predict the event relevance. The AVLN method is relatively simple to explore the cross-modality relationship. The DAM method temporally averages the query features that result in a confused global feature, and the cross-check mechanism is also somewhat coarse-grained. Therefore, we propose a dense cross modality relation (DCMR) network to elegantly explore the denser cross modality information by using a DIMA module. Furthermore, we average the features in the channel dimension instead of the temporal dimension, which is then used in a more fine-grained cross-check mechanism.

E. Multimodal Fusion Method

Three types of methods, namely, simple operation-based, attention-based, and bilinear pooling-based methods, are used to fuse multimodal representations.

Simple operation-based methods usually fuse multi-modal features via simple operations, such as addition, weighted sums [45], element-wise multiplication or concatenation [46], [47], [48], [32], [49]. An obvious advantage of these methods is that only a few or no parameters are often needed to conduct learning. These methods have been widely used in previous works on AVE localization.

Attention-based methods mainly update a feature via the weighted sum of a set of features with scalar weights, which are computed by modeling a certain correlation between two features. A classical self-attention-based method, *i.e.*, the transformer, is proposed by Vaswani *et al.* [29] to model the long-distance dependencies among words. Based on the transformer, Devlin *et al.* [38] introduce the well-known BERT (bidirectional encoder representations from transformers) for pre-training language representations. LXMERT [50] learn the intra-modality features for each modality by using independent encoders and learn the inter-modality features by using additional cross-attention encoder. OmniNet [51] fuse current modality features with those of other modalities by exploiting a gated multi-head attention model embedded in each decoder block. Recently, Xu *et al.* [28] combine self-attention and cross-attention into a multi-head attention module [29] to simultaneously model intra- and inter-modality relationships. However, the same correlation weight computation in this module causes the query modality to opt for the information of their own modality, which hampers the fusion of the two modalities. We introduce a dense fusion attention to enhance the audio-visual features and align the two modalities for better performance.

Bilinear pooling-based methods learn the joint representation space of two-modality feature vectors. Considering the high number of parameters in the project matrix of bilinear pooling, some approximation methods seeking to obtain compact bilinear representations have been proposed. The multimodal compact bilinear (MCB) pooling method combines the two-modality vectors by projecting them to higher dimensional space randomly and then convolving them using element-wise multiplication in the fast Fourier transform space [52]. To reduce the memory requirement of the MCB pooling method, Kim *et al.* [53] propose a low-rank bilinear pooling (MLB) method. The MLB method first uses linear mapping to project the two modality features into the same low-dimensional space and then applies element-wise multiplication and nonlinear activation to obtain fused features. Subsequently, Yu *et al.* [41] propose the MFB pooling method, which extends the MLB method with an expansion-and-squeeze operation. In this work, we utilize MFB in our AGTA module to model the fine-grained relationship between audio and visual features.

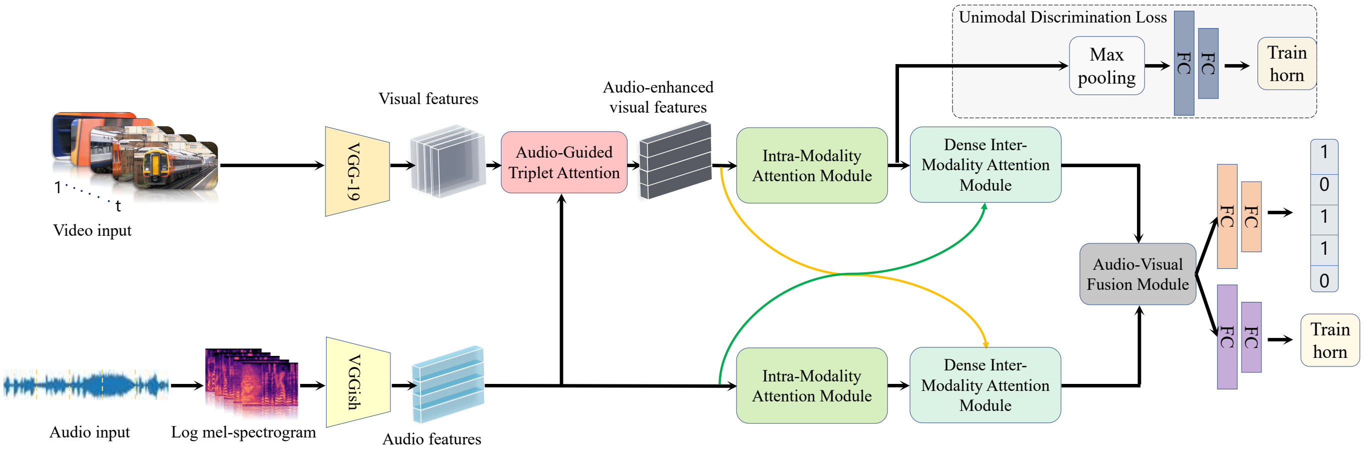

Unimodal Discrimination Loss
Max pooling
FC
FC
Train horn
Visual features
Audio-enhanced visual features
1 · · · · · · t
Video input
VGG-19
Audio-Guided Triplet Attention
Intra-Modality Attention Module
Dense Inter-Modality Attention Module
Audio-Visual Fusion Module
FC
FC
1
0
1
1
0
FC
FC
Train horn
Audio input
Log mel-spectrogram
VGGish
Audio features
Intra-Modality Attention Module
Dense Inter-Modality Attention Module

Fig. 3. The architecture of our proposed dense modality interaction network. The pre-extracted audio and visual features are given to the network. A novel audio-guided triplet attention is applied to enhance the visual features via the fine-grained attention conducted in the channel, temporal, and spatial dimensions. An intra-modality attention module is then used to exploit the intra-modality relationship. Afterwards, a dense inter-modality attention module is designed to elegantly explore the inter-modality relationship with a dense correlation weight computation. An audio-visual fusion module is finally adopted to learn the joint representation for the AVE localization. Meanwhile, we introduce a unimodal discrimination loss to enhance the classification capability of visual modality. In the inference stage, we use the event-relevant label and event category label from the two-modality fused features as the final prediction.

III. Problem Formulation

A. Notations

A video sequence with T non-overlapping segments is represented as $X = (V_t, A_t)_{t=1}^{T}$, where each segment lasts one second; and V_t and A_t represent the visual and audio components of the t-th segment, respectively.

B. Audio-Visual Event Localization

The audio-visual event localization task aims to identify whether a video segment contains an audio-visual event and the category of the event. An audible and visible event is an audio-visual event; otherwise, the event is defined as background.

In a fully-supervised setting, for each segment X_t, we can access the detailed segment-level event label $y_t = \{y_t^c \mid y_t^c \in \{0, 1\}, c = 1, ..., C, \sum_{c=1}^{C} y_t^c = 1\}$, where C is the total number of event categories plus one background label.

In a weakly-supervised setting, only the video-level label is provided for training, although we still intend to predict the segment-level labels in the testing stage. Specifically, a video-level label represents what event category is contained in the video.

C. Cross-Modality Localization

The goal of the cross-modality localization task is to predict the position of the synchronized content in one modality (visual/audio) with a segment of the other modality (audio/visual) [21]. The core of this task is to exploit the audio-visual correlations in the temporal dimension. Dense cross-modality relation information can be utilized to boost the performance.

For visual localization from audio (A2V), we use a K-second ($K < T$) audio segment $\tilde{A}$ from $\{(A_t)\}_{t=1}^{T}$ to localize its synchronized K-second visual segment within $\{(V_t)\}_{t=1}^{T}$. Similarly, for audio localization from a visual segment (V2A), we use a K-second ($K < T$) visual segment $\tilde{V}$ from $\{(V_t)\}_{t=1}^{T}$ to find its synchronized K-second audio segment within $\{(A_t)\}_{t=1}^{T}$. In this task, only event-relevant labels without event categories are provided.

IV. Audio-Visual Event Localization

The network architecture of our method is shown in Fig. 3. Given a video sequence, VGG-19 [54] pre-trained on ImageNet [55] is used to extract the visual features from the video input, and VGGish [56] pre-trained on AudioSet [57] is utilized to abstract the audio features from the log mel-spectrogram of the audio input. This feature extraction process is also the same as that of the state-of-the-art methods [21], [27], [28]. To improve the representation capability of the visual features, we propose a novel AGTA module, which takes the visual features as input and outputs their enhanced version with fine-grained guidance from the audio features. Then, symmetric intra-modality and dense inter-modality attention blocks are applied. Specifically, the audio features and enhanced visual features are respectively fed into two self-attention modules to model the intra-modality relationship. Next, the self-attentive audio features and enhanced visual features conduct the inter-modality fusion with our proposed dense inter-modality attention module. The DIMA module implements dense fusion attention by using the full channel pair product to delve into the relationship of the two modalities more deeply. In addition, the self-attentive visual features and the audio features perform the above process similarly. Finally, an audio-visual fusion module is utilized to fuse the cross-attentive visual and audio features to achieve joint representation. Similar to [27], the localization task is decoupled as two sub-tasks, namely, event relevance prediction and event category prediction, which are accomplished with the two multi-layer perceptions. Furthermore, a unimodal

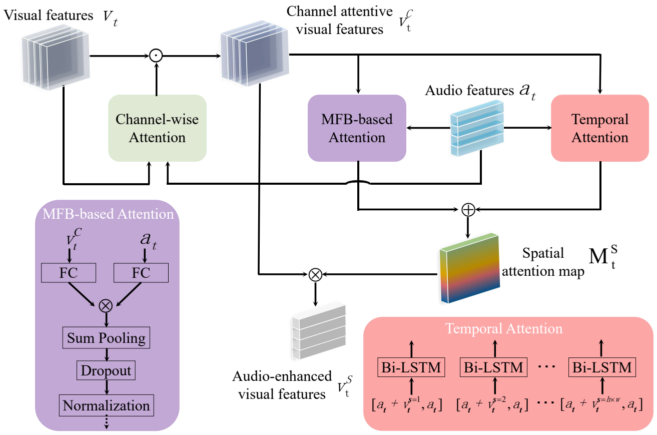


Fig. 4. Overview of our proposed AGTA module, which conducts the audio guidance with the attention in the channel, temporal, and spatial dimensions. $\odot$ stands for the element-wise multiplication, $\oplus$ means the element-wise addition, and $\otimes$ means the matrix multiplication.

discrimination loss is added to the self-attentive visual features to further enhance the discrimination capability of the visual modality. In the following sections, we describe the details of all the modules.

