MM-Pyramid: Multimodal Pyramid Attentional Network for Audio-Visual Event Localization and Video Parsing

Jiashuo Yu
School of Computer Science
Fudan University, China
jsyu19@fudan.edu.cn

Ying Cheng
Academy for Engineering and Technology, Fudan University, China
chengy18@fudan.edu.cn

Rui-Wei Zhao
Academy for Engineering and Technology, Fudan University, China
rwzhao@fudan.edu.cn

Rui Feng
School of Computer Science
Fudan University, China
fengrui@fudan.edu.cn

Yuejie Zhang
School of Computer Science
Fudan University, China
yjzhang@fudan.edu.cn

ABSTRACT
Recognizing and localizing events in videos is a fundamental task for video understanding. Since events may occur in auditory and visual modalities, multimodal detailed perception is essential for complete scene comprehension. Most previous works attempted to analyze videos from a holistic perspective. However, they do not consider semantic information at multiple scales, which makes the model difficult to localize events in different lengths. In this paper, we present a Multimodal Pyramid Attentional Network (MM-Pyramid) for audio-visual event localization and video parsing. Specifically, we first propose the attentive feature pyramid module. This module captures temporal pyramid features via several stacking pyramid units, each of them is composed of a fixed-size attention block and dilated convolution block. We also design an adaptive semantic fusion module, which leverages a unit-level attention block and a selective fusion block to integrate pyramid features interactively. Extensive experiments on audio-visual event localization and weakly-supervised audio-visual video parsing tasks verify the effectiveness of our approach.

CCS CONCEPTS
• Computing methodologies → Scene understanding.

KEYWORDS
Audio-Visual Learning, Pyramid Network, Event Localization.

ACM Reference Format:
Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng*, and Yuejie Zhang*. 2022. MM-Pyramid: Multimodal Pyramid Attentional Network for Audio-Visual Event Localization and Video Parsing. In Proceedings of ACM MULTIMEDIA CONFERENCE 2022 (MM’2022). ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547869

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM’2022, October 10–14, 2022, Lisboa, Portugal © 2022 Association for Computing Machinery. ACM ISBN 978-1-4503-9203-7/22/10...$15.00 https://doi.org/10.1145/3503161.3547869

*indicates corresponding authors. Rui Feng and Yuejie Zhang are also with Shanghai Collaborative Innovation Center of Intelligent Visual Computing. Rui Feng is also with Academy for Engineering and Technology, Fudan University.

1 INTRODUCTION
Video scene understanding in computer vision is fundamental for many real-world applications and it simulates the information perception process of human brain. According to the researches in cognitive neuroscience [6, 14], human perceives information from multiple modalities to obtain the overall comprehension. Similar to human brains, auditory and visual data can provide complementary cues from different perspectives for machine video understanding.

In recent years, some works [2–4, 7, 11, 21, 27, 28, 31] focus on the synergistic effects between auditory and visual modalities and acquire the joint multimodal representation, and some other works [1, 18, 49, 50] investigate on localizing sounding objects via self-supervised methods. However, the analysis of audio-visual events in videos, which is a crucial part of the video scene perception, is also in need of investigation. To this end, some tasks...
and corresponding methods are proposed to explore the impact of audio-visual cues on events. Specifically, Tian et al. [37] propose the audio-visual event localization task, which aims to classify and temporally localize an audio-visual event in a video clip. As shown in Fig. 1(a), the audio-visual event *frying* cannot be seen in the first and last two segments, thereby labeled as *background*. In the other seconds, the food frying can both be heard and seen, thus we label these as *frying*. To make this task more generalizable, Tian et al. [36] expand the task of localizing one event to multiple events scenarios and introduce the audio-visual video parsing task, which is illustrated in Fig. 1(b), given a video that includes several audible, visible, and audi-visible events, the audio-visual video parsing task aims to predict all event categories, distinguish the modalities perceiving each event, and localize their temporal boundaries. Since the process of labeling all event boundaries is cumbersome, this task is conducted in a weakly-supervised manner, which makes it more generalizable to real-world applications yet more challenging.

Some researchers tackle these problems by capturing contexts from a holistic perspective. For audio-visual event localization, prior works [24, 32, 37, 43–45, 51] explore the relationship between auditory and visual sequences via different kinds of attention mechanisms. For audio-visual video parsing, Tian et al. [36] propose a hybrid attention network to capture temporal context, which tends to focus more on the holistic content and is capable to detect the major event throughout the video. However, these methods are limited by some cases including when the lengths of target events are short, or videos include several events that have miscellaneous lengths. Since they focus more on the coarse-grained holistic content, detailed information is inclined to be neglected, which makes it difficult to localize short-term events. Despite several methods [41, 46, 48] proposed to capture temporal pyramid features, they can only tackle uni-modal scenarios and lack multi-modal interactions. Therefore, the necessity of exploring features both in different granularities and modalities emerges, which helps to localize multimodal events in different temporal sizes accurately and further leads to a comprehensive video understanding.

