Cross-Modal Fusion and Attention Mechanism for Weakly Supervised Video Anomaly Detection

Ayush Ghadiya*, Purbayan Kar*, Vishal Chudasama*, Pankaj Wasnik[†]

Media Analysis Group, Sony Research India, Bangalore, India

{ayush.ghadiya, purbayan.kar, vishal.chudasama1, pankaj.wasnik}@sony.com

Abstract

Recently, weakly supervised video anomaly detection (WS-VAD) has emerged as a contemporary research direction to identify anomaly events like violence and nudity in videos using only video-level labels. However, this task has substantial challenges, including addressing imbalanced modality information and consistently distinguishing between normal and abnormal features. In this paper, we address these challenges and propose a multi-modal WS-VAD framework to accurately detect anomalies such as violence and nudity. Within the proposed framework, we introduce a new fusion mechanism known as the Cross-modal Fusion Adapter (CFA), which dynamically selects and enhances highly relevant audio-visual features in relation to the visual modality. Additionally, we introduce a Hyperbolic Lorentzian Graph Attention (HLGAtt) to effectively capture the hierarchical relationships between normal and abnormal representations, thereby enhancing feature separation accuracy. Through extensive experiments, we demonstrate that the proposed model achieves state-of-the-art results on benchmark datasets of violence and nudity detection.

1. Introduction

In the modern technology era, kids are increasingly turning to online platforms for learning, fun, and connecting with others. However, this easy access also brings up worries about their exposure to harmful and unsuitable content, particularly content with violence and nudity. The potential adverse effects on a child's emotional well-being and psychological development underscores the importance of implementing robust mechanisms to detect violence and nudity. Detecting such anomalies in a video is a well-known computer vision problem that can also be useful in other real-world applications such as surveillance systems, crime prevention, and content moderation. Acquiring annotations for anomalies at the frame level in videos is costly and time-consuming. As a result, WS-VAD has emerged as a prominent area of research. WS-VAD focuses on learning abnormal events, such as violence and nudity, solely based on video-level binary labels. In this approach, a video is classified as normal if no anomalous event is detected. In contrast, it is classified as an anomaly if any form of abnormal events, such as violence or nudity, is present. WS-VAD methods usually employ Multiple Instance Learning (MIL) [17] for model training. Here, a regular video is seen as a negative bag with no anomalous segments, while an anomaly video is viewed as positive bag with one or more anomalous segments. The anomaly evaluation function is trained by optimizing the MIL loss to ensure positive bag has a higher anomaly value than negative (normal) bag.

Following MIL, recently, several WS-VAD methods have been proposed based on single-modality (i.e., video-based methods [10, 13, 28–30, 34, 35]) and multi-modality [1, 19, 21, 36–38, 40]. The multi-modal approaches have shown promising results compared to single-modality-based methods, which jointly learn audio and visual representations to improve performance by leveraging complementary information from different modalities. Although multi-modal methods show promising performance, they face two main challenges: 1) unbalanced modality information when combining audio-visual features and 2) inconsistent discrimination between normal and abnormal features. Recently, Peng *et al.* [21] found that the issue of modality imbalance is mainly due to noise in audio signals from real-world scenarios. To address this, they suggest that auditory information contributes less to anomaly detection than visual cues, leading to lower prioritization of audio features. However, this approach must be corrected when audio data is as crucial as visual data. To address another issue, i.e., inconsistent discrimination between normal and abnormal features, prior studies have utilized graph representation learning, where each instance is treated as a node in a graph. However, these methods still struggle to distinguish them accurately.

In this study, we propose a new framework to address

---

*Equal Contributions
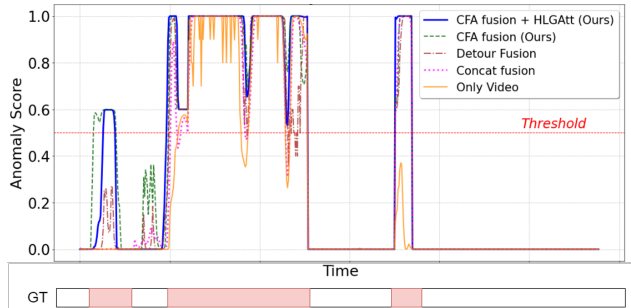[†]Corresponding Author.

Figure 1. Comparative analysis of our proposed method with prior video-based method as well as audio-video based fusion approaches [21, 36] on testing videos of XD-Violence dataset.

these challenges. We introduce a novel fusion module called a CFA to address the challenge of imbalanced modality information. It dynamically adjusts the influence of each modality by prioritizing the importance of audio features relative to the visual modality. This selective process ensures that only relevant audio features crucial for visual learning are being utilized. By adapting to select the most appropriate features relative to the visual modality, our approach enhances visual feature learning by incorporating relevant audio features. Furthermore, we introduce a hyperbolic graph convolution network-based HLGAtt mechanism to maintain consistent discrimination between normal and abnormal features. This mechanism operates in hyperbolic space to capture hierarchical relationships between normal and abnormal representations through spatial and temporal feature learning, which aids in distinguishing normal and abnormal features.

The proposed model accurately identifies anomaly events and outperforms existing state-of-the-art (SOTA) methods for violence and nudity detection tasks. Figure 1 shows the anomaly score analysis obtained from a few violent and normal instances of the XD-Violence dataset and compares it with various approaches such as only video-based method [36], Concate fusion [36], Detour fusion [21] approaches. Figure 1 shows that the proposed model accurately identifies anomalies compared to others. We summarize the contributions of this paper as follows:

- We propose a new WS-VAD framework to address the imbalance issue in audio-visual modality information and effectively distinguish abnormal features from normal ones so that anomaly events such as violence and nudity can be detected accurately.
- To address the imbalanced modality information issue, we introduce a novel fusion module called CFA, which helps the proposed framework to facilitate multi-modal interaction effectively by dynamically regulating the contribution of each modality.
- We introduce a novel attention mechanism called HLGAtt to capture the hierarchical relationships between normal

and abnormal representations, thereby enhancing the feature separation.