A. Audio-Guided Triplet Attention

Considering that the audio-visual event is audible and visible, we first enhance the importance of event-relevant visible objects and localize the sound sources in the visual representation with the guidance of audio signals. Previous works have proposed several AGVA modules [21], [40], [28] to guide visual signals with audio signals. Unfortunately, in these modules, audio features only participate in visual attention in limited dimensions, where AGVA [40] focuses on the spatial dimension and AGSVA [28] focuses on the spatial and channel dimensions. We propose a novel AGVA module called Audio-Guided Triplet Attention (AGTA) to exploit audio-guided visual attention in channel, temporal, and spatial dimensions. This module stresses informative features and every timely spatial region feature to improve the localization accuracy. We also introduce an attention module based on multi-modal factorized bilinear pooling (MFB) [41] that is embedded in every spatial region to emphasize the high-level semantic multi-modal fusion information.

Fig. 4 shows the architecture of the ATGA module. This module takes as the input the audio features $a_t \in \mathbb{R}^{d_a}$ and visual features $v_t \in \mathbb{R}^{d_v \times (HW)}$ (where H and W are the height and width of the feature maps, respectively), and outputs audio-enhanced visual features. In the following, we describe the details of AGTA module.

Channel-wise Attention. We first regulate the visual features with the channel-wise weights, which are computed by fusing the audio and visual features. Channel-wise attention can explicitly model the correlations among all of the elements across the channels. The aim is to select the event-relevant feature clues and suppress the others with the guidance of audio signals, while building the potential interactions for fine-grained audible events since the interactions usually correspond to different channels of the features. We follow [28] and implement channel-wise attention. First, the audio and visual features are projected and aligned with two non-linear transformations, and the channel-wise weights are obtained through a squeeze-and-excitation block [35]. This process can be formulated as:

$$\mathcal{M}_t^C = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1(AVP(\Phi_a^c a_t \odot \Phi_v^c v_t)))), \quad (1)$$

where $\Phi_a^c \in \mathbb{R}^{d_v \times d_a}$ and $\Phi_v^c \in \mathbb{R}^{d_v \times d_v}$ are fully-connected layers with ReLU; $\odot$ denotes the element-wise multiplication; AVP denotes the global average pooling in the spatial dimension; $\mathbf{W}_1 \in \mathbb{R}^{d_v \times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{d \times d_v}$ are two linear transformations with $d = 256$; δ and σ represent the ReLU and sigmoid activations, respectively; and the channel-wise attention map is denoted as $\mathcal{M}_t^C \in \mathbb{R}^{d_v \times 1}$.

Then, the channel attentive visual features are obtained via:

$$v_t^C = \mathcal{M}_t^C \odot v_t, \quad (2)$$

where the element-wise multiplication $\odot$ is conducted in the channel dimension of v_t.

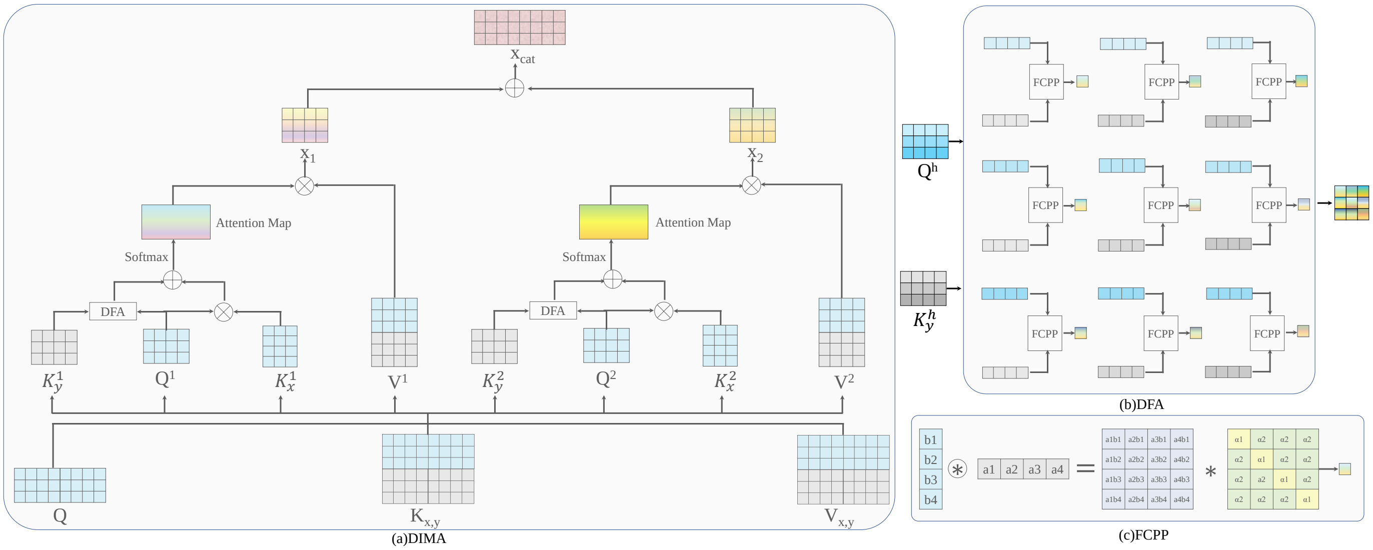
MFB-based Attention. We first exploit the audio-visual spatial correlation using the MFB method. MFB can build compact correlations between audio and visual features, which is better than simple element-wise addition and multiplication. Therefore, we apply MFB to the visual features of each visual region with audio features that model the fine-grained audio-visual relationships in a spatial-wise manner. Specifically, we project and align the audio features a_t and channel attentive visual features v_t^C with the same dimension kd_o using fully-connected layers with ReLU. Then, the projected features are fed into the MFB module to compute the correlation instead of using simple element-wise multiplication. We use the shared MFB block to proceed with the audio feature and the visual feature in each spatial location. Essentially, the d_o-dimensional correlation $\mathcal{M}_{av_t^s}^S$ is obtained by $a_t \mathbf{W} v_t^s$, where $\mathbf{W} \in \mathbb{R}^{d_a \times d_v \times d_o}$ is a learnable tensor, and $s \in [1, ..., HW]$ is the spatial location index of visual feature v_t. Due to the large number of parameters in $\mathbf{W}$, the MFB block utilizes the factorization trick and achieves the following reformulation:

$$\mathcal{M}_{av_t^s}^S = D(SP(\Phi^T a_t \odot \Psi^T v_t^s, k)), \quad (3)$$

where $\Phi \in \mathbb{R}^{d_a \times (kd_o)}$ and $\Psi \in \mathbb{R}^{d_v \times (kd_o)}$ are two learnable matrices factorized from $\mathbf{W}$, $\odot$ represents element-wise multiplication, $SP(f, k)$ represents the sum pooling operation with kernel size k and stride of k, and a dropout layer $D(\cdot)$ is used to prevent potential over-fitting. In addition, power and L_2 normalizations are also applied to stabilize the model training:

$$\mathcal{M}_{av_t^s}^S \leftarrow \mathrm{sign}(\mathcal{M}_{av_t^s}^S)\sqrt{\left|\mathcal{M}_{av_t^s}^S\right|}, \mathcal{M}_{av_t^s}^S \leftarrow \mathcal{M}_{av_t^s}^S / \left\|\mathcal{M}_{av_t^s}^S\right\|. \quad (4)$$

Temporal Attention. The aforementioned MFB-based attention primarily models the correlation between audio and visual features in the spatial dimension. Hereafter, we also consider temporal attention based on Bi-LSTM. Different from [40], which simply feeds globally averaged visual features and audio features into Bi-LSTM, we process every spatial location $s \in [1, ..., HW]$ in a dense manner. The motivation behind spatial-wise temporal attention is to build the temporal



(a)DIMA

(b)DFA

(c)FCPP

Fig. 5. Illustration of the dense inter-modality attention module (DIMA). DIMA contains a dense fusion attention module (DFA) with full channel pair product operation (FCPP) to compute the dense correlation for exploring inter-modality relationship. DIMA is also in multi-head setting, and the number of head in this illustrated example is 2 ($h = 1, 2$). The blue and gray colors separately denote two modality features. ⊕ means the concatenation along the temporal dimension, ⊗ means the matrix multiplication, and ⊛ means the outer product.

