ABSTRACT

Weakly-supervised audio-visual violence detection aims to distinguish snippets containing multimodal violence events with video-level labels. Many prior works perform audio-visual integration and interaction in an early or intermediate manner, yet overlooking the modality heterogeneity over the weakly-supervised setting. In this paper, we analyze the modality asynchrony and undifferentiated instances phenomena of the multiple instance learning (MIL) procedure, and further investigate its negative impact on weakly-supervised audio-visual learning. To address these issues, we propose a modality-aware contrastive instance learning with self-distillation (MACIL-SD) strategy. Specifically, we leverage a lightweight two-stream network to generate audio and visual bags, in which unimodal background, violent, and normal instances are clustered into semi-bags in an unsupervised way. Then audio and visual violent semi-bag representations are assembled as positive pairs, and violent semi-bags are combined with background and normal instances in the opposite modality as contrastive negative pairs. Furthermore, a self-distillation module is applied to transfer unimodal visual knowledge to the audio-visual model, which alleviates noises and closes the semantic gap between unimodal and multimodal features. Experiments show that our framework outperforms previous methods with lower complexity on the large-scale XD-Violence dataset. Results also demonstrate that our proposed approach can be used as plug-in modules to enhance other networks. Codes are available at https://github.com/JustinYuu/MACIL_SD.

KEYWORDS

Multi-Modality, Contrastive Learning, Violence Detection.

1 INTRODUCTION

Recent years have witnessed the extension of violence detection from a pure vision task [4, 18, 20, 30, 33, 45, 47, 54, 61, 67, 68] to an audio-visual multimodal problem [43, 44, 62], for which the corresponding auditory content supplements fine-grained violent cues. Despite numerous modality fusion and interaction methods have shown promising results, the modality discrepancy of the multiple instance learning (MIL) [38] framework under the weakly-supervised setting remains to be explored.

To alleviate the appetite for fine-labeled data, MIL is widely adopted for the weakly-supervised violence detection, where the output of each video sequence is formed into a bag containing multiple snippet-level instances. In the audio-visual scenarios, all prior works share a general scheme that regards each audio-visual snippet as an integral instance and averaging the top-K audio-visual logits as the final video-level scores. However, we analyze that this formula suffers from two defects: modality asynchrony and undifferentiated instances. Modality asynchrony indicates the temporal inconsistency between auditory and visual violence cues. Taking the typical violent event abuse in Figure 1(a) as an example, when the abuser hits the victim, the scream occurs afterward, and the entire procedure is regarded as a violent event. In this situation, scenes in part of the visual modality (2nd-3rd snippets) and audio modality (4th-5th snippets) contain violent clues. We argue that directly leveraging an audio-visual pair as an instance could introduce data noise to the video-level optimization. The other defect we discovered is undifferentiated instances, that...
is, picking the top-$K$ instances for optimization results in numerous disengaged instances. As shown in Figure 1(b), in a sequence of violent videos, the violent event can be reflected in some audio-visual instances. In contrast, others contain irrelevant elements such as background noises. On the contrary, in the videos of normal events, a few snippets contain elements of normal events, while others include background information. In this case, the $K$-max activation abandons the instances containing background elements, and the discrepancy between violent and normal instances is not explicitly revealed. To this end, we argue that adding contrastive constraints among the violent, normal, and background instances could contribute to the discrimination toward violent content.

Driven by preliminary analysis, we propose a simple yet effective framework constructed by modality-aware contrastive instance learning (MA-CIL) and self-distillation (SD) module. To address the modality asynchrony, we form the unimodal bags apart from the original audio-visual bags, compute unimodal logits, and cluster embeddings of top-$K$ and bottom-$K$ unimodal instances as semi-bags. To differentiate instances, we propose a modality-aware contrastive-based method. In detail, the audio and visual violent semi-bags are constructed as the positive pairs, while the violent semi-bags are assembled with embeddings of instances in the background and normal semi-bags as negative pairs. Furthermore, a self-distillation module is applied to distill unimodal knowledge to the audio-visual model, which closes the semantic gap between modalities and alleviates the data noise introduced by the abundant cross-modality interactions. In summary, our contributions are as follows:

- We analyze the modality asynchrony and undifferentiated instances phenomena of the widely-used MIL framework in audio-visual scenarios, further elaborating their disadvantages for the weakly-supervised audio-visual violence detection.
- We propose a modality-aware contrastive instance learning with self-distillation framework to introduce feature discrimination and alleviate modality noise.
- Equipped with a lightweight network, our framework outperforms the state-of-the-art methods on the XD-Violence dataset, and our model also shows the generalizability as plug-in modules.