In this paper, we introduce a novel Multimodal Pyramid Attentional Network (MM-Pyramid). To be specific, we first propose a novel attentive feature pyramid module composed of multiple pyramid units to acquire multi-level audio-visual features. In each pyramid unit, a fixed-size multi-scale attention block captures intra- and inter-modality interactions, together with a dilated convolution block to integrate segment-wise features and derive semantic information. To fuse pyramid units, we also design an adaptive semantic fusion module. This module explores the correlations among multi-level features and integrates pyramid units in a selective fusion way. By this means, the model can obtain more targeted representation, thereby resulting in better audio-visual event localization and video parsing performance. In summary, our contributions are as follows:

- We propose to exploit audio-visual pyramid features to learn multi-scale semantic information and localize events in different lengths, which is beneficial for a comprehensive video scene understanding.
- We develop a novel Multimodal Pyramid Attentional Network, which consists of an attentive feature pyramid module and an adaptive semantic fusion module to capture and integrate multi-level features, respectively.
- We conduct extensive experiments on two audio-visual tasks: audio-visual event localization on the AVE [37] dataset and weakly-supervised audio-visual video parsing on the LLP [36] dataset to verify the effectiveness of our proposed framework.

## 2 RELATED WORKS

### 2.1 Audio-visual representation learning

Audio-visual representation learning aims to acquire the informative multimodal representation by exploiting the correlations between auditory and visual modalities. Some works [2, 4, 7, 11, 21, 27, 28, 31] try to obtain the joint audio-visual representation by learning the correspondence of audio and visual streams in a self-supervised manner. Others [3, 17] leverage unsupervised clustering as the supervision to explore the cross-modal correlation. Besides, some other works [1, 10, 11, 49, 50] explore the relationship between the sound and dynamic motions of objects and enhance the capability of object localization. In this paper, we try to leverage the correlation between audio and visual content to enhance the performance of downstream applications.

### 2.2 Audio-visual event localization and video parsing

Audio-visual event localization [37] utilizes the synergy and relevance between auditory and visual streams to temporally localize events in the given video. Most prior works [24, 32, 37, 43–45] leverage the attention-based architecture to capture inter- and intra-modality interactions for holistic video understanding. Yu et al. [47] explores the differences of video-level classification and segment-level localization, and propose a multimodal parallel network to decrease the conflicts between global and local features. More recently, Zhou et al. [51] proposes a positive sample propagation strategy to utilize positive audio-visual pairs, thereby learning discriminative features for the classifier.

To expand the event localization task to multi-event scenarios, Tian et al. [36] propose a more generalizable and challenging task named audio-visual video parsing, which aims to classify and locate all audible, visible, and audi-visible events inside a video in a weakly-supervised manner. They also propose a hybrid attention network to capture multimodal contexts and a multimodal multiple instance learning method for the weakly supervised setting. Wu et al. [42] propose to obtain accurate modality-aware event supervision by swapping audio and visual tracks with other unrelated videos to address the modality uncertainty issue. In this paper, we propose to explore multi-scale audio-visual features for localizing events in multiple lengths, which is neglected by previous methods.

## 3 TASK FORMULATION

**Audio-visual event localization** aims to classify and localize an audio-visual event in a given video. The task can be tackled in the fully-supervised and weakly-supervised manners. For the fully-supervised setting, the event label for the $t_{th}$ video segment is given as $y_p = \{y_{tp}^p\}_{t=1}^C \in \{0, 1\}$, $p = 1, \ldots, C$, $\sum_{p=1}^C y_{tp}^p = 1$, where $C$ is the total number of audio-visual events plus one background.
category, while in the weakly-supervised setting, only video-level event categories are given during training yet temporal boundaries are still required during inference.

**Weakly-supervised audio-visual video parsing** aims to predict all event categories, distinguish the modalities perceiving each event, and localize their temporal boundaries. Given a video sequence \( \{V_t, A_t\}_{t=1}^N \) with \( N \) non-overlapping temporal segments, the event labels are given as \( y_t = \{ (y_t^a, y_t^v) \mid \{y_t^a\}_m, \{y_t^v\}_m \in \{0, 1\}, m = 1, ..., C \} \), where \( C \) is the total number of event categories. An event is labeled as audi-visible only when it is both audible and visible, thus the audi-visible label can be computed as \( y_t^{av} = y_t^a \cdot y_t^v \).