dependencies of each spatial region among all the audio-visual elements between temporal frames, and to highlight event-relevant visual features without potential loss of information. Specifically, we use the fully-connected layers with ReLU to project the audio features a_t and channel attentive visual features v_t^C into the same dimension d_o again. The formulation for obtaining the projected audio features $a_t^p \in \mathbb{R}^{d_o}$ and visual features $v_t^p \in \mathbb{R}^{d_o \times HW}$ is as follows:

$$a_t^p = \Phi_a a_t, v_t^p = \Phi_v v_t^C, \tag{5}$$

where $\Phi_a \in \mathbb{R}^{d_o \times d_a}$ and $\Phi_v \in \mathbb{R}^{d_o \times d_v}$ are linear operations with ReLU activation. Then, the spatial visual feature $v_t^{p_s}$ and audio feature a_t^p are reorganized as $\left(\left[v_1^{p_s} + a_1^p, a_1^p\right], ..., \left[v_T^{p_s} + a_T^p, a_T^p\right]\right)$. Then, the input $\left[v_t^{p_s} + a_t^p, a_t^p\right]$ is forwarded through Bi-LSTM. The d_o-dimensional correlation $\mathcal{M}_{av_t^s}^{\mathcal{T}}$ is calculated as:

$$\mathcal{M}_{av_t^s}^{\mathcal{T}} = \mathrm{BiLSTM}\left(\left[v_t^{p_s} + a_t^p, a_t^p\right]\right). \tag{6}$$

Spatial Attention. Spatial attention can build the audio-visual relationships among each spatial location, highlight the semantic objects in the visual modality under the guidance of audio signals, and finally obtain a compact and audio-related high-dimensional visual representation. In the AGTA module, temporal attention and MFB-based attention are conducted in a spatial-wise manner for fine-grained audio-visual correlation modeling, and then embedded in the spatial attention module for re-organizing spatial visual and audio features, which aims to build more fine-grained and temporally aligned audio-visual features. Specifically, we compute the spatial attention map ($\mathcal{M}_t^{\mathcal{S}} \in \mathbb{R}^{HW}$) by combining MFB-based attention and temporal attention. In this way, the bilinear model based fusion and LSTM can promote each other, and their advantages can be fully exploited in each spatial region from the spatial and temporal dimensions. The detailed formulation is as follows:

$$\mathcal{M}_t^{\mathcal{S}} = \mathrm{Softmax}\left(\tanh\left(\mathbf{W}_3\left(\mathcal{M}_{av_t}^{\mathcal{S}} + \mathcal{M}_{av_t}^{\mathcal{T}}\right)\right)\right), \tag{7}$$

where $\mathcal{M}_{av_t}^{\mathcal{S}}$ and $\mathcal{M}_{av_t}^{\mathcal{T}}$ are separately obtained by concatenating $\mathcal{M}_{av_t^s}^{\mathcal{S}}$ (Eqn.(4)) and $\mathcal{M}_{av_t^s}^{\mathcal{T}}$ (Eqn.(6)) in the spatial dimension ($s \in [1, ..., HW]$), and $\mathbf{W}_3 \in \mathbb{R}^{1 \times d_o}$ is the linear dimension reduction. Then, the channel attentive visual feature v_t^C is further updated using the spatial attention map $\mathcal{M}_t^{\mathcal{S}}$:

$$v_t^{\mathcal{S}} = \mathrm{sum}(\mathcal{M}_t^{\mathcal{S}} \odot v_t^C), \tag{8}$$

where sum is conducted in the spatial dimension.

To this end, we obtain the audio-enhanced visual features by making full use of the information from the channel, temporal, and spatial dimensions.

B. Intra-Modality Attention Module

After the AGTA module, the visual features are enhanced through audio-guided attention at the segment level. For visual features $v \in \mathbb{R}^{T \times d_v}$ and audio features $a \in \mathbb{R}^{T \times d_a}$, we use the linear layers to project them into the same dimension of $T \times d_m$, thereby yielding transformed visual and audio features as T_v and T_a, respectively. Then, we forward the transformed features into the intra-modality attention module to learn which segments are more informative and to explore the intra-modality information for each modality.

Technically, we adopt the encoder of the transformer to perform intra-modality attention, and we take the visual modality as an example to present the details of the intra-modality attention module. Specifically, we first project visual feature $T_v \in \mathbb{R}^{T \times d_m}$ into a query feature, key feature, and value feature, denoted as $Q \in \mathbb{R}^{T \times d_m}$, $K \in \mathbb{R}^{T \times d_m}$, and $V \in \mathbb{R}^{T \times d_m}$, respectively. Next, the intra-modality attention is calculated using the dot product operation in a multi-head setting. $Q^h \in \mathbb{R}^{T\times d_k}$, $K^h \in \mathbb{R}^{T\times d_k}$, and $V^h \in \mathbb{R}^{T\times d_k}$ respectively represent heads of related features, and d_k represents the dimension of each head. The formulation is written as:

$$\begin{aligned} &Q = T_v\mathbf{W}^Q, K = T_v\mathbf{W}^K, V = T_v\mathbf{W}^V,\\ &v_h = \text{Softmax}\left(\frac{Q^h\left(K^h\right)^T}{\sqrt{d_k}}\right)V^h, \\ &v_{cat} = \text{cat}(v_1, v_2, ..., v_n)\mathbf{W}^O, \end{aligned} \tag{9}$$

where $\mathbf{W}^Q \in \mathbb{R}^{d_m\times d_m}$, $\mathbf{W}^K \in \mathbb{R}^{d_m\times d_m}$, $\mathbf{W}^V \in \mathbb{R}^{d_m\times d_m}$, and $\mathbf{W}^O \in \mathbb{R}^{d_m\times d_m}$ are the projection matrices; and we employ $n = 4$ parallel attention heads. We utilize a residual connection followed by layer normalization to reduce degeneration, and a feed-forward layer is added to further fuse several parallel pieces of information.

$$\begin{aligned} v_r &= LayerNorm\left(v_{cat} + T_v\right),\\ v_f &= \delta\left(v_r\mathbf{W}_4\right)\mathbf{W}_5, \\ v_{self} &= LayerNorm\left(v_f + v_r\right), \end{aligned} \tag{10}$$

where $\mathbf{W}_4$ and $\mathbf{W}_5$ are learnable parameters and δ represents the ReLU activation function.

C. Dense Inter-Modality Attention Module

In [28], a cross-modality relation module is introduced to explore the relationship between audio and visual features via the decoder structure of the transformer. However, this module hinders the fusion of the two modalities because their query prefers to leverage the information of their own modality and does not fully exploit the relation information between audio and visual features. In this paper, we propose a novel *dense fusion attention* (DFA) to extend the traditional dot-product attention (DPA) [29] yielding a new dense inter-modality attention module (DIMA).

As shown in Fig. 3, two DIMA modules separately take (a, v_{self}) and (v, a_{self}) as the input pairs. We simply use (x, y), where $x \in \mathbb{R}^{T\times d_m}$ denotes one modality feature (*e.g.*, visual modality), and $y \in \mathbb{R}^{T\times d_m}$ denotes the other modality feature (*e.g.*, audio modality), to denote the input pair of the DIMA module. Feature $x \in \mathbb{R}^{T\times d_m}$ is projected into query feature $Q \in \mathbb{R}^{T\times d_m}$. Then, we temporally concatenate x with $y \in \mathbb{R}^{T\times d_m}$ to obtain a feature $F_{x,y} \in \mathbb{R}^{2T\times d_m}$, which is transformed into key feature $K_{x,y} \in \mathbb{R}^{2T\times d_m}$ and value feature $V_{x,y} \in \mathbb{R}^{2T\times d_m}$. DIMA is performed in a multi-head setting. The number of heads is denoted as n, and the dimension of each head is represented as d_k. In the DIMA module, we split the correlation of each head between Q^h and $K^h_{x,y}$ ($\mathcal{N}_{Q^h,K^h_{x,y}} \in \mathbb{R}^{T\times 2T}$) as $\mathcal{N}_{Q^h,K^h_x} \in \mathbb{R}^{T\times T}$ and $\mathcal{N}_{Q^h,K^h_y} \in \mathbb{R}^{T\times T}$, as shown in Fig. 5(a). The intra-modality correlation $\left(\mathcal{N}_{Q^h,K^h_x}\right)$ is obtained via classical matrix multiplication (the key point is the dot product of two vectors), and the inter-modality correlation $\left(\mathcal{N}_{Q^h,K^h_y}\right)$ is achieved through our proposed DFA to model the dense cross-modality correlation. The formula of DIMA is as follows:

$$\begin{aligned} &Q = x\mathbf{W}^Q, K_{x,y} = F_{x,y}\mathbf{W}^K, V_{x,y} = F_{x,y}\mathbf{W}^V,\\ &\mathcal{N}_{Q^h,K^h_x} = \frac{Q^h(K^h_x)^T}{\sqrt{d_k}}, \mathcal{N}_{Q^h,K^h_y} = \frac{DFA(Q^h, K^h_y)}{\sqrt{d_k}},\\ &x_h = \text{Softmax}\left(\text{cat}\left(\mathcal{N}_{Q^h,K^h_x}, \mathcal{N}_{Q^h,K^h_y}\right)\right)V^h_{x,y},\\ &x_{cat} = \text{cat}(x_1, x_2, ..., x_n)\mathbf{W}^O, \end{aligned} \tag{11}$$

where $\mathbf{W}^Q \in \mathbb{R}^{d_m\times d_m}$, $\mathbf{W}^K \in \mathbb{R}^{d_m\times d_m}$, $\mathbf{W}^V \in \mathbb{R}^{d_m\times d_m}$, and $\mathbf{W}^O \in \mathbb{R}^{d_m\times d_m}$ are the learnable parameters; and the number of parallel attention heads is 4 ($n = 4$). Similar to Eqn.(10), the same operations are conducted on x_{cat} to obtain the final cross-modality features (v_{cross} or a_{cross}).