2. Related Works

2.1. Violence Detection Works

Earlier, few unsupervised learning-based methods [14, 25] have been proposed for violence detection. These methods focus on one-class classification via learning what is normal and spotting anomalies by recognizing deviations from the norm. However, these methods are not well-suited for complex environments and often struggle due to the limited availability of abnormal video data during training.

Recently, WS-VAD methods [13, 28, 36, 37, 40] have been introduced utilizing video-level labels and achieved promising results over unsupervised VAD methods. A few video-based WS-VAD approaches [10, 13, 28–30, 34, 35] have been proposed to enhance the detection accuracy of violence events. However, these approaches overlooked audio information and cross-modality interactions, limiting the effectiveness of violence prediction. To address this issue, Wu *et al.* [36] introduced a large-scale audio-visual dataset named XD-Violence and established a baseline for audio-visual activities. Following this, many multi-modal approaches [1, 19, 21, 36–38, 40] have been proposed that outperforms video-based WS-VAD methods. Recently, Peng *et al.* [21] proposed a fusion mechanism for audio-visual data and introduced a hyperbolic graph convolution network-based model to efficiently capture the semantic distinctions via learning the embeddings in hyperbolic space. Recently, Zhou *et al.* [40] proposed a dual memory units module with uncertainty regulation emphasizing learning representations of abnormal and normal data. Salem *et al.* [1] introduced a new version of MIL that avoids the disadvantages of ranking loss by using margin loss instead.

Although these methods present promising results, their effectiveness is hindered by the integration of imbalanced audio-visual features. Moreover, they struggle to consistently differentiate between normal and abnormal features, limiting the detection accuracy. This paper addresses these issues and proposes a new multi-modal framework that detects violent events more accurately. In contrast to recent multi-modal approaches [21, 38, 40], we propose a new cross-modal fusion with modulation mechanism to learn and fuse audio modality with relative visual features adaptively. Furthermore, we introduce Lorentzian attention-based hyperbolic graph mechanism to learn hierarchical relationships between normal and abnormal features and discriminate them effectively.

2.2. Nudity Detection Works

In video-based nudity detection, researchers have devised various methods to tackle the task of identifying explicit

content. A common strategy involves detecting skin color in video frames [5, 9, 12, 22]. Samal *et al.* [26] proposed a model that combines attention-enabled pooling with a Swin transformer-based YOLOv3 architecture for obscenity detection in images and videos. Jin *et al.* [8] employed a weakly supervised multiple instance learning approach for generating a bag of properly sized regions with minimal annotations to tackle the detection of private body parts based on local regions. Wang *et al.* [15] incorporated an attention-gated mechanism with a deep network, demonstrating its efficacy in performance enhancement. Several studies have proposed deep learning architectures considering local and global context jointly [18, 33]. Utsav *et al.* [27] proposed a domain adaptation-based method to filter adult content in streaming video. Tran *et al.* [31] proposed an additional training-based approach on pseudo labels using Mask R-CNN for sexual object detection.

However, above methods focus on image-based approaches or utilize uni-modal approaches; the audio-visual-based approaches have not been extensively explored. This paper seeks to address this gap by employing audio-visual data, aiming to enhance the accuracy of nudity detection in videos.

3. Methodology

3.1. Problem Statement

Given a set of N videos, $X = \{X_i\}_{i=1}^N$ and the corresponding ground-truth video-level labels $Y = \{Y_i\}_{i=1}^N \in \{1, 0\}$ where $Y_i = 1$ denotes the presence of any abnormal event in the video while $Y_i = 0$ signifies the absence of any abnormal event, we aim to accurately detect abnormal events such as violence and nudity within the videos in a weakly supervised manner. Specifically, each video X_i is initially divided into 16-frame based T non-overlapping multimodal segments ($M = \{M_i^V, M_i^A\}_{i=1}^T$), which are processed by a pre-trained CNN network to extract the corresponding visual features $F_V \in \mathbb{R}^{T \times D_V}$ and audio features $F_A \in \mathbb{R}^{T \times D_A}$, where D_V and D_A represents the feature dimensions of video and audio modality. Here, M_i^V and M_i^A denote the video and audio segments, respectively. These extracted visual and audio features are then forwarded into the proposed framework which identifies whether the input video contains any abnormal events or not.

To identify abnormal events accurately, we propose a new framework as shown in Figure 2 in which we introduce novel cross-modal fusion module and hyperbolic Lorentzian graph attention mechanism. Details of these modules are discussed in subsequent subsections.

3.2. Cross-modal Fusion Adapter (CFA)

The CFA module consists of a prefix-tuned-based bottleneck attention and a modulation mechanism. The prefix-

tuned bottleneck attention helps in efficient multi-modal interaction between audio and visual modalities. The modulation mechanism dynamically regulates the contribution of each modality during the fusion process, taking into account the importance of the audio features to the visual modality.

Prefix-Tuning bottleneck attention mechanism: This mechanism incorporates prior knowledge into the feature transformation process by combining the learned representations with initialized parameters through the prefix-tuning operation. To do this, the process involves concatenating the keys K and values V obtained from audio features F_A with prefixes P_k & P_v, resulting in prefix-tuned keys K_p and values V_p, respectively. The parameters P_k & P_v are initialized as zero matrices with dimensions of $\mathbb{R}^{B \times D_A \times D_p}$, where B, D_A & D_p represent the batch size, audio feature dimension, and prefix dimension, respectively.