2 RELATED WORKS

2.1 Weakly-Supervised Violence Detection

Weakly-supervised violence detection requires identifying violent snippets under video-level labels, where the MIL framework is widely used for denoising irrelevant information. Some previous works [4, 18, 20, 30, 45, 47, 61, 67, 68] regard violence detection as a pure vision task and leverage CNN-based networks to encode visual features. Among these methods, various feature integration and amelioration methods are proposed to enhance the robustness of MIL. Tian et al. [54] propose RTFM, a robust temporal feature magnitude learning method to refine the capacity of recognizing positive instances. Li et al. [33] design a Transformer-based multi-stream learning network to reduce the probability of instance selection errors. However, these models neglect the corresponding auditory information as well as the cross-modality interactions, thereby restricting the performance of violence prediction.

Recently, Wu et al. [62] curate a large-scale audio-visual dataset XD-Violence and establish an audio-visual benchmark. However, they integrate audio and visual features in an early fusion way, thereby limiting further inter-modality interactions. To facilitate multimodal fusion, Pang et al. [43] propose an attention-based network to adaptively integrate audio and visual features with mutual learning module in an intermediate manner. Different from prior methods, we perform inter-modality interactions via a lightweight two-stream network and conduct discriminative multimodal learning via modality-aware contrast and self-distillation.

2.2 Contrastive Learning

Contrastive learning is formulated by contrasting positive pairs against negative pairs without data supervisory. In the unimodal field, several visual methods [10, 23, 25, 35] leverage the augmentation of visual data as a contrast to increase model discrimination. Furthermore, some natural language processing methods utilize the token- and sentence-level contrasts to enhance the performance of pre-trained models [15, 50] and supervised tasks [17, 46]. For the multimodal fields, some works introduce modality-aware contrasts to vision-language tasks, such as image captioning [16, 58], visual question answering [9, 60], and representation learning [34, 49, 59, 66]. Moreover, recent literature [1, 2, 14, 32, 37, 39, 40, 42] utilizes the temporal consistency of audio-visual streams as contrastive pretext tasks to learn robust audio-visual representations. Based on existing instance-level contrastive frameworks [12, 63], we put forward the concept of semi-bags and leverage the cross-modality contrast to obtain model discrimination.
Violent or Normal?

Visual Stream

Audio-Visual Stream

\( \mathbf{y} \)

visual model via two similar lightweight networks, which reduces RGB representations via hallucination networks to address the noise and close the semantic gap between unimodal and multimodal features. The entire framework is trained jointly in a weakly supervised manner, and we adopt the multiple instance learning (MIL) strategy for optimization.

2.3 Cross-Modality Knowledge Distillation

Knowledge distillation is first proposed to transfer knowledge from large-scale architectures to lightweight models [5, 28]. However, the cross-modality distillation aims to transfer unimodal knowledge to multimodal models for alleviating the semantic gap between modalities. Several methods [21, 29] distill depth features to the RGB representations via hallucination networks to address the modality missing and noisy phenomena. Chen et al. [13] propose an audio-visual distillation strategy, which learns the compositional embedding and transfers knowledge across semantic-uncorrelated modalities. Recently, Multimodal Knowledge Expansion [65] is proposed as a two-stage distillation strategy, which transfers knowledge from unimodal teacher networks to the multimodal student network by generating pseudo labels. Inspired by the methodology of self-distillation [6, 8, 11, 19, 52, 64], we propose the parameter integration paradigm to transfer visual knowledge to our audio-visual model via two similar lightweight networks, which reduces the modality noise and benefits robust audio-visual representation.

3 PRELIMINARIES

Given an audio-visual video sequence \( S = (S^A, S^V) \), where \( S^A \) is the audio channel, and \( S^V \) denotes the visual channel, the entire sequence is divided into \( T \) non-overlapping segments \( \{s^A_t, s^V_t\}_{t=1}^{N} \). For an audio-visual pair \( (s^A_t, s^V_t) \), weakly-supervised violence detection task requires to distinguish whether it contains violent events via an event relevance label \( y_t \in \{0, 1\} \), where \( y_t = 1 \) means at least one modality in the current segment includes violent cues. In the training phase, only video-level labels \( y \) are available for optimization. Hence, a general scheme is to utilize the multiple instance learning (MIL) procedure to satisfy the weak supervision.