As shown in Fig. 5(b), DFA calculates the correlation of two modalities ($\mathcal{N}$) via a variant of matrix multiplication, where $\mathcal{N}_{i,j}$ is obtained by applying the *full channel pair product* (FCPP) instead of the dot product for query feature Q^h_i and key feature $\left(K^h_y\right)_j$. This process is formulated as:

$$\begin{aligned} \mathcal{N} &= DFA\left(Q^h, K^h_y\right),\\ \mathcal{N}_{i,j} &= FCPP\left(Q^h_i, \left(K^h_y\right)_j\right). \end{aligned} \tag{12}$$

As illustrated in Fig. 5(c), FCPP refers to the full channel pair product of query and key features. It is calculated as:

$$FCPP\left(Q^h_i, \left(K^h_y\right)_j\right) = \text{sum}\left(\left(Q^h_i \circledast \left(K^h_y\right)_j\right) \odot \mathbf{W}\right), \tag{13}$$

where $\circledast$ is the outer product and $\mathbf{W}$ is the weight matrix. We split the elements of the matrix $\left(Q^h_i \circledast \left(K^h_y\right)_j\right)$ into two groups: the diagonal elements (corresponding to the original inner product), and the remaining elements. Accordingly, the diagonal elements of $\mathbf{W}$ are α and the remaining elements of $\mathbf{W}$ are $(1-\alpha)\frac{1}{d_k-1}$, where $\frac{1}{d_k-1}$ is the normalization factor. Intuitively, we assign equal weights, *i.e.*, $\alpha = 0.5$ in our experiments, to these two groups.

D. Audio-Visual Fusion Module

To obtain a compact and joint representation from audio and visual modalities for the event localization task, we follow [28] and fuse audio and visual features using an audio-visual fusion module. The query F_{av} is the correlation obtained by the element-wise product. The key is the temporal concatenation of a_{cross} and v_{cross}. Considering that the query and key are high-level comprehensive representations of the two modalities, we replace the DFA in DIMA with DPA. Layer normalization and a residual connection are added to reduce transmission loss. The final multi-modality representation F_o is obtained as follows:

$$\begin{aligned} F_{av} &= a_{cross} \odot v_{cross},\\ O &= DIMA\left(F_{av}, \text{cat}\left(a_{cross}, v_{cross}\right)\right),\\ F_o &= LayerNorm\left(F_{av} + O\right). \end{aligned} \tag{14}$$

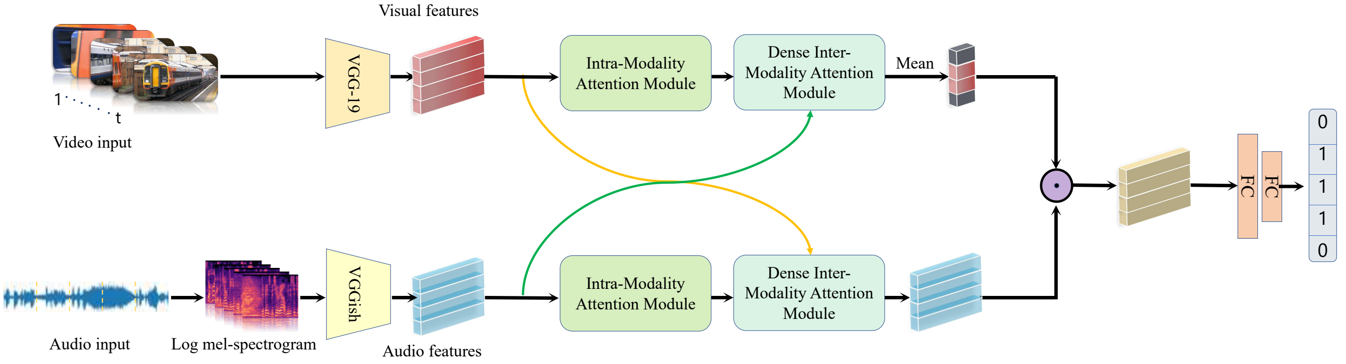

Visual features
Video input
1 · · · · · t
VGG-19
Intra-Modality Attention Module
Dense Inter-Modality Attention Module
Mean
FC
FC
0
1
1
1
0
Audio input
Log mel-spectrogram
VGGish
Audio features
Intra-Modality Attention Module
Dense Inter-Modality Attention Module

Fig. 6. The architecture of our proposed dense cross modality relation network for V2A. The "Mean" operation is averaging the features in channel dimension. When solving the A2V, the "Mean" operation is correspondingly conducted in the audio modality. The pink segments denote the event-relevant query segments and the black segments denote the background segments. And the final correlated features are obtained by multiplying the event-relevant query segments with the whole segments in the other modality in a sliding window manner. ⊙ means the element-wise multiplication.

E. Loss Functions

Previous methods mainly focus on the final two-modality fused features with the event localization loss. Through various attention modules, which are commonly used in existing methods, some useful discrimination information of the unimodal features might be destroyed or ignored. In this work, we introduce a unimodal discrimination loss (UDL), which adds constraints on the middle-level visual features. Our UDL can be used in fully- and weakly-supervised settings. In the following, we describe the details of the loss functions.

Fully-Supervised Losses. For the fully-supervised setting, we can access the segment-level labels. The training objective for fully-supervised event localization consists of an event-relevant loss on F_o and event category losses on v_{self} and F_o. The event-relevant loss is defined as:

$$s = \text{Sigmoid}(FC(F_o)),$$
$$\mathcal{L}_{ero}^{f} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} s_i^t \log(s_i^t) + (1-s_i^t)\log(1-s_i^t), \tag{15}$$

where N is the number of training samples, and FC is the classifier.

In addition, the event category loss on v_{self} is formulated as:

$$\hat{s} = \text{Softmax}(FC(\text{MaxPooling}(v_{self}))),$$
$$\mathcal{L}_{ecm}^{f} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{1}\{y^i = k\}\log(\hat{s}_k), \tag{16}$$

where K is the number of event categories, $\mathbb{1}\{\cdot\}$ is the indicator function ($\mathbb{1}\{True\} = 1$ and $\mathbb{1}\{False\} = 0$). Similarly, $\mathcal{L}_{eco}^{f}$ is obtained by replacing v_{self} with F_o in Eqn.(16).

In summary, the objective for the fully-supervised setting is:

$$\mathcal{L}_{fs} = \mathcal{L}_{ero}^{f} + \lambda \cdot \mathcal{L}_{ecm}^{f} + (1-\lambda)\cdot \mathcal{L}_{eco}^{f}, \tag{17}$$

where λ is a balancing factor.

Weakly-Supervised Losses. In a weakly-supervised setting, we can leverage only video-level labels, and this is usually modeled as a multiple instance learning (MIL) task [58]. In this case, our training objective includes event category losses on v_{self} and F_o. To simultaneously obtain the event-relevant and event category predictions in the inference stage (like in the fully-supervised setting), we aggregate these two types of predictions to achieve video-level prediction using MIL pooling [59]. Then, the event category losses on v_{self} ($\mathcal{L}_{ecm}^{w}$) and F_o ($\mathcal{L}_{eco}^{w}$) are separately calculated based on the multi-label soft margin loss. To this end, the final objective is

$$\mathcal{L}_{ws} = \lambda \cdot \mathcal{L}_{ecm}^{w} + (1-\lambda)\cdot \mathcal{L}_{eco}^{w}. \tag{18}$$

V. Cross-Modality Localization

To evaluate the model's capacity for exploring audio-visual relationships, we propose a *dense cross modality relation* network (DCMR), which is a deformation of our dense modality interaction network, to solve the cross-modality localization (CML) task. CML aims to find the position of synchronized event-relevant content between two modalities. The architecture of DCMR is shown in Fig. 6.