This task is conducted in a weakly supervised manner. We only have all event categories that appeared in the given video for training, but need to predict which segments contain those events and which modalities perceive them during inference.

### 4 METHODOLOGY

In this section, we introduce our Multimodal Pyramid Attentional Network, which is shown in Fig. 2. We first propose the attentive feature pyramid module to obtain temporal pyramid features, which is introduced in Sec. 4.1. Then we propose an adaptive semantic fusion module for an interactive pyramid feature fusion in Sec. 4.2.

#### 4.1 Attentive Feature Pyramid Module

The attentive feature pyramid module is composed of a few stacked units in different scales. Pyramid units in different modalities are connected interactively. The detailed structure of two linked audio and visual pyramid units in the same size are shown in Fig. 3. In each unit, we first propose the fixed-size attention mechanism to introduce intra- and inter-modality interactions, then perform feature integration via a dilated residual convolution block. The size of each unit is different and the outputs of all units are preserved as pyramid-like multimodal features.

**Attentive feature interaction.** Self-attention (SA) and cross-modal attention (CMA) are used to provide temporal feature interactions. We reform the encoder part of Transformer [39]. Specifically, the attention scores between different video snippets is computed by the scaled dot-product attention \( att(q, k, v) = \text{softmax}(\frac{qk^T}{\sqrt{d}})v \), where \( q, k, v \) denotes the query, key, and value vectors, \( d_m \) is the dimension of query vectors, \( T \) denotes the matrix transpose operation. The self-attention block learns uni-modal temporal relationships via \( \text{sat}(f) = att(fW_q, fW_k, fW_v), \) where \( W_q, W_k, W_v \) are learnable parameters, \( f \) is the input feature. For the cross-modal attention block, we assign features in current modality as the query vectors, while the key and value vectors are from features of the other modality. The formulations can be defined as \( \text{cma}(f_a, f_v) = att(f_aW_q, f_aW_k, f_aW_v) \) and \( \text{cma}(f_a, f_v) = att(f_aW_q, f_aW_k, f_aW_v) \), where \( f_a \) is the audio feature, \( f_v \) is the video feature, \( W_q, W_k, W_v \) are learnable parameters. The parameter matrices in the cross-modal attention block are shared. This parameter-efficient setting can project audio and visual features into the same subspaces, which facilitates further interactions of uni-modal and multimodal features. Then the features...
are processed by a feed-forward layer. We adopt the layer normalization [5] for regularization, and the residual connections [15] for the identity mapping to avoid overfitting.

Since we intend to obtain pyramid features for localizing different lengths of events, the temporal interacting size of each unit ought to be diverse. To this end, we set an interaction window to constrain the interacting size of the self-attention and cross-modal attention layer. Concretely, we propose a fixed-size attention block as shown in Fig. 4, which restricts the interaction windows by adding masks to areas that should not be involved. In this way, the fixed-size attention can be computed as follows,

$$sa(f, d) = att(fW_q, S_t(f, d)W_k, S_t(f, d)W_v),$$

(1)

$$cma(f_a, f_d, d) = att(f_0W_q, S_1(f, d)W_k, S_1(f, d)W_v),$$

(2)

$$cma(f_d, f_0, d) = att(f_0W_q, S_1(f, d)W_k, S_1(f, d)W_v),$$

(3)

$$S_t(x, d) = [x_{t-d}, ..., x_{t+d}].$$

(4)

where $S(t)$ indicates creating interaction windows for the $t$-th segment, $d$ denotes the size of the interaction window.

Different from some prior works [7, 36] where self-attention and cross-modal attention blocks are connected in serial, we adopt a parallel arrangement. The inputs of the two kinds of attention blocks are both the output of the previous pyramid unit. We then leverage the channel-wise attention to interconnect and integrate uni-modal and multimodal features. To be specific, the output of self-attention and cross-modal attention blocks are firstly concatenated along the channel dimension. Then channel-wise attention scores are computed to refine raw features via a linear layer followed by a sigmoid function. The final fused features are calculated by the summation of the refined uni-modal and multimodal features. The fusion process of visual modality is formulated as below,

$$F_{fused} = σ(W_{sa}F^v_{sa} + b_1)F^a_{a} + σ(W_{cma}F^v_{cma} + b_2)F^v_{cma},$$

(5)

$$F^v_c = Concat(F^v_{sa}, F^v_{cma}).$$

(6)

**Figure 3: Detailed structure of two linked audio and visual pyramid units in the same level. $x$ is the channel-wise multiplication. SA and CMA denote self-attention and cross-modal attention.**

![Pyramid Units](image)

**Figure 4: Detailed structure of the fixed-size attention mechanism (the size of the interaction window is 2).**

where $W_{sa}, W_{cma}, b_1, b_2$ are learnable weights. The formulation of auditory modality fusion is highly similar, thus we omit it for concise writing.