In the MIL framework, each video sequence \( S \) is regarded as a bag, and video segments \( \{s^A_t, s^V_t\}_{t=1}^{N} \) are taken as instances. Then instances are aggregated via a specific feature-level/score-level pooling method to generate video-level predictions \( p \). In this paper, we utilize the \( K \)-max activation with average pooling rather than attention-based methods [41, 53] and global pooling [51, 67] as the aggregation function. To be specific, given the audio and visual feature \( f_a, f_v \) extracted from CNN networks, we use a multimodal network to generate unimodal logits \( l_a, l_v \) and audio-visual logits \( l_{av} \). The embeddings of audio and visual instances are symbolized as \( h_a \) and \( h_v \). Then we average \( K \) maximum logits and use the sigmoid activation to generate the video-level prediction \( p \). Due to the additional constraint of our proposed contrastive learning method, we define the unimodal bags \( B_a, B_v \). In each unimodal bag, instances are clustered into several semi-bags \( B_{m} \), \( m \in \{a, v\} \) based on their intrinsic characteristics, and the corresponding semi-bag representations are noted as \( \mathbf{B}_m \) for each bag.

4 METHODOLOGY

Our proposed framework consists of three parts, a lightweight two-stream network, modality-aware contrastive instance learning (MA-CIL), and the self-distillation (SD) module. An illustration of our framework shown in Figure 2 is detailed as follows.

4.1 Two-Stream Network

Considering prior methods suffer from the parameter redundancy of the large-scale networks, we design an encoder-agnostic lightweight architecture to achieve feature aggregation and modality interaction. Taking the visual and auditory feature \( f_a, f_v \) extracted by pre-trained networks (e.g., I3D and VGGish for visual and audio features, respectively) as input, our proposed network consists of three
parts, linear layers to keep the dimension of input features identical, cross-modality attention layer to perform inter-modality interactions, and MIL module for the weakly-supervised training. Among these modules, the cross-modality attention layer is ameliorated from the encoder part of Transformer [57], which includes the multi-head self-attention, feed-forward layer, residual connection [26], and layer normalization [3]. In the raw self-attention block, features are projected by three different parameter matrices as query, key, and value vectors, respectively. Then the scale dot-product attention score is computed by $\text{attr}(q, k, v) = \text{softmax}(\frac{qk^T}{\sqrt{d_m}})v$, where $q, k, \nu$ denotes the query, key, and value vectors, $d_m$ is the dimension of query vectors, $T$ denotes the matrix transpose operation. To enforce cross-modality interactions, we change the key and value vectors of the self-attention block to features in other modalities: $h_a = \text{attr}(f_aW_Q, f_aW_K, f_aW_V)$, $h_v = \text{attr}(f_vW_Q, f_vW_K, f_vW_V)$, where $h_a, h_v$ are updated audio and visual features, $W_Q, W_K, W_V$ are learnable parameters. We adopt the sharing parameter strategy for feature projection to reduce computation.

We adopt the MIL procedure under the weakly-supervised setting to obtain video-level scores. Unlike prior works, we process unimodal features individually to alleviate modality asynchrony. To be specific, fully-connected layers are used in each modality to generate unimodal logits. Then we take the summation of unimodal logits as the fused audio-visual logits while reserving the unimodal logits for the following contrastive learning. Finally, the top-K audio-visual logits are average-pooled and put into a sigmoid activation to generate video-level scores for optimization. The entire procedure is formulated as:

$$L_{v2a} (B^{\text{audio}}_a(i)) = \frac{1}{\sum_{n=1}^{K_{\text{audio}}} e^{\Theta(\sigma(h_a(i) + h_v(i))/\tau)} - \log e^{\phi(B^{\text{audio}}_a(i), h^{\text{audio}}_a(i))/\tau} + \sum_{n=1}^{K_{\text{audio}}} e^{\phi(B^{\text{audio}}_a(i), h^{\text{audio}}_a(n))/\tau}},$$

where $\phi$ denotes cosine similarity function, $\tau$ is the temperature hyperparameter, $K_{\text{audio}}$ denotes the normal instances number in the whole mini-batch. Similarly, the objective for audio violent semi-bag representation $B^{\text{audio}}_a(i)$ against visual background instances embeddings $(h^{\text{bgd}}_v(n))_{n=1}^{K_{\text{bgd}}}$ is formulated as:

$$L_{a2b} (B^{\text{audio}}_a(i)) = \frac{1}{\sum_{n=1}^{K_{\text{bgd}}} e^{\phi(B^{\text{audio}}_a(i), h^{\text{bgd}}_v(i))/\tau)} - \log e^{\phi(B^{\text{audio}}_a(i), h^{\text{audio}}_a(i))/\tau} + \sum_{n=1}^{K_{\text{bgd}}} e^{\phi(B^{\text{audio}}_a(i), h^{\text{bgd}}_v(n)/\tau)},$$

where $K_{\text{bgd}}$ denotes the background instance number in the whole mini-batch. The visual-against-audio counterparts are highly similar, thus we omit these for concise writing.