In the training stage, event-relevant labels are provided. The audio feature $a_t \in \mathbb{R}^{d_a}$ and visual feature $v_t \in \mathbb{R}^{d_v}$ are fed into the intra-modality attention module to obtain the self-attentive features a_{self} and v_{self}. Then, the most important module DIMA uses one modality feature as the query to jointly explore the intra- and inter- modality relationship generating a_{cross} and v_{cross}. These two features are utilized to further perform cross matching. Taking V2A as an example (see Fig. 6), we first average v_{cross} in channel dimension: $v_{mean} = Mean(v_{cross})$, where Mean is the channel-wise average pooling operation. Furthermore, the query feature v_{query} is obtained from v_{mean} by removing the background segments according to the event-relevant labels. Then, we conduct the cross-matching operation by multiplying the audio feature with the query visual feature in a sliding window manner, which can shorten the distance of matched segments

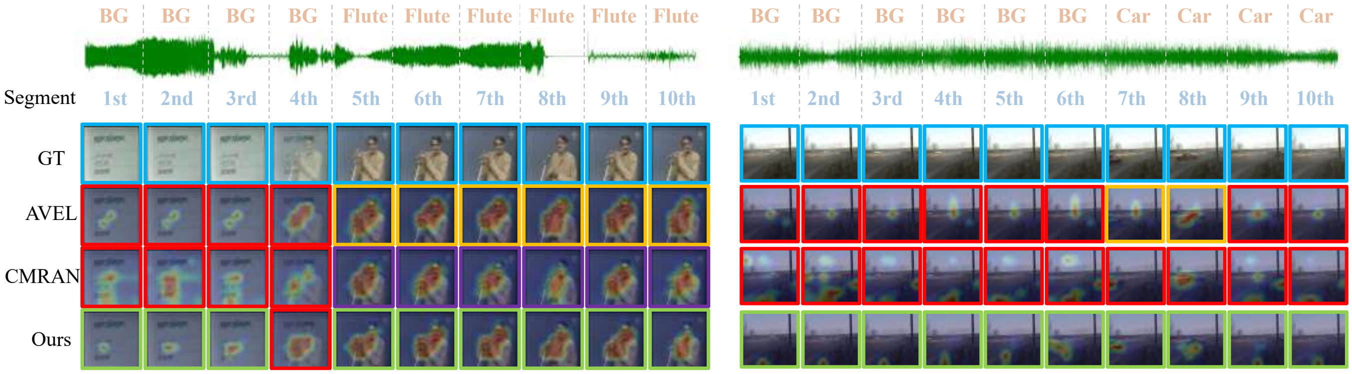


Fig. 7. Qualitative comparisons of our method with AVEL [21] and CMRAN [28]. The blue box refers to the ground truth. The orange, purple, and green boxes separately represent the correct results of AVEL, CMRAN, and our method, while the red box represents the wrong result.

while lengthening the distance of mismatched segments. This process is computed as:

$$f_{av} = v_{query} \odot \underset{k=1,\ldots,\mathcal{K}}{a_{cross}}\,, \tag{19}$$

where $\mathcal{K}$ is the number of K-second segments in a_{cross}. Given that only an event-relevant label is provided for the cross-modality localization task, DCMR is trained with the event-relevant loss, as shown in Eqn.(15).

In the inference stage, we examine the final prediction scores of the T-length candidate segments, and output the K-length segment with the maximum contiguous sum as the final localization prediction.

VI. Experiments

In this section, we first describe the dataset and implementation details. Then, we compare our method with the most advanced methods on the AVE dataset. Ablation studies are conducted to validate and analyze our method.

A. Dataset

Following previous works [26], [27], [23], [28], [25], we use the AVE dataset [21] to evaluate our method. The AVE dataset is a subset of AudioSet [57]. The AVE dataset contains 4,143 videos, each of which has a duration of 10 seconds. Each video contains one event category and is temporally labeled with AVE boundaries. The number of event categories in the AVE dataset is 28, including dog barking, acoustic guitar, female speech, and so on. The audio-visual event in each video lasts from 2 seconds to 10 seconds. We adopt the original splitting [21] where the ratio of training/validation/testing samples is 8:1:1. To evaluate the cross-modality localization task, only short-event videos, where the duration of the event is strictly between 2 seconds and 9 seconds, are sampled.

B. Implementation Details

Feature Extraction. Following the common setting, for the visual features used in the audio-visual event localization task, we apply VGG-19 network [54] pre-trained on ImageNet [55] to extract *pool5* feature maps from 16 sampled RGB frames within each one-second video segment and then use global average pooling on the 16 frames to obtain a 512×7×7 visual feature. For the cross-modality localization task, global average pooling is also used in the spatial dimension to generate a 512-D visual feature vector. In addition, we extract the audio features via the VGGish [56] pre-trained on AudioSet [57]. Each one-second audio segment is first transformed into log-mel spectrograms and then a $128-D$ feature vector is extracted in both tasks.

Training Details. Our experiments are implemented with PyTorch 1.2.1. The GPU is a NVIDIA TITAN RTX. Adam [60] optimizer is used to train our models. We set the batch size as 32. The base learning rate is initialized to 0.0005, divided by 2 every 10 epochs and fixed after 30 epochs.

Evaluation Metric. In the audio-visual event localization task, our network predicts the event label for each one-second video segment. For a t-th segment, if the event-relevant score $s_t \geqslant 0.5$, the event label is assigned according to the event category score $\hat{s}$, otherwise, this segment is predicted as background. For fully- and weakly-supervised event localization, we follow previous works [21], [26], [27], [23], [28], [25], which apply the overall classification accuracy as the evaluation metric. The cross-modality task has two parts, namely, audio localization from visual content (V2A) and visual localization from audio content (A2V). Following previous works [21], [27], we evaluate the performance by calculating the proportion of correct matches in all testing samples, where a correct matching is when the predicted audio/visual segment is exactly matched with the other modality as the ground truth.

C. Comparisons with State-of-the-art Methods

1) Audio-Visual Event Localization: In this section, we compare our proposed dense modality interaction network with several of the most advanced methods quantitatively and qualitatively.

Quantitative Evaluation. We compare our method with state-of-the-art methods under fully- and weakly-supervised settings. For fair comparisons, we use the same settings as those in previous methods. The results are reported in Table I. For fully-supervised AVE localization, our method achieves the highest accuracy of 79.6%, which surpasses the second best method PSP [25] by 1.8%. For the weakly-supervised setting,

TABLE I
COMPARISONS OF ACCURACY (%) ON THE AVE DATASET FOR THE FULLY- AND WEAKLY-SUPERVISED SETTINGS. * AND ** MEAN THAT THE RESULTS ARE REPRODUCED BY [28] AND [25], RESPECTIVELY.

Methods	Fully-supervised	Weakly-supervised
AVEL (Only Video) [21]	55.3	52.9
AVEL (Only Audio) [21]	59.5	53.4
AVSDN* [26]	72.6	66.8
AVEL [21]	72.7	66.7
CMAN** [23]	73.3	70.4
DAM [27]	74.5	-
AVRB [40]	74.8	68.9
AVIN [22]	75.2	69.4
AVT [24]	76.8	70.2
CMRAN [28]	77.4	72.9
PSP [25]	77.8	73.5
Ours	**79.6**	**74.3**

our method still obtains the best performance of 74.3%, which is higher than that of PSP [25] by 0.8%. These two results indicate the superiority of our proposed method for the AVE localization task.

Qualitative Evaluation. Fig. 7 shows two examples of audio-visual event localization. Red refers to the wrong results, and the other colors mean that the classification is correct. As shown in the AGTA attention weight map (i.e., $\mathcal{M}_t^S$ in Eqn.(7)) in each segment (the last row), when the segments contain an audio-visual event, such as the last six segments in the first example, the attention can focus on the sounding objects. When the segment does not contain an audio-visual event, such as the first three segments in the first example, where the flute source is not visible, the attention map of AGTA is smaller than those of the other methods, which means that the correlation between the two modalities is smaller, which is helpful for background inference. For the BG segments in the second example, the two other methods somehow focus on the noisy visual regions, thereby resulting in the incorrect prediction of a train horn (CMRAN) or truck (AVEL). In contrast, AGTA can pay attention to the unrelated background region, which is helpful for prediction. For the seventh and eighth segments in the second example, AGTA can capture the car and ignore the distant objects, and the two other methods still focus on the noisy distant visual regions. These analyses validate that AGTA can capture a salient map with sounding objects when the two modalities are related; otherwise, AGTA can prevent the interference of semantic objects when the two modalities are unrelated.

In addition, we rank the attention weights to analyze the most associated segment. In the two examples, we find that DIMA would opt for the information of the other modality to improve modality fusion performance. However, CMRA [28] prefers to utilize the information of its own modality which limits the modality fusion and the prediction accuracy.

2) Cross-Modality Localization: In this part, we perform the quantitative and qualitative evaluations to compare our DCMR method with other methods.

Quantitative Evaluation. Table II shows the comparisons of our DCMR method with three state-of-the-art methods, namely, DCCA [61], AVDLN [21], and DAM [27] on the cross-modality localization task. For a fair comparison, we use the same settings as previous methods. AVDLN essentially measures the Euclidean distance of audio and visual features. AVDLN and DCCA utilize the local segments to compute the correlation of two modalities, and DAM uses the global features obtained by temporally averaging self-attentive query features to check each segment of the other modality. Unfortunately, these methods neglect to explore the dense inter-modality relationship. Our method deeply explores the cross-modality correlation based on our DIMA module. Specifically, our method improves the accuracy from 47.1% to 56.0% on the A2V task, and from 48.5% to 56.2% on the V2A task.

TABLE II
COMPARISONS OF ACCURACY (%) OF OUR METHOD WITH DCCA, AVDLN, AND DAM ON THE CROSS-MODALITY LOCALIZATION TASK.

Methods	A2V	V2A	Average
DCCA [61]	34.1	34.8	34.5
AVDLN [21]	35.6	44.8	40.2
DAM [27]	47.1	48.5	47.8
Ours	**56.0**	**56.2**	**56.1**

TABLE III
EFFECT OF THE AGTA MODULE ON THE ACCURACY(%).