**Dilated temporal convolution.** Directly utilizing outputs of the fixed-size attention will lead to two problems: First, though interactions among temporal segments have been performed sufficiently, features still need to be amalgamated in temporal dimension to perceive semantic information. Second, since positional encoding is not performed in the attention blocks, temporal order of the sequence has not been modeled, which is important for event understanding. Therefore, a temporal convolution block is used to inject positional information and derive semantic representation.

Temporal convolutional network has been widely applied in speech synthesis [30] and action segmentation [9, 22, 23]. They can provide multi-grained information via multiple dilated convolution layers. The dilation size of each layer is increasing exponentially, which expands the receptive fields at each layer, thus the network can focus on information in distinct temporal lengths. Following [23], we adopt the dilated residual block for our temporal convolution block. Each dilated residual block contains a 3×3 dilated convolution, a ReLU [13] activation, a 1×1 convolution, and the residual connection. Moreover, instead of causal convolution adopted in some temporal forecasting tasks, we use acausal convolution with kernel size 3 since it can take more contextual information of the current segment into consideration. The operations in each dilated residual block can be described as follows,

$$\hat{F}^l_t = ReLU(W^l_1F^l_t + W^l_2F^l_{t-1} + W^l_3F^l_{t-2} + b_3),$$

(7)

$$F^l_t = \hat{F}^l_{t-1} + V + \hat{F}^l_t + b_4,$$

(8)

where $F^l_t$ is the output of the $t$-th segment in the $l$-th pyramid unit, $d$ denotes the dilated size, $\{W^l_i\}_{i=1}^4 \in \mathbb{R}^{D \times D}$ are convolution filter parameters, $b_3 \in \mathbb{R}^D$ is the bias vector, $*$ denotes the $1 \times 1$ convolution operation, $V \in \mathbb{R}^{D \times D}$ and $b_4 \in \mathbb{R}^D$ are convolutional weights and bias.

We keep the dilation size of each unit equal to the interactive size of the fixed-size attention mechanism. Besides, the size of each unit is different, thereby guaranteeing that features obtained by these units contain contexts in multiple scales. Finally, we preserve the
output of all units \( \{ F^i_t, F^i_s \}_{i=1}^L \) as the multimodal pyramid features, where \( L \) is the total number of pyramid units.

### 4.2 Adaptive Semantic Fusion Module

A simple way to integrate pyramid features is to conduct pooling over the unit level. However, these pooling methods lack interactions between different levels, resulting in the incompatibility of multi-scale semantic information. To address this problem, we put forward an adaptive semantic fusion module, which consists of a unit-level attention block to explore the correlation of pyramid features and a selective fusion block for adaptively feature integration.

**Unit-level attention.** Since the interaction size of each pyramid unit is strictly restricted, pyramid units focus on areas in distinct scopes and generate semantic information at different levels, which results in a large semantic gap between pyramid features. To this end, we introduce the unit-level attention to provide contextual interactions and refine the pyramid features. The unit-level attention considers the similarity of contents in different units, and the interacting results vary with the characteristics of the captured features. The interacting process can be formulated as follows,

\[
r_t = \text{att}(\overline{T}_t, W_q, \overline{T}_t W_b, \overline{T}_t W_o),
\]

where \( \overline{T}_t \in \mathbb{R}^{L \times D} \) is the outputs of all pyramid units in the \( t \)-th segment, \( W_q, W_b, W_o \) are learnable parameters.

**Selective fusion.** Since the type and length of events are uncertain, the model is supposed to pay more attention to features at suitable levels. Therefore, outputs of pyramid units are fused by a selective fusion block after building the relation-aware connections. Specifically, we perform a linear projection on each modality to gain the fusion weights of each unit. By doing so, the selective fusion block can dynamically assign weights on pyramid features in different granularities, and the fusion results vary with the characteristics and the event's type of the given video. The fused features are computed by the weighted summation,

\[
\hat{r}_t = \sum_{l=1}^L w^l_t r^l_t,
\]

\[
w^l_t = \sigma(W_q f^l_t + b_s),
\]

where \( W_q \) and \( b_s \) denote the linear projection parameters, \( \sigma \) denotes the sigmoid function. We use sigmoid instead of softmax since the characteristics of pyramid features are not mutually exclusive.