Moreover, we argue that both normal and violent videos contain background snippets, and learning the difference between event-related segments and background noises could benefit the localization. Therefore, we select the bottom-K instances of the whole mini-batch as the background semi-bag $B^{\text{bgd}}_m = \{h_m(n)\}_{n=1}^{K_{\text{bgd}}}$, $m \in \{a, v\}$. In each mini-batch, the model should contrast violent audio-visual instances against negative pairs constructed by violent instances and other instances (background and normal).

An intuitive way is to randomly pick intra- and inter-semi-bag instances in the opposite modality as positive and negative pairs. However, we argue that audio and visual violent instances with diverse positions could be semantically mismatched, such as expressing the beginning and ending of a violent event, respectively. Therefore, it is unnatural to assume that they share the same implication. In contrast, we conduct average pooling to embeddings of all violence instances in each bag and form a semi-bag-level representation $B^{\text{semi}}_m(m \in \{a, v\})$. By doing so, the audio and visual representation both express event-level semantics, thereby alleviating the noise issue. To this end, we construct semi-bag-level positive pairs, which are assembled by audio and visual violent semi-bag representations $B^{\text{audio}}_a, B^{\text{vgd}}_v$. We also construct semi-bag-to-instance negative pairs to maintain numerous contrastive samples, where violent semi-bag representations are combined with background and normal instance embeddings $h^{\text{bgd}}_v, h^{\text{nor}}_m(m \in \{a, v\})$ in the opposite modality as negative pairs.

We use the InfoNCE [55] as the training objective of this part, which closes the distance between positive pairs and enlarges the distance between negatives. The objective for audio violent semi-bag representation $B^{\text{audio}}_a(i)$ against visual normal instance embeddings $(h^{\text{nor}}_v(n))_{n=1}^{K_{\text{nor}}}$ is formulated as:

$$L_{c2a} (B^{\text{audio}}_a(i)) = \frac{1}{\sum_{n=1}^{K_{\text{nor}}} e^{\phi(B^{\text{audio}}_a(i), h^{\text{nor}}_v(n))/\tau)} - \log e^{\phi(B^{\text{audio}}_a(i), h^{\text{audio}}_a(i))/\tau} + \sum_{n=1}^{K_{\text{nor}}} e^{\phi(B^{\text{audio}}_a(i), h^{\text{nor}}_v(n)/\tau)},$$

where $\phi$ denotes cosine similarity function, $\tau$ is the temperature hyperparameter, $K_{\text{nor}}$ denotes the normal instances number in the whole mini-batch. Similarly, the objective for audio violent semi-bag representation $B^{\text{audio}}_a(i)$ against visual background instances embeddings $(h^{\text{bgd}}_v(n))_{n=1}^{K_{\text{bgd}}}$ is formulated as:

$$L_{a2b} (B^{\text{audio}}_a(i)) = \frac{1}{\sum_{n=1}^{K_{\text{bgd}}} e^{\phi(B^{\text{audio}}_a(i), h^{\text{bgd}}_v(n))/\tau)} - \log e^{\phi(B^{\text{audio}}_a(i), h^{\text{audio}}_a(i))/\tau} + \sum_{n=1}^{K_{\text{bgd}}} e^{\phi(B^{\text{audio}}_a(i), h^{\text{bgd}}_v(n)/\tau)},$$

where $K_{\text{bgd}}$ denotes the background instance number in the whole mini-batch. The visual-against-audio counterparts are highly similar, thus we omit these for concise writing.

### 4.3 Self-Distillation

The audio-visual interactions provided by the former parts could introduce abundant modality noises, and modality asynchrony also results in the semantic mismatch of multimodal and unimodal features in the same temporal position. To address these issues, we argue that training a similar visual network simultaneously enables the model to ensemble unimodal and multimodal knowledge. With a
controllable co-distillation strategy, our proposed module warrants modality noise reduction and robust modality-agnostic knowledge.

Specifically, we propose an analogous unimodal network that contains comparable architecture with our two-stream network. The cross-modality attention block is substituted by the standard transformer encoder block including self-attention. During training, the unimodal network is trained with a relatively small learning rate, and parameters of the same layers are infused into the audio-visual network with an exponential moving average strategy:

\[
\theta_{uv} \leftarrow m\theta_{uv} + (1 - m)\theta_u
\]

where \(\theta_{uv}\) and \(\theta_u\) denotes parameters of the audio-visual model and visual model, respectively, \(m\) denotes the control hyperparameter following a cosine scheduler that increases from the original value \(m\) to 1 during training.