Methods	Fully-supervised	Weakly-supervised
w/o AGTA	76.6	71.7
AVTrans[24]	75.8	68.1
AGVA [21]	77.1	72.1
AGSCA [28]	78.0	72.6
AGTA w/o LSTM	78.0	72.3
AGTA w/o MFB	77.6	73.0
AGTA w/o channel	78.8	71.9
AGTA w/o spatial	77.5	69.5
AGTA	**79.6**	**74.3**

Qualitative Evaluation. We show two qualitative results of the cross-modality localization task in Fig. 8. The blue box represents the query segments; and the red and green boxes denote incorrect and correct predictions, respectively. In the A2V example, the performer plays the accordion in the first five segments. Only the first five segments contain the accordion sound; however, all ten segments contain the accordion object, and the similarity of each visual segment increases the difficulty of locating the synchronized visual content given the audio query. In the V2A example, a helicopter exists in the first few visual segments, but it makes sound for all ten segments, and the similarity of each audio segment makes finding the synchronized audio content given the visual query more difficult. For these two challenging cases, where the segments in the target modality have relatively high similarity, our DCMR method can achieve the correct matching. Compared with sparse cross-modality correlation modeling in CMRA [28], the dense relationship of audio and visual features exploited by our DCMR is more effective in locating synchronized segments.

D. Ablation Studies

In this section, we verify the effectiveness of the AGTA module, the DIMA module, the DFA module, and the UDL through ablation studies.

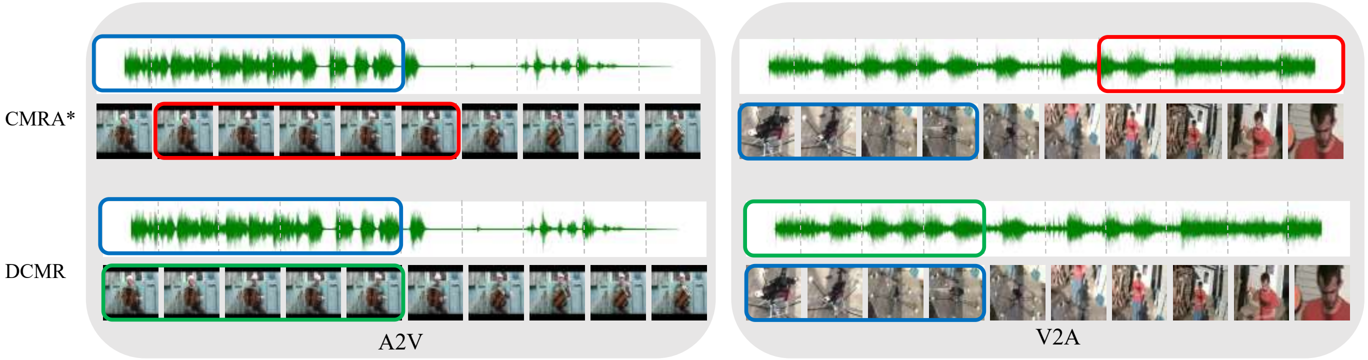


Fig. 8. Qualitative comparisons of DCMR with CMRA*. The CMRA* is the deformation of CMRAN [28] to solve the CML task. The deformed network is same as the DCMR expect that replacing DFA with DPA. The blue box refers to query segments. The green and red boxes represent the right and wrong predictions, respectively.

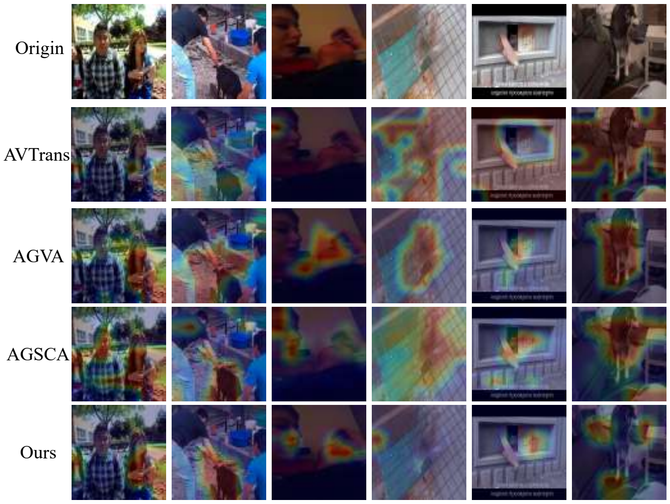


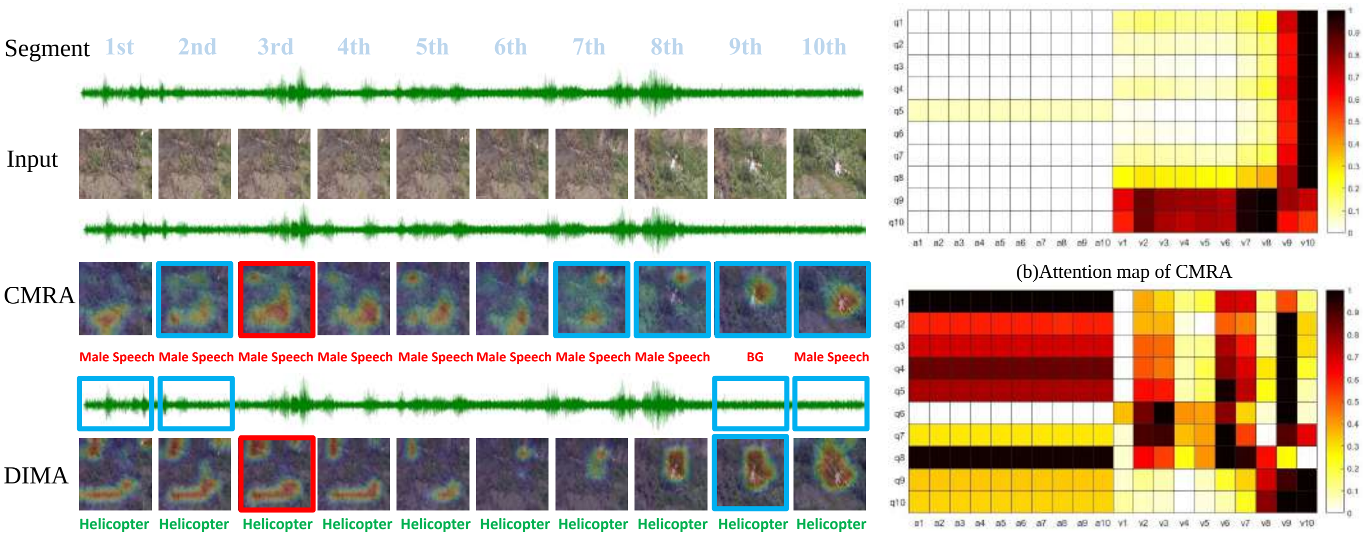
Fig. 9. Qualitative comparisons of AVTrans [24], AGVA [21], AGSCA [28], and our AGTA. The examples in the first three columns are the audio-visual event segments, where the audio and visual semantic contents are same. The ukulele, sheep, and baby are making sound in these examples, and our method can more exactly and tightly localize the sound sources compared with other methods. In addition, the examples in the last three columns are the background segments, where the audio and visual contents are different. In details, the rat, cat, and dog occur in the last three columns' visions, but they do not make a sound. Our AGTA can prevent from the interference of the semantic objects in these cases, which is helpful to infer the background segments.

Effect of the AGTA Module. In our work, the AGTA module is proposed to enhance the visual features with comprehensive guidance from audio signals. To validate this module, we compare our method with several different settings for audio-guided attention, and the corresponding results are reported in Table III. "w/o AGTA" refers to replacing the AGTA module with spatial average pooling. "AGVA", "AGSCA", and "AVTrans" represent replacing the AGTA module with the audio-guided attention module from [21], [28], and [24], respectively. When removing the AGTA module, the accuracy of our method drops by 3.0% and 2.6% for fully- and weakly-supervised settings, respectively. These performance drops are proof that the effective guidance with audio information can remarkably improve the localization capability of the deep model. In addition, the performance of AGTA is obviously higher than those of AGVA and AGSCA by at least 1.6% and 1.7% for fully- and weakly-supervised settings, respectively. This finding indicates that more fine-grained guidance can obtain better visual representations. In addition, compared with AGTA, the performance of "AVTrans" remarkably decreases by 3.7% and 6.2% in the two respective settings. The reason is that the method ignores channel-wise attention and temporal modeling is limited to adjacent segments. Moreover, we deeply analyze our AGTA module by deleting the attention block separately: "AGTA w/o LSTM", "AGTA w/o MFB", and "AGTA w/o channel". The experimental results show that every attention block contributes to the final performance of our proposed AGTA module (comparing the last four rows in Table III).

Furthermore, we qualitatively compare the attention maps of AGVA [21], AGSCA [28], AVTrans [24] and our AGTA as shown in Fig. 9. For the first three columns, compared with AGVA, AGSCA, and AVTrans, our AGTA method can localize the sound sources more exactly and tightly. However, the attentions of AGVA, AGSCA, and AVTrans are spread over different background regions, hence limiting the performance of predicting event classes. The last three columns illustrate the cases that do not contain an audio-visual event. The fourth and sixth columns are the out-of-screen examples, which mean that the rat and dog do not make a sound, but a person's voice is heard out of the screen. Compared with the three other methods, AGTA can prevent interference from semantic objects, which is helpful in inferring the background segments.