Finally, the probabilities in each level can be computed by the sigmoid and the softmax function for multiple events and single event, respectively. For audio-visual events, since the event is both audible and visible in the same segment, the event probability in the \( t \)-th segment \( p^a_t \) can be computed by the logical conjunction of uni-modal predictions, which is formulated as \( p^a_t = p^a_t \wedge p^v_t \).

### 5 EXPERIMENTS

#### 5.1 Audio-Visual Event Localization

**Dataset and metrics.** Audio-Visual Event (AVE) [37] dataset is an audio-visual event dataset with 4,183 video clips of 29 classes. Each video clip is 10 seconds with event category annotation per second. We follow the original setting as [37] and divide the dataset as 80%/10%/10% for training, validation, and test, respectively. The overall segment-wise accuracy is used for evaluation, which is the percentage of all matching segments.

**Implementation details.** Since there is only one audio-visual event in a video, we decompose the task into video-level category predictions and segment-level relevance predictions as prior methods [43, 44, 47] do. Video-level categories are predicted by a temporal average pooling layer followed by a linear classifier, while segment-level relevance predictions are obtained by a segment-wise event-related binary classifier. We employ the VGG-19 [34] network pre-trained on ImageNet [8] and the VGGish [16] network pre-trained on AudioSet [12] for feature extraction. We use Adam [19] as optimizer. The initial learning rate is 2e-5 and divided by 10 after 50 epochs. We complement more details in Appendix A.

**Comparison with the state-of-the-arts.** We compare our model with all prior methods as shown in Tab. 1. Results show that our model achieves comparable results with the state-of-the-art method PSP [51]. Since more than 60% of events occur throughout the entire video in this dataset, the advantages of our model for detecting events of different lengths are not fully embodied in the full AVE dataset. Therefore, we conduct additional experiments on the subset where events are in multiple lengths. Results show that our model obtains higher performance on the multiple length events subset both in fully and weakly supervised settings. This proves our declaration that our model can detect more events with different lengths via the pyramid setting.

#### 5.2 Audio-Visual Video Parsing

**Dataset and metrics.** Look, Listen, and Parse (LLP) [36] dataset derived from AudioSet [12] is constructed for the audio-visual video parsing task. It contains 11,849 videos of 25 event categories. Each

---

**Table 1: Overall accuracy (%) compared with prior methods in both fully and weakly supervised manner. * denotes results re-implemented by the same feature extractor. Subset means the subset of the AVE dataset where events have multiple lengths and do not occur throughout the whole video.**

| Method               | Sup. (%) | W-Sup. (%) |
|----------------------|----------|------------|
| AVE [37]             | 68.6     | 66.7       |
| AVDSN [24]           | 72.6     | 67.3       |
| DAM [43]             | 74.5     | -          |
| AVRB [33]            | 74.8     | 68.9       |
| AVIN [32]            | 75.2     | 69.4       |
| CMAN [45]            | 73.3*    | 70.4*      |
| AVT [25]             | 76.8     | 70.2       |
| MPN [47]             | 77.6     | 72.0       |
| CMRAN [44]           | 77.4     | 72.9       |
| PSP [51]             | 77.8     | 73.5       |
| MM-Pyramid (Ours)    | 77.8     | 73.2       |
| MPN on Subset        | 62.1     | 53.8       |
| CMRAN on Subset      | 62.3     | 54.2       |
| PSP on Subset        | 62.7     | 54.4       |
| MM-Pyramid (Ours) on Subset | 63.9     | 55.3       |

---

**MM-Pyramid: Multimodal Pyramid Attentional Network for Audio-Visual Event Localization and Video Parsing MM'2022, October 10–14, 2022, Lisboa, Portugal**
Table 2: Audio-visual video parsing F-score results (%) in comparison with recent weakly-supervised methods.