4.4 Learning Objective

The entire framework is optimized in a joint-training manner. For the video-level prediction \(p\), we leverage binary cross-entropy \(L_B\) as the training objective and use a linearly growing strategy to control the weight of contrastive loss. The total objective is:

\[
L_{au} = \frac{\lambda_{ovi}(t)}{K_{ovi}} \sum_i (L_{ct}^{ovi}(B_u^{vio}(i)) + L_{ct}^{ovi}(B_u^{vio}(i))) +
\frac{\lambda_{ovi}(t)}{K_{ovi}} \sum_i (L_{ct}^{ovi}(B_u^{vio}(i)) + (L_{ct}^{ovi}(B_u^{vio}(i)) + L_B)
\]

\[
\lambda(t) = \min(r \times t, \Lambda)
\]

where \(K_{ovi}\) denotes the number of channels of video streams in the whole mini-batch, \(\lambda(t)\) is a controller to increase weight within a few epochs linearly, \(r\) denotes the growing ratio, \(t\) is the current epoch, and \(\Lambda\) denotes the maximum weight.

The visual network is optimized via the BCE loss with video-level labels to distill unimodal knowledge. The two objectives are optimized simultaneously during training while in the inference phase, only the audio-visual network is used for prediction.

5 EXPERIMENT

We design experiments to verify our model from two perspectives, the end-to-end framework compared with state-of-the-art methods and assembling with other networks as plug-in modules. Experimental details and analyses are introduced as follows.

5.1 Dataset and Evaluation Metric

XD-Violence [62] dataset is by far the only available large-scale audio-visual dataset for violence detection, which is also the largest dataset compared with other unimodal datasets. XD-Violence consists of 4,757 untrimmed videos (217 hours) and six types of violent events, which are curated from real-life movies and in-the-wild scenes on YouTube. Although previous methods adopt some popular datasets [36, 51] as benchmarks, we argue that these datasets only contain unimodal visual contents, which cannot perform cross-modality interactions and further verify our proposed multimodal framework. Hence, following [43, 62], we select the large-scale audio-visual dataset XD-Violence as benchmark. During inference, we utilize the frame-level average precision (AP) as evaluation metrics following previous works [43, 54, 62].

5.2 Implementation Details

To make a fair comparison, we adopt the same feature extracting procedure as prior methods [43, 54, 61, 62]. Concretely, we use the I3D [7] network pretrained on the Kinetics-400 dataset to extract visual features. Audio features are extracted via the VGGish [22, 27] network pretrained on a large YouTube dataset. The visual sample rate is set to be 24 fps, and visual features are extracted by a sliding window with a size of 16 frames. For the auditory data, we first divide each audio into 960-ms overlapped segments and compute the log-mel spectrogram with 96 × 64 bins.

The entire network is trained on an NVIDIA Tesla V100 GPU for 50 epochs. We set the batch size as 128 and the initial learning rate as 4e-4, which is dynamically adjusted by a cosine annealing scheduler. For the visual distillation network, the learning rate is set as 5e-5. We use Adam [31] as the optimizer without weight decay. During optimization, the weighted hyperparameter \(r\), \(\Lambda_{ovi}, \Lambda_{ovi}\) are 0.1, 1.5, and 1.5, respectively. The initial distillation weight \(m\) is set to 0.91. The temperature \(\tau\) of InfoNCE [55] is set to be 0.1. The hidden dimension of our two-stream network is 128, and the dropout rate is 0.1. For the MIL, we set the value \(K\) of K-max activation as \(\frac{T + 1}{T}\), where \(T\) denotes the length of input feature.

5.3 Comparisons with State-of-the-Arts

We compare our proposed approach with state-of-the-art models, including (1) unsupervised methods: SVM baseline, OSCVM [48], and Hasan et al. [24]; (2) unimodal weakly-supervised methods: Sultani et al. [51], RTFM [54], Li et al. [33], and Wu et al. [61]; (3) audio-visual weakly-supervised methods: Wu et al. [62] and Pang et al. [43]. We report the AP results on XD-Violence dataset in Table 1.