Effect of the DIMA Module. We conduct an ablation study to verify the effectiveness of the DIMA module by comparing the following different settings: removing the DIMA module ("w/o DIMA"); directly replacing the DIMA module with HAN [62] and CMCA [23]; replacing the DIMA module with modified FGNL [43] (FGNL*), where we use the dot product to compute the intra-modality correlation of Q and K_x, and use the pairwise function in [43] to compute the inter-modality correlation of Q and K_y. These two correlations are concatenated along the temporal dimension to output the total correlation, which is fed into a softmax function to obtain the attention map. We multiply the attention map and $V_{x,y}$ to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3150469, IEEE Transactions on Multimedia



(b)Attention map of CMRA

(a)Qualitative comparisons of CMRA and DIMA

(c)Attention map of DIMA

Fig. 10. Figure (a) shows the comparisons of CMRA [28] and our method DIMA. A query segment is highlighted with the read box, and the segments with top-5 attention weights are highlighted with blue boxes. Figure (b) and (c) separately denote the attention maps of CMRA and DIMA for each visual segment. The vertical axis represents the segments of visual query (Q), and the horizontal axis represents the segments of key, where a_1-a_{10} (K_y) denote the audio part of key, and v_1-v_{10} (K_x) denote the visual part of key. For each row, the original attention weight is renormalized in $[0, 1]$ just for visualization. We can find that our DIMA enables to obtain richer information and build the inter-relationship between two modalities. However, CMRA tends to leverage the segments of the own modality limiting the fusion of two modalities.

obtain the relative response, and repeat the above operation by rolling K_y along the channel axis yielding the final response. The final features are achieved by applying the MoSE [43] to the final response; replacing the DIMA with modified GNL [42] (GNL*), where we take the Q and $cat(K_x, K_y)$ as inputs of GNL; replacing the DIMA module with CMRA [28] ("CMRA"), in other words, replacing the full channel pair product (FCPP) operation with the dot product operation; and our method ("DIMA"). The corresponding results are shown in Table IV. When removing "DIMA", the performance drops for fully- and weakly-supervised localization. This implies the importance of inter-modality attention. Comparing the rows of "HAN" with those of "DIMA" in Table IV, the accuracy of "DIMA" is apparently higher. HAN [62] simply sums up the results of the self-attention and cross-attention to model the cross-modality relationships. In contrast, our DIMA temporally concatenates the two modality features and introduces a DFA module to densely model the cross-modality correlation and extract the audio-visual channel correlation. In addition, the performance of "CMCA" [23] is also limited. First, "CMCA" uses LSTM, which has a limited temporal receptive field, to model the temporal correlation between two modalities, whereas our DIMA considers all the temporal inputs when computing the attention. Second, we introduce a full channel pair product to model the audio-visual correlation, which is omitted by "CMCA".

Comparing the rows of "FGNL*" with those of "DIMA" in Table IV shows that if we replace DIMA with FGNL*, the performance will decline from 79.6%/74.3% to 77.2%/73.7%. The main reason for this reduction is that DIMA can build correlations simpler and more effectively. In detail, FGNL* divides the channel-wise correlations into multiple groups and each of them is independently processed to update the features many times; however, our DIMA updates the features only one time by summing up these correlations with different weights and then linearly combining the two modality features. In addition, a modified squeeze-and-excitation scheme is used to restore the original channel dimensions in FGNL*, which may result in confusion. However, DIMA can keep the original feature architecture without a similar operation which can obtain more accurate localization. For GNL*, there is also a performance drop, *i.e.*, from 79.6%/74.3%("DIMA") to 78.0%/73.7% ("GNL*"). The reason might be that taking as the key each element obtained by collapsing all dimensions into one dimension may introduce potential confusion, whereas the DIMA uses the feature vector as the key, which can retain the original feature architecture.

TABLE IV
EFFECT OF THE DIMA MODULE ON THE ACCURACY (%).

Methods	Fully-supervised	Weakly-supervised
w/o DIMA	77.5	71.2
HAN [62]	76.1	72.1
CMCA [23]	76.4	66.8
FGNL* [43]	77.2	73.7
GNL* [42]	78.0	73.7
CMRA [28]	78.2	72.7
DIMA	**79.6**	**74.3**

Comparing the rows of "CMRA" with those of "DIMA" in Table IV shows that our method is superior to [28] in terms of the performance improvement (1.4% and 2.5% for two supervised settings). The reason is that the full channel pair product (FCPP) applied for dense fusion attention (DFA) in DIMA can obtain richer information than the dot product operation used for multi-head attention (MHA) in CMRA, thereby enhancing the representation capability.

Next, we qualitatively compare the cross attention modules of CMRA [28] and DIMA as shown in Fig. 10. Essentially,

TABLE V
EFFECT OF THE DFA MODULE ON THE ACCURACY (%).

Q&K_x	Q&K_y	Fully-supervised	Weakly-supervised
MHA	MHA	78.2	72.7
DFA	MHA	78.7	73.1
DFA	DFA	78.5	73.6
MHA	**DFA**	**79.6**	**74.3**

this experiment compares the effect of the fully channel pair product and dot product operation in DIMA. A query segment is highlighted with the red box, and the segments with the top-5 attention weights are highlighted with blue boxes. We also illustrate the attention maps of the two models in the right of Fig. 10. In Fig. 10(b), we observe that CMRA tends to leverage the segments of the own modality and utilize very little information from the other modality. This phenomenon is intuitive because the query would like to correspond with the own modality part of the key, which is not helpful for modality fusion. In contrast, our proposed DIMA (see Fig. 10(c)) enables one to obtain richer information and exploit the inter-relationship between two modalities by using full channel pair product computation. As shown in Fig. 10(a), the helicopter is visually small in the third segment, CMRA mostly uses only the visual features (all blue boxes are in the visual modality), whereas our DIMA can leverage the strong audio information of the helicopter (four green boxes are in the audio modality), which is useful for event predictions.

Effect of the DFA Module. In our DIMA, we use MHA for Q and K_x, and DFA for Q and K_y. The reason is that Q and K_x are the same modalities, and Q and K_y are different modalities. We apply DFA to explore the dense inter-modality relationship and try to align K_x and K_y. Table V reports the corresponding ablation study. The comparison of the fourth row with the last row shows that if DFA is used for Q with K_x and K_y, the same correlation weight computation would harm the performance. In addition, comparing the second row with the third and fourth rows, DFA still has an advantage for localizing the event. Overall, the present method is beneficial for aligning the two modalities and obtains the best accuracy (see the last row).

Effect of the UDL. In this work, we introduce a unimodal discrimination loss on the middle-level visual features, which is combined with the common event localization losses with a parameter λ. Table VI shows the accuracy of our network with different $\lambda = [0, 0.02, 0.04, 0.06, ..., 0.1]$. When $\lambda = 0.1$ and $\lambda = 0.08$, our method achieves the best performance for fully- and weakly-supervised event localization. When removing the UDL, *i.e.*, $\lambda = 0$, the accuracy will drop by 1.2% and 0.6% for the fully- and weakly-supervised settings, respectively. This observation verifies the effectiveness of the UDL. By analyzing and comparing the classification results without and with the UDL, we observe that some categories with similar sounds are better classified with the UDL. This finding further shows that the representations learned by our network with the UDL are more discriminative. In addition, we add the unimodal discrimination loss to the audio branch, but we do not obtain the expected effect. The reason might be that visual features (in four dimensions) have more information than audio features (in two dimensions). In other words, compared with audio features, visual features have a larger excavation space via the unimomal discrimination loss.

TABLE VI
IMPACT OF THE WEIGHT λ IN UDL ON THE ACCURACY (%).

Weights	Fully-supervised	Weakly-supervised
$\lambda = 0.00$	78.4	73.7
$\lambda = 0.02$	78.5	72.8
$\lambda = 0.04$	78.2	72.6
$\lambda = 0.06$	78.1	73.4
$\lambda = 0.08$	79.2	**74.3**
$\lambda = 0.1$	**79.6**	72.7

VII. CONCLUSION AND FUTURE WORK

In this work, we propose a dense modality interaction network for audio-visual event localization. Our standing point is to fuse the audio and visual modalities elegantly and deeply in the stages of audio-guided fusion and cross-modality fusion. We first propose an audio-guided triplet attention module to highlight the event-relevant regions in the visual features by applying fine-grained attention in the channel, temporal, and spatial dimensions. Then, we develop a dense inter-modality attention module to effectively fuse the audio-visual features via dense fusion attention, where the sparse correlation weight computation method is enhanced with our proposed full channel pair product. To exploit the localization capability of the unimodal and fused features simultaneously, we also introduce a unimodal discrimination loss function, which is combined with the common AVE event localization losses. Various experiments show that our network is superior to state-of-the-art methods in fully- and weakly-supervised AVE localization by a large margin. To further validate the superiority of our dense inter-modality attention for exploring the cross-modality correlation, we propose a dense cross modality relation network for cross-modality localization. The experimental results illustrate that our method achieves better performance.