| Event Type | Methods          | Segment Level | Event Level |
|------------|------------------|---------------|-------------|
| Audio      | Kong et. al 2018 [20] | 39.6          | 29.1        |
|            | TALNet [41]      | 50.0          | 41.7        |
|            | AVE [37]         | 47.2          | 40.4        |
|            | AVDSN [24]       | 47.8          | 34.1        |
|            | HAN [36]         | 60.1          | 51.3        |
|            | Ours             | 60.9          | 52.7        |
|            | HAN+MA [42]      | 60.3          | 53.6        |
|            | Ours+MA          | 61.1          | 53.8        |
| Visual     | STPN [29]        | 46.5          | 41.5        |
|            | CMCS [26]        | 48.1          | 45.1        |
|            | AVE [37]         | 44.9          | 34.7        |
|            | AVDSN [24]       | 52.9          | 46.3        |
|            | HAN [36]         | 52.9          | 48.9        |
|            | Ours             | 54.4          | 51.8        |
|            | HAN+MA [42]      | 60.0          | 56.4        |
|            | Ours+MA          | 60.3          | 56.7        |
| Audio-Visual| AVE [37]        | 35.4          | 31.6        |
|            | AVDSN [24]       | 37.1          | 26.5        |
|            | HAN [36]         | 48.9          | 43.0        |
|            | Ours             | 50.0          | 44.4        |
|            | HAN+MA [42]      | 55.1          | 49.0        |
|            | Ours+MA          | 55.8          | 49.4        |
| Type@AV    | AVE [37]         | 39.9          | 35.5        |
|            | AVDSN [24]       | 45.7          | 35.6        |
|            | HAN [36]         | 54.0          | 47.7        |
|            | Ours             | 55.1          | 49.9        |
|            | HAN+MA [42]      | 58.9          | 53.0        |
|            | Ours+MA          | 59.7          | 54.1        |
| Event@AV   | AVE [37]         | 41.6          | 36.5        |
|            | AVDSN [24]       | 50.8          | 37.7        |
|            | HAN [36]         | 55.4          | 48.0        |
|            | Ours             | 57.6          | 50.5        |
|            | HAN+MA [42]      | 57.9          | 50.6        |
|            | Ours+MA          | 59.1          | 51.2        |

Event@AV is generated by considering all types of events in each video and computing the composite F-score results.

**Implementation details.** The outputs of MM-Pyramid $p^i_t$, $p^u_t$, and $p^m_t$ represent the uni-modal and multimodal parsing results. Since this task is performed in a weakly supervised manner, we follow [36] to leverage an attentive MMIL pooling to generate video-level predictions. We also use label smoothing [35] to alleviate label noises of the weakly supervised setting. Following [36], we employ ResNet-152 [15] and R(2+1)D [38] to extract visual features, and VGGish network to extract audio features. We use Adam [19] optimizer and set the learning rate as $1e$-4, which is degraded by a factor of 5 after 10 epochs. More details are listed in Appendix B.

**Comparison with the state-of-the-arts.** We compare with the state-of-the-art methods HAN [36], AVE [37] and AVDSN [24], as well as several competitive weakly supervised event detection methods, which is shown in Tab. 2. To be specific, we choose temporal action localization methods STPN [29] and CMCS [26], sound event detection methods TALNet [40] and Kong et al. [20]. For the new modality-aware method MA [42], since they do not conduct optimization from the network perspective and use the same hybrid attention network (HAN) as [36], our method is not mutually exclusive with their strategy. Therefore, we provide results both using the raw training strategy and the new label refinement and contrastive learning strategy. Results show that our model outperforms baseline methods on all evaluation metrics in a large margin. For the raw training strategy, our model yields up to 2.9% higher on the unimodal metrics (Visual&Event-level) and up to 2.5% higher on the multimodal metrics (Event@AV&Event-level). This proves that the insight of capturing and integrating multimodal pyramid features enables the localization of events in multiple lengths precisely, which further results in better video parsing performance.

5.3 **Ablation Studies**

**Do pyramid units help?** We first investigate the impact of the pyramid units. As shown in Tab. 3, "MM-Pyramid-Last" means only the output of the last pyramid unit is preserved. "MM-Unpyramid" denotes the sizes of all pyramid units are identical and equal to the size of the last pyramid units in the raw framework. "Hybr-Trans w/PE" denotes the 4-layer hybrid transformer encoder with positional encoding. Results show that our full model outperforms all ablated models, indicating the significance of capturing multi-level features via our pyramid settings. We argue that the multimodal information learned by different levels of contexts resolves videos at multiple granularities. We also make comparisons among different transformer-based structures, which are shown in Appendix B.

**Does dilated residual convolution help?** We also explore the efficacy of our dilated convolution block. "MM-Pyramid w/o conv" indicates that the entire dilated block is removed, "MM-Pyramid w/o residual" means the dilated residual block is replaced with a single $3 \times 3$ convolution layer. Without dilated convolution block, the performance declines significantly, showing the integration of pyramid features is necessary for semantics information. The performance also declines when using the vanilla convolution layer, proving that the residual connections and $1 \times 1$ convolution can enhance the expressiveness of integrated features.
Table 3: Ablation studies with different components on the audio-visual video parsing task. We propose several variants to investigate the impact of multimodal pyramid setting, attentive feature pyramid module, and adaptive semantic fusion module.