## Table 1: Comparison of the frame-level AP performance with unsupervised and weakly-supervised baselines. † denotes results re-implemented by integrating logits of two identical networks with audio and visual inputs, and ‡ indicates re-implemented by fusing audio and visual features as inputs.

| Manner | Method | Modality | AP (%) | Param. |
|--------|--------|----------|--------|--------|
| Unsup. | SVM baseline | V | 50.78 | / |
| | OSCVM [48] | V | 27.25 | / |
| | Hasan et al. [24] | V | 30.77 | / |
| Sultani et al. [51] | V | 73.20 | / |
| Wu et al. [61] | V | 75.90 | / |
| RTFM [54] | V | 77.81 | 12.067M |
| RTFM † [54] | A-V | 78.10 | 13.510M |
| RTFM‡ [54] | A-V | 78.54 | 13.190M |
| Li et al. [33] | V | 78.28 | / |
| Wu et al. † [62] | A-V | 78.66 | 1.539M |
| Pang et al. [43] | A-V | 81.69 | 1.876M |
| Ours (light) | A-V | 82.17 | 0.347M |
| Ours (full) | A-V | 83.40 | 0.678M |
which suggests that modality-aware interactions are indispensable. This approach for-

bids the intermediate modality interaction in the network, aiming to show the performance of simply integrating multimodal data. Considering some networks may be unsuitable for multimodal inputs, we put forward another strategy to train two unimodal networks simultaneously and generate audio and visual logits, respectively. The audio-visual predictions are generated by fusing logits of two identical networks with audio and visual inputs, respectively. The model surpasses prior results with a minimum of 5.12%, showing the necessity of utilizing multimodal cues for violent detection.

To further demonstrate the efficacy of our modality-aware contrastive instance learning and cross-modality distillation, we select state-of-the-art methods [43, 62] as audio-visual baselines and re-implement SOTA unimodal MIL method [54] with two modality-expansion strategies. First, following [62], we fuse the audio and visual features in an early way as model inputs. This approach forbids the intermediate modality interaction in the network, aiming to show the performance of simply integrating multimodal data. Considering some networks may be unsuitable for multimodal inputs, we put forward another strategy to train two unimodal networks simultaneously and generate audio and visual logits, respectively. The audio-visual predictions are generated by fusing unimodal logits. Results show that our framework achieves 1.71% higher performance against state-of-the-art method Pang et al. [43], which verifies that our modules are practical for violence detection. Our method outperforms RTFM* and Wu et al. [62] by 1.43% and 2.30%, respectively. Pang et al. [46] can only be amalgamated with MA-CIL in the two-stream network manner (†), while SD should be assembled in an early modality fusion way (*), we can only combine these frameworks with our modules separately. For the multimodal approach [43], we both testify the joint and independent enhancement performances of our MA-CIL and SD modules.

We report the results on the XD-Violence dataset in Table 2. First, we observe that MA-CIL boosts the unimodal baselines Wu et al. [62] and RTFM [54] for 1.32% and 1.46%, respectively, showing that our contrastive learning method improves the discrimination of models. We also note that equipped with the SD module, the performances of [54, 62] also gain an increase of 1.43% and 2.30%. For the multimodal baseline [43], we remove the mutual loss and multimodal fusion modules and leverage the vanilla attention-based variant (‡) for comparison. Results show that enhanced with MA-CIL and SD separately or jointly both achieve accuracy boosts. In summary, we conclude that integrating our MA-CIL and SD modules is beneficial to numerous networks and our modules can be utilized flexibly depending on specific usages.

### 5.5 Complexity Analysis

As we mentioned before, we propose a computation-friendly framework that does not introduce too many parameters. To support our claims, we compare parameter amounts with previous methods, which are shown in the Param. column of Table 1, 2. In Table 1, we report the parameter amounts of previous works we re-implement and our proposed framework, where Ours (light) denotes the ablated model without self-distillation, and Ours (full) indicates the full model with MA-CIL and SD. In Table 2, we provide parameter amounts of the raw methods and our enhancement variants.

From the comparison with other methods, we observe that Ours (light) holds the smallest model size (0.347M) while outperforming all previous methods. Combined with the SD module, our full model still has fewer parameter amounts and achieves the best performance. This result demonstrates the efficiency of our framework, which leverages a much simpler network yet gains better performance. As shown in Table 2, we note that the MA-CIL method does not include any parameters, which exploits the intrinsic prior of multimodal instances and obtains model discrimination with