In this work, we experimentally found that the unimodal discrimination loss on the middle-level visual features can work well, but no improvement was observed when adding this loss to the audio features. In the future, we would like to explore more appropriate methods to utilize the unimodal loss. We would also like to collect a large-scale AVE localization dataset that contains a larger number of videos and event categories to facilitate related studies. Furthermore, it would be interesting to extend our attention methods to other audio-visual problems, *e.g.*, audio-visual separation and representation learning.

ACKNOWLEDGMENT

This work was partially supported by the National Natural Science Foundation of China (62102418 and 62172415), the National Key R&D Program of China (2019YFB2204104), and the Alibaba Group through Alibaba Innovative Research Program.

References

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

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

[3] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *JSTSP*, vol. 14, no. 3, pp. 478–493, 2020.

[4] H. Zhu, M. di Luo, R. Wang, A. hua Zheng, and R. He, "Deep audio-visual learning: A survey," *IJAC*, vol. 18, pp. 351–376, 2021.

[5] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, "Seeing through noise: Visually driven speaker separation and enhancement," in *ICASSP*, 2018, pp. 3051–3055.

[6] R. Lu, Z. Duan, and C. Zhang, "Listen and look: Audio–visual matching assisted speech source separation," *SPL*, vol. 25, no. 9, pp. 1315–1319, 2018.

[7] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba, "The sound of motions," in *ICCV*, 2019, pp. 1735–1744.

[8] D. Surís, A. Duarte, A. Salvador, J. Torres, and X. Giró-i Nieto, "Cross-modal embeddings for video and audio retrieval," in *ECCV*, 2019, pp. 711–716.

[9] J. Zhang and Y. Peng, "Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval," *TMM*, vol. 22, no. 1, pp. 174–187, 2020.

[10] A. Zheng, M. Hu, B. Jiang, Y. Huang, Y. Yan, and B. Luo, "Adversarial-metric learning for audio-visual cross-modal matching," *TMM*, pp. 1–1, 2021.

[11] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *TMM*, vol. 2, no. 3, p. 141–151, 2000.

[12] J.-S. Lee and C. H. Park, "Robust audio-visual speech recognition based on late integration," *TMM*, vol. 10, no. 5, pp. 767–779, 2008.

[13] D. Hu, X. Li, and X. Lu, "Temporal multimodal learning in audiovisual speech recognition," in *CVPR*, 2016, pp. 3574–3582.

[14] P. Zhou, W. Yang, W. Chen, Y. Wang, and J. Jia, "Modality attention for end-to-end audio-visual speech recognition," in *ICASSP*, 2019, pp. 6565–6569.

[15] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *ICASSP*, 2018, pp. 6548–6552.

[16] F. Tao and C. Busso, "End-to-end audiovisual speech recognition system with multitask learning," *TMM*, vol. 23, pp. 1–11, 2021.

[17] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *CVPR*, 2016, pp. 2405–2413.

[18] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, "Visual to sound: Generating natural sound for videos in the wild," in *CVPR*, 2018, pp. 3550–3558.

[19] C. Gan, D. Huang, J. B. Chen, Peihao nd Tenenbaum, and A. Torralba, "Foley music: Learning to generate music from videos," in *ECCV*, 2020, pp. 758–775.

[20] S. Ghose and J. J. Prevost, "Autofoley: Artificial synthesis of synchronized sound tracks for silent videos with deep learning," *TMM*, vol. 23, pp. 1895–1907, 2021.

[21] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *ECCV*, 2018.

[22] J. Ramaswamy, "What makes the sound?: A dual-modality interacting network for audio-visual event localization," in *ICASSP*, 2020, pp. 4372–4376.

[23] H. Xuan, Z. Zhang, S. Chen, J. Yang, and Y. Yan, "Cross-modal attention network for temporal inconsistent audio-visual event localization," in *AAAI*, 2020, pp. 279–286.

[24] Y.-B. Lin and Y.-C. F. Wang, "Audiovisual transformer with instance attention for audio-visual event localization," in *ACCV*, 2020.

[25] J. Zhou, L. Zheng, Y. Zhong, S. Hao, and M. Wang, "Positive sample propagation along the audio-visual event line," in *CVPR*, 2021, pp. 8436–8444.

[26] Y.-B. Lin, Y.-J. Li, and Y.-C. F. Wang, "Dual-modality seq2seq network for audio-visual event localization," in *ICASSP*, 2019, pp. 2002–2006.

[27] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, "Dual attention matching for audio-visual event localization," in *ICCV*, 2019, pp. 6292–6300.

[28] H. Xu, R. Zeng, Q. Wu, M. Tan, and C. Gan, "Cross-modal relation-aware networks for audio-visual event localization," in *ACMMM*, 2020, pp. 3893–3901.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, p. 6000–6010.

[30] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement based on multimodal deep convolutional neural networks," *arXiv preprint arXiv:1709.00944*, 2017.

[31] B. Li and A. Kumar, "Query by video: Cross-modal music retrieval," in *ISMIR*, 2019, pp. 604–611.

[32] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "Centralnet: a multilayer approach for multimodal fusion," in *ECCV*, 2018, pp. 575–589.

[33] H. M. Fayek and A. Kumar, "Large scale audiovisual learning of sounds with weakly labeled data," in *IJCAI*, 2020, pp. 558–565.

[34] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *CVPR*, 2020.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[36] S. Woo, J. Park, J.-Y. Lee, and I.-S. Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.

[37] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," 2017, pp. 6298–6306.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019, pp. 4171–4186.

[39] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," 2018, pp. 7794–7803.

[40] J. Ramaswamy and S. Das, "See the sound, hear the pixels," in *WACV*, 2020, pp. 2959–2968.

[41] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *ICCV*, 2017, pp. 1821–1830.

[42] K. Yue, M. Sun, Y. Yuan, F. Zhou, E. Ding, and F. Xu, "Compact generalized non-local network," in *NeurIPS*, 2018, pp. 6510–6519.

[43] I. Kuo, W. Wei, and J. Lin, "Positions, channels, and layers: Fully generalized non-local network for singer identification," in *AAAI*, 2021, pp. 8217–8225.

[44] T.-I. Hsieh, Y.-C. Lo, H.-T. Chen, and T.-L. Liu, "One-shot object detection with co-attention and co-excitation," in *NeurIPS*, 2019, pp. 2721–2730.

[45] J. Perez-Rua, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "Mfas: Multimodal fusion architecture search," in *CVPR*, 2019, pp. 6966–6975.

[46] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *ICMI*, 2016, p. 284–288.

[47] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *ICME*, 2017, pp. 949–954.

[48] A. Anastasopoulos, S. Kumar, and H. Liao, "Neural language modeling with visual features," *ArXiv*, vol. abs/1903.02930, 2019.

[49] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," *ArXiv*, vol. abs/1512.02167, 2015.

[50] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *EMNLP*, 2019, pp. 5099–5110.

[51] S. Pramanik, P. Agrawal, and A. Hussain, "Omninet: A unified architecture for multi-modal multi-task learning," *ArXiv*, vol. abs/1907.07804, 2019.

[52] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *EMNLP*, 2016, pp. 457–468.

[53] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *ICLR*, 2017, pp. 247–263.

[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[55] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[56] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *ICASSP*, 2017, pp. 131–135.

[57] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017, pp. 776–780.

[58] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *NIPS*, 1998, pp. 570–576.

[59] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *CVPR*, 2015, pp. 3460–3469.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[61] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," *ICML*, vol. 28, no. 3, pp. 1247–1255, 2013.

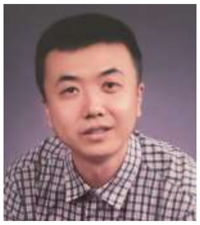
[62] Y. Tian, D. Li, and C. Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," in *ECCV*, 2020, pp. 436–454.

Shuo Liu is currently pursuing the M.S. degree in the School of Artificial Intelligence, University of Chinese Academy of Sciences. She received the B.S. degree of information security from Hunan University in 2019. Her research interests include image processing and multi-modal learning.

Weize Quan received his Ph.D. degree from Institute of Automation, Chinese Academy of Sciences (China) and Université Grenoble Alpes (France) in 2020, and his Bachelor's degree from Wuhan University of Technology in 2014. He is currently an assistant professor at the National Laboratory of Pattern Recognition of the Institute of Automation, Chinese Academy of Sciences. His research interests include computer graphics and image processing.

Chaoqun Wang received his M.S. degree from School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, in 2020, and his Bachelor's degree from Jianghan University, Wuhan, in 2018. His research interests include image processing and computer vision.

Yuan Liu received her B.S degree in Microelectronics and M.S degree in Information and Communication Engineering from Shanghai Jiao Tong University, China in 2012 and 2015. She is currently a staff algorithm engineer in DAMO Academy of Alibaba. Her current research interests include speech recognition, sound event detection and deep learning.



Bin Liu received his B.S. degree and M.S. degree from Beijing Institute of Technology, Beijing, China, in 2007 and 2009 respectively. He received Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015. He is currently an Associate Professor in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include affective computing and audio signal processing.

Dong-Ming Yan is a professor in National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences(CAS). He received his Ph.D. degree in computer science from Hong Kong University in 2010, and his master and bachelor degrees in computer science and technology from Tsinghua University in 2005 and 2002, respectively. His research interests include image processing, geometric processing, and visualization.