These prefix-tuned keys K_p and values V_p along with the query Q, i.e., visual features F_V, are then passed on to the cross-modal multi-head attention module [23]. This module enables the interaction between the prefix-tuned features of the audio and visual modalities, allowing them to selectively and contextually focus on each modality's relevant information. In this process, the attention scores are computed based on queries, prefixed tuned keys and values. The mathematical formulation of the cross-modal multi-head attention module function (i.e., f_{CMA}) can be formulated as

$$F_{Att} = f_{CMA}(Q, K_p, V_p) = Softmax\left(\frac{Q \cdot K_p^T}{\sqrt{D_{K_P}}}\right) \times V_p,$$

(1)

where, D_{K_P} represents the dimensionality of the key vectors (K_p). The attention features F_{Att} are subsequently passed to the bottleneck adapter module. In this stage, the bottleneck adapter ensures smooth interaction between modalities while preserving modality-specific characteristics. It comprises down-scaled fully connected layers (i.e., f_{down}) followed by Gaussian Error Linear Unit (GELU) activation (i.e., f_{GELU}) and up-scaled fully connected layers (i.e., f_{up}). This can be expressed mathematically as

$$\hat{F}_{Att} = f_{up}(f_{GELU}(f_{down}(F_{Att}))), \qquad (2)$$

Here, the GELU activation function introduces non-linearity, allowing intricate feature transformations. This careful design ensures that the adapter module effectively adjusts input features to the shared bottleneck representation, promoting context-aware fusion.

Modulation Mechanism: In the proposed CFA module, we introduce modulation factors that dynamically adjust the impact of individual modalities by considering the importance of their audio features relative to the visual modality. This mechanism is facilitated by a learnable modulation function that operates on audio features F_A to select
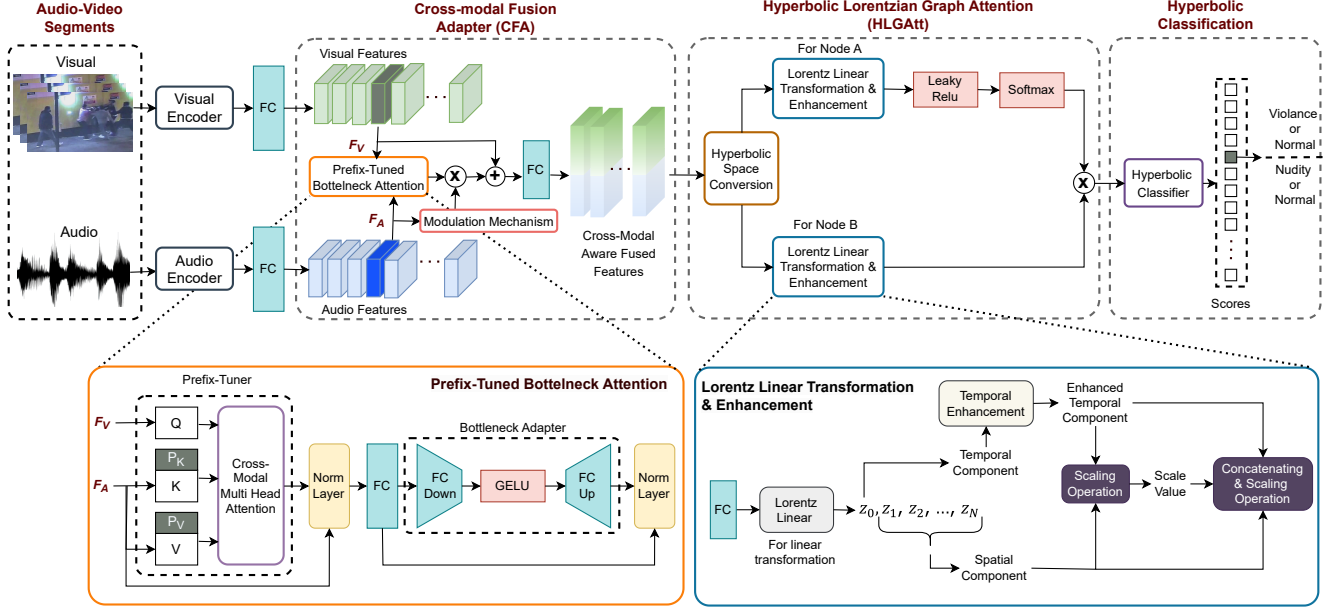
Figure 2. **Overview of the proposed framework.** It takes audio and visual features extracted from pre-trained encoder networks as input, which are further fused through the proposed Cross-Modal Fusion Adapter (CFA) module to learn multi-modal interaction effectively, followed by the introduced Hyperbolic Lorentzian Graph Attention (HLGAtt) mechanism to capture hierarchical relationships between visual and audio representations, ensuring consistency in distinguishing normal and abnormal features during training. Finally, the outcome features are passed in a hyperbolic classifier to predict anomaly events for each instance.

relevant audio features that are important to visual modality. The resulting modulated features F_{Mod} are defined as

$$F_{Mod} = f_{MF}(F_A) = \sigma(W_{mod} \cdot F_A). \qquad (3)$$

Here, σ represents the sigmoid activation, while W_{mod} stands for the weights associated with the modulation function. The sigmoid activation function ensures that modulation factors range between 0 and 1, thereby regulating the degree of modulation applied to the fused representation.

Next, the fusion and refinement process is used, where it first fuses the modulated features with the output of the prefix-tuning bottleneck attention and then refines the fused representation through a fully connected layer. This operation can be expressed mathematically as

$$F_{Fused} = f_{FC}(F_V + (\hat{F}_{Att} \times F_{Mod})). \qquad (4)$$

The modulation mechanism $f_{MF}()$ modifies the output of the prefix-tuning bottleneck attention based on the significance of their audio features to the visual modality. Through the fusion and refinement process, the final fused representation is carefully crafted to capture the most relevant information from both modalities, simultaneously reducing noise and preserving the modality-specific characteristics.