### Table 2: Results on proposed MA-CIL and SD modules as plug-in modules.

| Method          | MA-CIL | SD | AP (%) | Param. |
|-----------------|--------|----|--------|--------|
| Wu et al. [62]  | X      | X  | 78.64  | 0.843M |
| Wu et al.† [62]|       | X  | 78.66  | 1.539M |
| Wu et al. [62]  | X      | ✓  | 80.07 (1.437) | 1.612M |
| Wu et al.† [62]| ✓      | ✓  | 79.98 (1.327) | 1.539M |
| RTFM [54]       | X      | X  | 77.81  | 12.067M |
| RTFM* [54]      | X      | X  | 78.10  | 13.510M |
| RTFM† [54]      | X      | X  | 78.54  | 13.190M |
| RTFM* [54]      | ✓      | ✓  | 80.40 (2.307) | 25.577M |
| RTFM† [54]      | ✓      | ✓  | 80.00 (1.467) | 13.190M |
| Pang et al. [46]| X      | X  | 81.69  | 1.876M |
| Pang et al.‡ [46]| X     | X  | 80.03  | 1.086M |
| Pang et al. [46]| X      | ✓  | 81.21 (1.187) | 2.138M |
| Pang et al. [46]| ✓      | X  | 80.90 (0.877) | 1.086M |
| Pang et al. [46]| ✓      | ✓  | 82.21 (2.187) | 1.613M |

### Table 3: Ablation studies on different components of our proposed framework.

| Index | Two-Stream | MA-CIL | SD | AP (%) |
|-------|------------|--------|----|--------|
| 1     | ✓          | X      | X  | 71.37  |
| 2     | ✓          | X      | ✓  | 74.01  |
| 3     | ✓          | ✓      | ✓  | 82.17  |
| 4     | ✓          | ✓      | ✓  | 83.40  |
5.6 Ablation Studies

To further investigate the contribution of our proposed modules, we conduct ablation experiments to demonstrate how each aspect of our framework affects the overall performance.

We first conduct experiments on the effectiveness of each component, and the results are shown in Table 3. The vanilla two-stream network without MA-CIL and SD achieves a performance of 71.37%. We argue that the limited performance is driven by the small-scale analogy model structure. This suggests that our proposed modules are flexible to be adapted to multimodal networks.

We then investigate the role of our SD module. Combining our SD module to the raw two-stream network and network with MA-CIL, the ablated models achieve the AP increase of 2.64% and 1.23%, respectively. This indicates that the SD module is effective both with and without no computation cost. When boosting the multimodal model [43], the enhanced model has comparable size to the raw model due to the analogous model structure. This suggests that our proposed modules are flexible to be adapted to multimodal networks.

Table 4: Ablation study for the hyperparameters in the proposed modality-aware contrastive instance learning.

| Index | $\Lambda_{o2a}$ | $\Lambda_{o2b}$ | ratio ($\tau$) | AP (%) |
|-------|-----------------|-----------------|----------------|--------|
| 1     | 1.0             | 1.0             | 0.1            | 82.62  |
| 2     | 1.0             | 1.0             | 0.3            | 82.67  |
| 3     | 3.0             | 3.0             |                | 82.09  |
| 4     | 1.5             | 1.0             | 0.1            | 82.95  |
| 5     | 1.5             | 1.0             | 0.3            | 81.37  |
| 6     | 3.0             | 3.0             |                | 81.25  |
| 7     | 1.0             | 1.5             | 0.1            | 83.21  |
| 8     | 1.0             | 1.5             | 0.3            | 82.62  |
| 9     | 3.0             | 3.0             |                | 81.68  |
| 10    |                 |                 | 0.1            | 83.40  |
| 11    | 1.5             | 1.5             | 0.3            | 81.61  |
| 12    | 3.0             | 3.0             |                | 82.14  |

5.7 Qualitative Analysis

We first visualize the variation of the training loss and video-level prediction accuracy. The ranges of BCE loss and contrastive loss are shown in blue and green curves, respectively.

Figure 4: Illustration of the accuracy and loss curves in 50 epochs during training. The red curve denotes the video-level prediction accuracy. The ranges of BCE loss and contrastive loss are shown in blue and green curves, respectively.

Figure 5: Feature space visualizations of the vanilla features and the output of our model on XD-Violence testing videos. Contrastive learning, and the two modules complement each other for a better violence detection performance.

Finally, we investigate the control hyperparameter $m$ of the self-distillation block in our proposed method as shown in Figure 3. Results show that the best performance achieves at $m = 0.91$.

Figure 3: Ablation studies of different settings for control hyperparameter $m$ in our self-distillation module.