| Model               | Audio | Visual | Audio-Visual | Type@AV | Event@AV |
|---------------------|-------|--------|--------------|---------|----------|
|                     | Seg   | Eve    | Seg          | Eve     | Seg      | Eve      | Seg | Eve |
| MM-Pyramid-Last     | 59.6  | 51.4   | 53.6         | 50.1    | 49.2     | 43.2     | 54.1 | 48.2 | 56.4 | 48.6 |
| MM-Unpyramid        | 59.7  | 49.6   | 53.3         | 50.0    | 47.7     | 41.5     | 53.6 | 47.0 | 57.4 | 48.1 |
| Hybr-Trans w/ PE    | 60.1  | 51.9   | 53.0         | 50.1    | 48.4     | 43.7     | 54.5 | 47.6 | 56.1 | 48.5 |
| MM-Pyramid w/o conv | 60.0  | 50.8   | 52.4         | 49.2    | 47.9     | 41.7     | 53.1 | 47.0 | 56.4 | 47.8 |
| MM-Pyramid w/o residual | 60.4  | 51.5   | 52.5         | 49.5    | 47.7     | 41.8     | 53.5 | 47.6 | 57.1 | 48.7 |
| MM-Pyramid w/o ULA  | 60.6  | 52.1   | 53.4         | 49.8    | 48.8     | 43.6     | 54.3 | 48.5 | 56.8 | 49.2 |
| MM-Pyramid w/o SF   | 60.5  | 51.8   | 53.6         | 49.9    | 48.7     | 43.1     | 54.3 | 48.3 | 56.9 | 49.1 |
| MM-Pyramid (full)   | 60.9  | 52.7   | 54.4         | 51.8    | 50.0     | 44.4     | 55.1 | 49.9 | 57.6 | 50.5 |

Does adaptive semantic fusion module help? Multimodal pyramid features are fused interactively. To reveal the contribution of our adaptive semantic fusion module, we propose two ablated models “MM-Pyramid w/o ULA” and “MM-Pyramid w/o SF”. The first model is constructed by removing the unit-level attention, while the other fuses pyramid features via an average pooling layer instead of the selective fusion block. The performance of “MM-Pyramid w/o ULA” declines, which indicates that building the relation-aware connections of multi-scale features is effective. Our model also outperforms “MM-Pyramid w/o SF”. We argue that this proves the insight of integrating pyramid features selectively helps the acquisition of complete video scene understanding.

Impact of parameter sharing setting. To investigate the performance of the parameter sharing strategy of cross-modal attention blocks, we complement a non-sharing ablated model as shown in Tab. 4. The result indicates that the sharing matrices get comparable performance with lower computing complexity compared with the non-sharing setting.

Table 4: Segment-level and event-level f1-scores(%) comparison with different parameter sharing strategy.

| Model           | Type@AV | Event@AV |
|-----------------|---------|----------|
|                 | Segment | Event    | Segment | Event |
| Non-Sharing     | 54.8    | 50.1     | 57.2    | 50.0  |
| Sharing (ours)  | 55.1    | 49.9     | 57.6    | 50.5  |

5.4 Qualitative Results

Role of selective fusion. To illustrate the effectiveness of our selective fusion block in the attentive semantic fusion module, we conduct qualitative results as shown in Fig. 6. The sample video in the picture consists of some long events as well as a small audio-visual event “speech”. It should be noticed that since the characteristics of pyramid features are not mutually exclusive, we use the sigmoid function as the substitution of softmax to generate fusion weights, thus the sum of weights is not 1. Results show that the selective fusion module assigns relatively high scores on the pyramid units of large scales, which indicates the effectiveness of our feature integration method. The sample video below the picture consists of one visual event of medium length yet several audio events in miscellaneous lengths. Therefore, the selective fusion block focuses more on the visual pyramid units with medium lengths and disperses weights into all audio pyramid units.

Capability of detecting events in multiple lengths. We also illustrate our model’s capability of capturing multiple events. As shown in Fig. 5, we conduct qualitative experiments on the audio-visual video parsing task in comparison with HAN [36]. The green labels are the ground truths, and the yellow and blue labels indicate the predictions of HAN and our model, respectively. We found that though the HAN model can precisely predict events that exist throughout the whole video, it fails to detect a short-term event.
Figure 6: Qualitative comparison with the weakly-supervised audio-visual video parsing method HAN. The red dotted box includes the visual predictions, and the audio predictions are in the purple dotted box. The green, yellow, and blue labels denote the ground-truth, predictions of HAN, and predictions of our MM-Pyramid, respectively.

(singing audio events from 4th to 10th seconds) and predicts incorrect event temporal boundary (the speech event in the first second). However, our MM-Pyramid model tends to recognize all events of different sizes and provide predictions with only small deviations (one-second errors in speech and singing events). This result reveals that our model is capable of exploring features in different granularities, which further leads to localizing events in diverse lengths precisely. We provide more qualitative results in Appendix E, including additional visualization results and error analysis.