3.3. Hyperbolic Lorentzian Graph Attention (HLGAtt) Mechanism

In the proposed framework, we introduce a hyperbolic graph convolution network based on a new attention mechanism called HLGAtt. The proposed HLGAtt uses a hyperbolic Lorentz graph attention mechanism that learns layer-wise curvature parameters to capture the hierarchical structure of the input graph, thereby enhancing the hierarchical relationship between normal and abnormal representations compared to existing graph-based [21, 36] or transformer-based [40] approaches. It consists of a hyperbolic space conversion operation, a Lorentz linear transformation & enhancement module process on parallel nodes, and a fusing operation.

Initially, we convert the fused audio-visual features F_{Fused} into the hyperbolic space using an exponential function. As a result, we obtain the converted fused features maps $F_H \in \mathbb{R}^{T \times 2D_H}$, wherein T denotes the number of segments and D_H represents the hyperbolic dimension.

Recently, Zhang *et al.* [39] proposed a hyperbolic graph attention mechanism that utilized a parallel branch process to learn different features and patterns in respective branches for prediction tasks. Inspired by this [39], we process the converted hyperbolic feature maps on two parallel branches, i.e., node A and node B, to learn specific patterns from the input feature maps. Separating the branches en-

sures that features with similar characteristics are directed to their respective nodes. This allows each branch to learn the unique properties of normal and abnormal features, enabling more precise discrimination between them.

The converted hyperbolic feature maps are passed through the Lorentzian linear transformation & enhancement module in each node. Here, we employ the Lorentzian linear transformation [4, 21] for feature transformation and its transformed temporal and spatial features are further enhanced using the proposed enhancement mechanism. In Lorentzian linear transformation, we first establish the adjacency matrix $A \in \mathbb{R}^{T \times T}$ to capture hyperbolic feature similarities. Here, each entry A_{ij} can be calculated as

$$
\begin{aligned}
A_{ij} &= f_{sim}(F_{H,i}, F_{H,j}) \\
&= Softmax(\exp(-d_L(F_{H,i}, F_{H,j})),
\end{aligned} \tag{5}
$$

where, f_{sim} represents the hyperbolic feature similarity measure, which evaluates how closely snippets i and j resemble each other based on their Lorentzian intrinsic distance d_L. The exponential and Softmax functions are employed to maintain non-negativity and restrict the values of A within the range of $[0, 1]$.

Next, we incorporate a hyperbolic Lorentz linear (i.e., $f_{HL}()$), followed by neighborhood hyperbolic aggregation operation [24] for feature transformation. These transformed hyperbolic features of the ith snippet at the layer l (i.e., z_i^l) can be expressed as

$$
z_i^l = F_{H,i}^l = \frac{\sum_{j=1}^T A_{ij} f_{HL}\left(F_{H,i}^{l-1}\right)}{\sqrt{-\eta} \left|\left\|\sum_{k=1}^T A_{ik} f_{HL}\left(F_{H,i}^{l-1}\right)\right\|_{\mathcal{L}}\right|}, \tag{6}
$$

where, η indicates the negative curvature constant.

To enhance these transformed features z further, they are processed based on temporal and spatial information. The initial component of the input vector $z[0]$ signifies the temporal aspect within hyperbolic space [4]. This component is processed via a sigmoid activation function followed by exponential scaling and shifting operations. Through this procedure, temporal features (i.e., T_{nodeA} and T_{nodeB}) are computed for both node A and node B as

$$
\begin{aligned}
T_{nodeA} &= \sigma(z_{nodeA}[0]) \times e^{\gamma} + 1.1 \\
T_{nodeB} &= \sigma(z_{nodeB}[0]) \times e^{\gamma} + 1.1
\end{aligned} \tag{7}
$$

where, γ is a trainable parameter. The remaining elements of input vector z can be considered as the spatial features [4] for node A and node B (i.e., S_{nodeA} and S_{nodeB}). Mathematically, they can be formulated as

$$
\begin{aligned}
S_{nodeA} &= [z_{nodeA}[1], z_{nodeA}[2], ..., z_{nodeA}[n]] \\
S_{nodeB} &= [z_{nodeB}[1], z_{nodeB}[2], ..., z_{nodeB}[n]]
\end{aligned} \tag{8}
$$

These features encapsulate the intricate spatial features in hyperbolic space, which are critical for capturing the hierarchical structure and relationships within the graph.

To ensure the alignment of spatial components with the hyperbolic model, a scaling factor, referred to as Υ is computed. This factor takes into account the temporal and spatial complexities of each node. It ensures that the spatial components are appropriately scaled to fit within the hyperbolic space.

$$
\begin{aligned}
\Upsilon_{nodeA} &= \frac{T_{nodeA}^2 - 1}{\sum_{i=1}^n (S_{nodeA}[i])^2 + \epsilon} \\
\Upsilon_{nodeB} &= \frac{T_{nodeB}^2 - 1}{\sum_{i=1}^n (S_{nodeB}[i])^2 + \epsilon}
\end{aligned} \tag{9}
$$

The temporal and scaled spatial components are concatenate, resulting in enhanced feature vectors (i.e., $\hat{F}_H^{nodeA}$ and $\hat{F}_H^{nodeB}$). Mathematically, this process can be expressed as

$$
\begin{aligned}
\hat{F}_H^{nodeA} &= Concat\left[T_{nodeA}, S_{nodeA} \times \sqrt{\Upsilon_{nodeA}}\right] \\
\hat{F}_H^{nodeB} &= Concat\left[T_{nodeB}, S_{nodeB} \times \sqrt{\Upsilon_{nodeB}}\right]
\end{aligned} \tag{10}
$$