Figure 5: Feature space visualizations of the vanilla features and the output of our model on XD-Violence testing videos.
where the red curve denotes the video-level accuracy, and the blue and green curves denote BCE loss and contrastive loss, respectively. For the prediction accuracy, we observe a sudden decrease in the first 10 training epochs, where the contrastive learning constraints are gradually applied with the increasing weights. After learning the discrimination for a few epochs, the training accuracy begins to increase and finally outperforms previous results. A similar conclusion also appears in the loss curves. The reduction of the BCE loss comes from the first few epochs, where the model is required to generate high-quality embeddings. The contrastive loss has a lasting decline in dozens of epochs, which means the constraints enforce the model to differentiate instances for a long training period. These curves also denote that the two objectives are co-optimized without interfering with each other. We argue that contrastive learning plays a complementary role to traditional MIL learning, and this insight further demonstrates the generalizability of our methods.

We also provide t-SNE [56] visualizations about the distributions of audio and visual features on the XD-Violence test set. Results are shown in Figure 5, where yellow dots denote background segments and purple dots are violent features. We can find that the violent and non-violent features are clearly clustered, and the distance between uncorrelated features is enlarged after the training procedure. This reveals that aided by our proposed network, instances are successfully differentiated in both audio and visual modalities, further indicating the effectiveness of our proposed framework.

Finally, we provide visualizations of prediction results presented in Figure 6. Our model accurately localizes the anomalous events and even identifies normal events of a very short duration between two violent events. In non-violent videos, the magnitudes between normal and background segments are also evident. Scores of normal events will be a little higher than the background segments yet far less than the violent segments, and our method generates nearly zero predictions for the background snippets. These results show that our proposed approach enables the model to perceive the discrepancy between segments in different types (violent, normal, and background), and further contribute to the violent detection.

6 CONCLUSION

In this paper, we investigate the model asynchrony and undifferentiated instances phenomena of MIL under audio-visual scenarios, and further show the impact on weakly-supervised audio-visual learning. Then a modality-aware contrastive instance learning with a self-distillation framework is proposed to address these issues. To be specific, we design a lightweight two-stream network to generate audio and visual embedding and logits. Furthermore, a cross-modality contrast is applied to audio and visual instances of different semantics, which involves more unused instances for better discrimination and alleviates the modality inconsistency. To diminish training noises, a self-distillation module is leveraged to transfer visual knowledge to the audio-visual network, by which the semantic gaps between unimodal and multimodal features are narrowed. Our framework outperforms previous methods on the XD-Violence dataset with minor expenses. Besides, assembled with our contrastive learning and self-distillation modules, several prior methods achieve higher detection accuracy, showing the capability as plug-in modules to ameliorate other networks.

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 the SPMI Innovation and Technology Fund Projects (SAST2020-110).
vision and pattern recognition. 6479–6488.

[52] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive representation distillation. arXiv preprint arXiv:1910.10699.

[53] Yapeng Tian, Dingzeyu Li, and Chendiang Xu. 2020. Unified multisensory perception. Weakly-supervised audio-visual video parsing. In European Conference on Computer Vision. Springer, 436–454.

[54] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 4975–4986.

[55] Aaron Van den Oord, Yazhe Li, Oriol Vinyals, et al. 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 2, 3 (2018), 4.

[56] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. Journal of machine learning research 9, 11 (2008).

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems 30.

[58] Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. 2022. On Distinctive Image Captioning via Comparing and Reweighting. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

[59] Keyu Wen, Jin Xia, Yuanyuan Huang, Linyang Li, Jiayan Xu, and Jie Shao. 2021. COOKIE: Contrastive Cross-Modal Knowledge Sharing Pre-training for Vision-Language Representation. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2208–2217.

[60] Spencer Whitehead, Hui Wu, Heng Ji, Rogerio Feris, and Kate Saenko. 2021. Separating skills and concepts for novel visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 5632–5641.

[61] Peng Wu and Jing Liu. 2021. Learning causal temporal relation and feature discrimination for anomaly detection. IEEE Transactions on Image Processing 30, 3513–3527.

[62] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In European Conference on Computer Vision. Springer, 322–339.

[63] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3733–3742.

[64] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 10687–10698.

[65] Zihui Xue, Sucheng Ren, Zhengqg Gao, and Hang Zhao. 2021. Multimodal knowledge expansion. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 854–863.

[66] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Lipun Chen, Belinda Zeng, Trushul Chilum, and Junzhou Huang. 2022. Vision-Language Pre-Training with Triple Contrastive Learning. arXiv preprint arXiv:2202.10401 (2022).

[67] Jiangong Zhang, Laiyun Qing, and Jun Miao. 2019. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 4030–4034.

[68] Tao Zhang, Zhijie Yang, Wenjing Jia, Baoqing Yang, Jie Yang, and Xiangjian He. 2016. A new method for violence detection in surveillance scenes. Multimedia Tools and Applications 75, 12 (2016), 7327–7349.