6 LIMITATION
Though our MM-Pyramid framework shows the efficacy of detecting multiple events in different lengths, the advantage is limited when detecting events that occur throughout the whole video compared with other methods. This can be shown in the experimental results of the audio-visual event localization task, in which task the majority (66.4%) of events span over the whole video. In that situation, we suggest injecting our model or our multimodal pyramid feature methodology into other single-shot event detection methods as an enhancement for detecting multiple events. To this end, finding a flexible way to assemble our proposed multimodal pyramid paradigm with some widely-adopted temporal localization methods could be a promising research direction.

7 CONCLUSION
In this paper, we propose a novel Multimodal Pyramid Attentional Network (MM-Pyramid) for audio-visual event localization and weakly-supervised audio-visual video parsing. Our model captures and integrates multimodal pyramid features in distinct temporal scales for comprehensive scene understanding. To acquire features in different granularities, we propose a novel attentive feature pyramid module, which is composed of the fixed-size attention mechanism and dilated convolution block. Furthermore, we propose an adaptive semantic fusion module to refine and fuse pyramid features in an interactive and selective way. Extensive experiments on the AVE and LLP datasets demonstrate the effectiveness of our proposed approach on localizing events in multiple lengths. In future works, we plan to expand our multimodal pyramid architecture to more audio-visual scenarios such as violence detection, representation learning, and multimodal reasoning.

ACKNOWLEDGMENTS
This work was supported by National Natural Science Foundation of China (No. 62172101, No. 61976057). This work was supported (in part) by the Science and Technology Commission of Shanghai Municipality (No. 21511101000, No. 21511100602), and SPMI Innovation and Technology Fund Projects (SAST2020-110)
REFERENCES

[1] Triantafyllos Aifouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. 2020. Self-Supervised Learning of Audio-Visual Objects from Video. In ECCV.

[2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. arXiv preprint arXiv:2006.16228 (2020).

[3] Hurnam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Chanem, and Du Tran. 2020. Self-supervised learning by cross-modal audio-video clustering. In NeurIPS, Vol. 33.

[4] Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In ICCV. 649–657.

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. arXiv preprint arXiv:1607.06450 (2016).

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

[7] Ying Cheng, Ruize Wang, Zhaohao Pan, Rui Feng, and Yuejie Zhang. 2020. Look, listen and attend: Co-attention network for self-supervised audio-visual representation learning. In ACM MM. 3884–3892.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In CVPR. 248–255.

[9] Yazan Abu Farha and Jurgen Gall. 2019. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In CVPR. 3575–3584.

[10] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. 2020. Music gesture for visual sound separation. In CVPR. 10478–10487.

[11] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. 2019. Self-supervised moving vehicle tracking with stereo sound. In ICCV. 7055–7062.

[12] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and humandabeled dataset for audio events. In ICASSP. 776–780.

[13] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In Proceedings of the fourteenth international conference on artificial intelligence and statistics. 315–323.

[14] Aviva I Goller, Leun J Otten, and Jamie Ward. 2009. Current opinion in neurobiology 19, 6 (2009), 10478–10487.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In CVPR. 770–778.

[16] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Pratt, Rf A Saurous, Bryan Seyboldt, et al. 2017. CNN architectures for large-scale audio classification. In ICASSP. 131–135.

[17] Di Hu, Fenqing Nie, and Xielong Li. 2019. Deep multimodal clustering for unsupervised audiovisual learning. In CVPR. 9248–9257.

[18] Di Hu, Rui Qian, Minuye Jiang, Xiao Tan, Shilie Wen, Erruz Ding, Weiyao Liu, and Dejing Dou. 2020. Discriminative Sounding Objects Localization via Self-supervised Audio-Visual Matching. In NeurIPS, Vol. 33.

[19] Diederek P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

[20] Quqiang Kong, Yong Xu, Wenhui Wang, and Mark D Plumbley. 2018. Audio-visual scene analysis with cross-modal models. In ICASSP.

[21] Shuang Ma, Xingyue Chen, Zhipeng Li, and Jiaying Yu. 2016. Rethinking the inception architecture for computer vision. In CVPR. 2818–2826.

[22] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In CVPR. 2970–2979.

[23] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Chanem, and Du Tran. 2020. Self-supervised learning by cross-modal audio-video clustering. In NeurIPS, Vol. 33.

[24] Evan Jastrzebski, Jurgen Gall, and Antonio Torralba. 2019. Audio-visual transformer with instance attention for audio-visual event localization. In ACCV.