The enhanced feature maps in the node A branch passed through Leaky-ReLU activation and softmax normalization operations to introduce non-linearity and ensure standardization across the enhanced feature maps. This ensures that distinct patterns, representing normal and abnormal data, are learned at each node. By doing so, the model is encouraged to learn different sets of features from those processed by other node (i.e., node B). Finally, the enhanced feature maps from node A and node B are processed via matrix multiplication to compute attention, followed by a ReLU activation to generate the output feature maps. This outcome of the proposed HLGAtt module can be formulated as

$$
F_H^{final} = f_{ReLU}(\hat{F}_H^{nodeA} \cdot \hat{F}_H^{nodeB}). \tag{11}
$$

3.4. Hyperbolic Classifier & Learning Objective

Following [21], we also utilize the hyperbolic classifier, which takes the output of the HLGAtt module as input and predicts the confidence scores for normal and abnormal events. The final score $Score$ can be represented as

$$
Score = f_{Hyp-cls}(F_H^{final}) \tag{12}
$$

In order to train the proposed model end-to-end, we employ the MIL-based learning objective adopted in [20, 21, 28, 36], which calculates the mean value of the top $k-$max predictive scores within a video. The high-scoring positive predictions indicate the presence of abnormal events, while the $k-$max negative scores usually represent hard samples. This learning objective function can be formulated as

$$
L_{MIL} = \frac{1}{N} \sum_{i=1}^N -Y_i \cdot \log(\overline{Score}). \tag{13}
$$

Here, $\overline{Score}$ indicates the average of the $k-$max scores in the video, and Y_i represents the binary video-level label.

4. Experiments and Results

4.1. Implementation Details

The proposed model is trained/tested on benchmark XD-Violence dataset [36] for violence detection task, on NPDI pornography dataset [2, 3] for nudity detection task. The details of these datasets are mentioned below:

- **XD-Violence for violence detection:** The XD-Violence dataset [36] is a diverse compilation of 4754 raw videos (equivalent to 217 hours) gathered from real-world sources, including movies, web videos, sports broadcasts, security cameras, and CCTVs. It consists of six types of violent events, such as abuse, auto crashes, and shootings, with corresponding video-level annotations. The testing set comprises 300 normal and 500 violent videos, while the training set includes 2049 normal and 1905 violent videos, all labeled at the video level.

- **NPDI for Nudity Detection :** The NPDI Pornography benchmark dataset [2, 3] comprises around 80 hours of video content extracted from 400 movies. These contents are classified as pornographic or non-pornographic, with an equivalent amount of videos in each category. Within the non-pornographic section, there are 200 videos labeled as either "easy" or "difficult". The "easy" videos were randomly selected, while the "difficult" ones were obtained through textual search queries such as "beach," "wrestling" and "swimming". Although the "difficult" videos may contain body skin, they do not include explicit nudity or pornographic content.

Training / Evaluation Details: The proposed model is trained on datasets mentioned above using the multi-instance learning-based loss function (i.e., Eq. 13) with a batch size of 128. During the training process, we adopt the Adam optimizer with a learning rate of 5×10^{-4} varied using a cosine annealing scheduler and trained for 50 epochs. For fair comparison with existing SOTA methods, the proposed framework also employs a pre-trained I3D model [11] to extract the visual features (F_V), while the VGGish network [7] is utilized to extract the audio features (F_A). In the proposed framework, we use the LeakyReLU activation function with a negative slope of -2. In the Prefix-Tuner of the CFA module, we empirically chose the prefix dimension as 64. The bottleneck adapter has a size of 256 and utilizes the GELU activation function with a dropout rate of 0.1. The constant representing negative curvature (η) is set to -1 during training.

For comparison on violence detection task, we choose unsupervised methods (i.e., SVM baseline, and Hasan *et al.* [6]), video modality-based weakly supervised methods [13, 28, 28–30, 34, 36, 38, 40], and audio-visual modality-

Table 1. Comparison against SOTA methods on XD-Violence Dataset for violence detection. Best result is **bolded** and second best result is underlined.

Method	Publication	Modality	AP (%)
Unsupervised learning based methods			
SVM baseline	NIPS'99	Video	50.78
Hasan *et al.* [6]	CVPR'16	Video	30.77
Weakly supervised learning based methods			
Sultani *et al.* [28]	CVPR'18	Video	75.68
Wu *et al.* [36]	ECCV'20	Audio + Video	78.66
Wu *et al.* [35]	ICIP'21	Video	75.90
Pang *et al.* [19]	ICASSP'21	Audio + Video	81.69
RTFM [30]	ICCV'21	Video	77.81
MSL *et al.* [13]	AAAI'22	Video	78.28
S3R [34]	ECCV'22	Video	80.26
MACIL-SD [37]	ACM'22	Audio + Video	83.40
HyperVD [21]	arXiv'23	Audio + Video	<u>85.67</u>
UR-DMU [40]	AAAI'23	Audio + Video	81.77
Zhang *et al.* [38]	CVPR'23	Audio + Video	81.43
Salem *et al.* [1]	WACVW'24	Audio + Video	71.40
Tan *et al.* [29]	WACVW'24	Video	82.10
REWARD-E2E [10]	WACV'24	video	80.30
Proposed	—	**Audio + Video**	**86.34**

based weakly supervised methods [1, 19, 36–38, 40]). The frame-level average precision (AP) metric is adopted to compare these methods, whereas a higher AP measure means better performance. For the nudity detection task, we compare the proposed method with existing methods [8, 16, 21, 26, 27, 31]. However, these methods have utilized uni-modal approaches in their network. Additionally, we re-train the recent multi-modal SOTA method called HyperVD [21] on the NPDI dataset. For comparison, we use the standard evaluation metrics, i.e., AP, accuracy, precision, and recall, where higher measures of these evaluation metrics indicate superior performance.

All the experiments were implemented using PyTorch and the network was trained on a 40GB NVIDIA A100 GPU with batch size of 128.

4.2. Result Analysis on Violence Detection task

Table 1 compares state-of-the-art methods on the XD-Violence testing dataset in terms of AP metric. Notably, our proposed method outperforms both video modality-based and audio-video modality-based methods. It achieves an AP score of 86.34%, which is 0.67% higher than the previous best-performing method HyperVD [21]. Compared to video-modality-based methods, our proposed approach shows a 4.24%

Figure 3 displays the visual prediction analysis of our method when compared to existing methods, i.e., HyperVD [21] and Wu *et al.* [36]. The comparison is based on the anomaly score obtained from a few videos of the XD-Violence testing dataset [36]. Here, one can observe that the proposed method not only identifies violent event regions
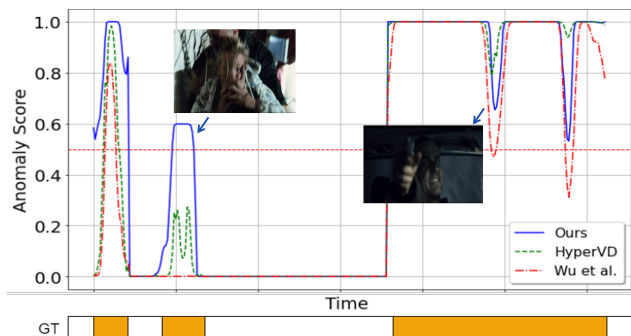
Figure 3. Visual comparison in terms of anomaly score curves on sample video of XD-Violence dataset. Yellow regions are the temporal ground-truths of violent events.
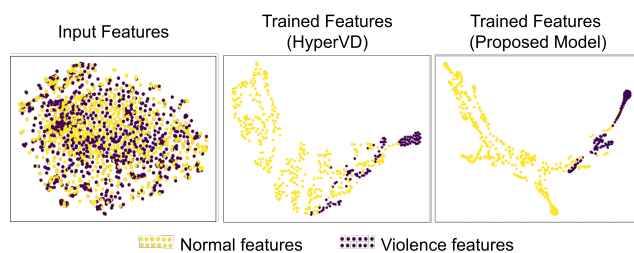

Normal features   Violence features

Figure 4. Visual comparison on normal and violence features of the proposed and HyperVD [21] methods on XD-Violence dataset.

but also yields superior and more precise anomaly scores compared to other methods.

Additionally, Figure 4 provides a comparison between the proposed and HyperVD [21] methods in terms of t-SNE visualization [32] of normal and violent features distributions on the XD-Violence dataset testing videos. One can find that the proposed method effectively clusters the violent and non-violent features and also enlarges the distance between uncorrelated features after the training procedure as compared to the HyperVD [21] method.

4.3. Result Analysis on Nudity Detection task

This section provides an analysis of the Nudity detection task, comparing it with existing methods [8, 16, 21, 26, 27, 31] on the NPDI testing dataset [2, 3]. Table 2 shows the comparison in terms of AP, accuracy, precision, and recall. Here, it can be noticed that the proposed model outperforms other methods in all evaluation metrics by a significant margin. To be specific, our model achieves an AP of 99.45%, an accuracy of 94.12%, a precision of 95%, and a recall of 93.75%. Notably, it demonstrates improvements of at least 1.95% in AP, 0.42% in accuracy, 2.2% in precision, and 3% in recall compared to other methods.

Additionally, Figure 5 demonstrates the visualization of the anomaly score obtained from the proposed and HyperVD [21] methods on the NPDI dataset. The visualization

Table 2. Comparison against SOTA methods on NPDI Dataset for nudity detection. The best result is **bolded** and the second best result is underlined. Here, * indicates the re-trained method.

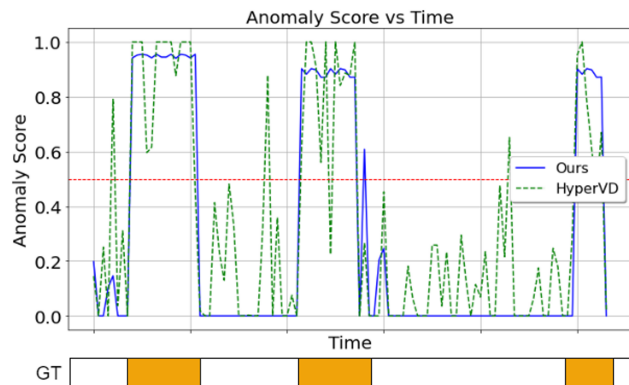| Methods	Modality	AP	Accuracy	Precision	Recall
OpenYahoo [16]	Video	79.0	—	—	—
Deep Region-based CNN [8]	Video	87.8	—	—	—
Deep Part Detector [8]	Video	87.0	—	—	—
Deep MIL [8]	Video	86.0	—	—	—
Weighted MIL [8]	Video	97.5	—	—	—
Tran *et al.* [31]	Video	—	90.43	—	—
WD-based adaptation [27]	Video	96.92	93.70	—	—
ASYv3 [26]	Video	89.87	89.35	89.38	89.55
HyperVD* [21]	Audio + Video	96.45	92.19	92.80	90.75
Proposed	Audio + Video	**99.45**	**94 .12**	**95.00**	**93.75**

Figure 5. Visual comparison between proposed model and HyperVD [21] in terms of Anomaly Score vs Time. Yellow regions are the temporal ground-truths of nudity events.

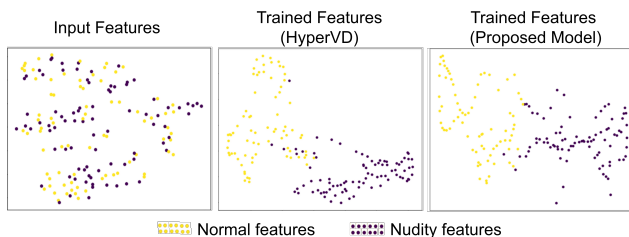Normal features   Nudity features

Figure 6. Visual comparison on normal and nudity event features of the proposed and HyperVD [21] methods on NPDI dataset.

demonstrates that the proposed method generates minimal predictions for regular segments in normal footage while effectively handling extreme situations within nudity content. This analysis also proves that the proposed method not only accurately identifies specific regions but also provides more precise anomaly scores compared to anomaly predictions of the HyperVD method [21].

We also provide t-SNE [32] based visual comparison of the proposed and HyperVD [21] methods in Figure 6. Here, we compare the proposed and re-trained HyperVD [21] methods with corresponding normal and violent feature distributions obtained from the NPDI dataset testing videos. It can be clearly observed that the proposed method

Table 3. Ablation studies on introduced components i.e., CFA & HLGAtt of the proposed framework. Here, PT indicates Prefix-Tuning, MM indicate Modulation Mechanism.

Cases	CFA w/o PT & MM	CFA w/o PT	CFA	FHGCN [21]	HLGAtt	AP (%)
1	✓			✓		81.02
2		✓		✓		81.51
3			✓	✓		83.42
4	✓				✓	85.71
5		✓			✓	86.09
6			✓		✓	**86.34**

performs better in clustering the nudity and normal features as compared to the HyperVD [21] method.

4.4. Ablation Analysis

In order to ensure a fair comparison, all ablation experiments were conducted using the XD-Violence testing dataset for the violence detection task.

Analysis of proposed components: A series of ablation experiments were conducted to validate the efficacy of the proposed components, i.e., CFA and HLGAtt, within the proposed framework. Few experiments have been performed on the CFA module with and without Prefix-Tuning (*PT*) and Modulation Mechanism (*MM*) modules. The results are outlined in Table 3. It is evident that the inclusion of the *PT* and *MM* modules enhances the performance of the CFA module. Furthermore, experiments were carried out using the attention mechanism FHGCN, as proposed in [21], to validate the effectiveness of the HLGAtt module. In Table 3, we present the results performed using the FHGCN module in Cases 1 - 3, while Cases 4 - 6 show the result obtained from the proposed HLGAtt module. The HLGAtt module demonstrates superior performance compared to the FHGCN [21] module by a significant margin. For instance, there is a 4.69% improvement from Case 1 to Case 4, a 4.58% improvement from Case 2 to Case 5, and a 2.92% improvement from Case 3 to Case 6. This analysis proves the efficacy of the proposed modules.

Analysis on prefix dimension in the Prefix-Tuner module: We also thoroughly analyze the prefix dimension within the Prefix-Tuner module featured in the CFA module. The experiment tested different prefix dimensions including 34, 48, 64, 80, 128, and 256. The corresponding results are illustrated in Figure 7 where one can observe that the optimal performance is achieved as AP of 86.34%, with a prefix dimension of 64. However, the performance of AP is diminished when the prefix dimension is increased beyond 64. As a result, we set the prefix dimension as 64 in the proposed CFA module.

Analysis of different fusion mechanism: A series of ablation experiments were conducted to evaluate the efficacy of the proposed CFA fusion mechanism in comparison to alternative fusion methods like Detour fusion [21], Concat fusion, and Gated fusion. The impact of the AP measure obtained from these experiments during the training process
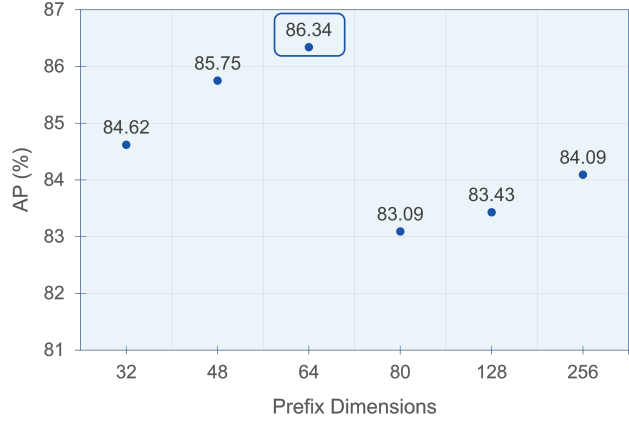

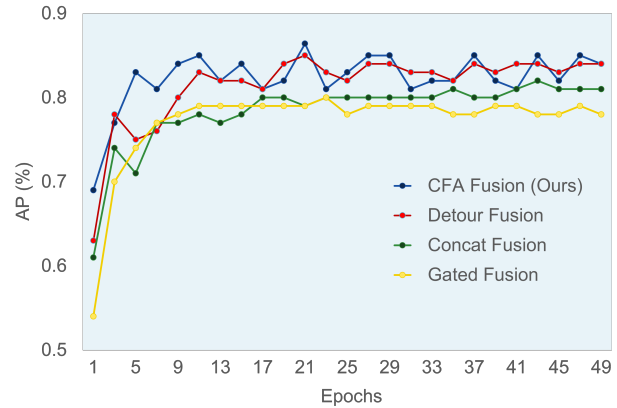Figure 7. Ablation studies on different setting of prefix dimensions D_p on our proposed CFA module

Figure 8. Comparative analysis (Epoch vs AP) with existing fusion methods with our proposed CFA fusion.

is illustrated in Figure 8. It can be seen from this analysis that the proposed CFA fusion module consistently achieves superior performance across all fusion mechanisms.

5. Conclusion

This research presents a new WS-VAD framework with a Cross-modal Fusion Adapter (CFA) module and a Hyperbolic Lorentzian Graph Attention (HLGAtt) module to detect anomaly events such as violence and nudity accurately. The CFA module addresses the imbalanced modality information issue and effectively facilitates multi-modal interaction by dynamically selecting the relevant audio features with corresponding visual features. Additionally, the HLGAtt module captures the hierarchical relationships within normal and abnormal representations, thereby improving the accuracy of separating normal and abnormal features. Through extensive experiments and ablation studies, it has been demonstrated that the proposed model outperforms existing violence and nudity detection methods.

References

[1] Salem AlMarri, Muhammad Zaigham Zaheer, and Karthik Nandakumar. A multi-head approach with shuffled segments for weakly-supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 132–142, 2024. 1, 2, 6

[2] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araújo. Bossa: Extended bow formalism for image classification. In *2011 18th IEEE International Conference on Image Processing*, pages 2909–2912, 2011. 6, 7

[3] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo De A AraúJo. Pooling in image representation: The visual codeword point of view. *Computer Vision and Image Understanding*, 117(5):453–465, 2013. 6, 7

[4] Weize Chen, Xu Han, Yankai Lin, Hexu Zhao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fully hyperbolic neural networks. *arXiv preprint arXiv:2105.14686*, 2021. 5

[5] David Alexander Forsyth and Margaret M. Fleck. Automatic detection of human nudes. *International Journal of Computer Vision*, 32:63–77, 1999. 3

[6] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016. 6

[7] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 6

[8] Xin Jin, Yuhui Wang, and Xiaoyang Tan. Pornographic image recognition via weighted multiple instance learning. *IEEE transactions on cybernetics*, 49(12):4412–4420, 2018. 3, 6, 7

[9] Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. *Int. J. Comput. Vision*, 46(1):81–96, 2002. 3

[10] Hamza Karim, Keval Doshi, and Yasin Yilmaz. Real-time weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6848–6856, 2024. 1, 2, 6

[11] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[12] Wayne Kelly, Andrew Donnellan, and Derek Molloy. Screening for objectionable images: A review of skin detection techniques. page 151–158, USA, 2008. IEEE Computer Society. 3

[13] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multisequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1395–1403, 2022. 1, 2, 6

[14] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013. 2

[15] WANG Liyuan, ZHANG Jing, YAO Jiacheng, and ZHUO Li. Porn streamer recognition in live video based on multimodal knowledge distillation. *Chinese Journal of Electronics*, 30(6):1096–1102, 2021. 3

[16] Jay Mahadeokar and Gerry Pesavento. Open sourcing a deep learning solution for detecting nsfw images. *Retrieved August*, 24:2018, 2016. 6, 7

[17] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997. 1

[18] Xinyu Ou, Hefei Ling, Han Yu, Ping Li, Fuhao Zou, and Si Liu. Adult image and video recognition by a deep multicontext network and fine-to-coarse strategy. 8(5), 2017. 3

[19] Wen-Feng Pang, Qian-Hua He, Yong-jian Hu, and Yan-Xiong Li. Violence detection in videos based on fusing visual and audio information. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2260–2264. IEEE, 2021. 1, 2, 6

[20] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579, 2018. 5

[21] Xiaogang Peng, Hao Wen, Yikai Luo, Xiao Zhou, Keyang Yu, Yigang Wang, and Zizhao Wu. Learning weakly supervised audio-visual violence detection in hyperbolic space. *arXiv preprint arXiv:2305.18797*, 2023. 1, 2, 4, 5, 6, 7, 8

[22] Christian Platzer, Martin Stuetz, and Martina Lindorfer. Skin sheriff: a machine learning solution for detecting explicit images. page 45–56, New York, NY, USA, 2014. Association for Computing Machinery. 3

[23] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021. 3

[24] Eric Qu and Dongmian Zou. Hyperbolic convolution via kernel point aggregation. *arXiv preprint arXiv:2306.08862*, 2023. 5

[25] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3379–3388, 2018. 2

[26] Sonali Samal, Yu-Dong Zhang, Thippa Reddy Gadekallu, and Bunil Kumar Balabantaray. Asyv3: Attention-enabled pooling embedded swin transformer-based yolov3 for obscenity detection. *Expert Systems*, 2023. 3, 6, 7

[27] Utsav Shah, Muhammad Aqmar, Mitsuru Nakazawa, and Björn Stenger. Content filtering in streaming video using domain adaptation. In *2021 17th International Conference on Machine Vision and Applications (MVA)*, pages 1–6. IEEE, 2021. 3, 6, 7

[28] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 1, 2, 5, 6

[29] Weijun Tan, Qi Yao, and Jingfeng Liu. Overlooked video classification in weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 202–210, 2024. 6

[30] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4975–4986, 2021. 1, 2, 6

[31] Hoang-Loc Tran, Quang-Huy Nguyen, Dinh-Duy Phan, Thanh-Thien Nguyen, Khac-Ngoc-Khoi Nguyen, and Duc-Lung Vu. Additional learning on object detection: A novel approach in pornography classification. In *Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications: 7th International Conference, FDSE 2020, Quy Nhon, Vietnam, November 25–27, 2020, Proceedings 7*, pages 311–324. Springer, 2020. 3, 6, 7

[32] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 7

[33] Xizi Wang, Feng Cheng, Shilin Wang, Huanrong Sun, Gongshen Liu, and Cheng Zhou. Adult image classification by a local-context aware network. *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2989–2993, 2018. 3

[34] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *European Conference on Computer Vision*, pages 729–745. Springer, 2022. 1, 2, 6

[35] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30:3513–3527, 2021. 1, 2, 6

[36] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer, 2020. 1, 2, 4, 5, 6

[37] Jiashuo Yu, Jinyu Liu, Ying Cheng, Rui Feng, and Yuejie Zhang. Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6278–6287, 2022. 2, 6

[38] Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16271–16280, 2023. 1, 2, 6

[39] Yiding Zhang, Xiao Wang, Chuan Shi, Xunqiang Jiang, and Yanfang Ye. Hyperbolic graph attention network. *IEEE Transactions on Big Data*, 8(6):1690–1701, 2021. 4

[40] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. *arXiv preprint arXiv:2302.05160*, 2023. 1, 2, 4, 6