Prompt-Enhanced Multiple Instance Learning for Weakly Supervised Video Anomaly Detection

Junxi Chen[1], Liang Li[2]*, Li Su[1,3]*, Zheng-Jun Zha[4], Qingming Huang[1,2,5]
[1]University of Chinese Academy of Sciences
[2]Key Lab of Intell. Info. Process., ICT, CAS, [3]Peng Cheng Lab,
[4]University of Science and Technology of China, [5] Key Lab of AI Safety, CAS
chenjunxi22@mails.ucas.ac.cn, liang.li@ict.ac.cn, {suli,qmhuang}@ucas.ac.cn

Abstract

Weakly-supervised Video Anomaly Detection (wVAD) aims to detect frame-level anomalies using only video-level labels in training. Due to the limitation of coarse-grained labels, Multi-Instance Learning (MIL) is prevailing in wVAD. However, MIL suffers from insufficiency of binary supervision to model diverse abnormal patterns. Besides, the coupling between abnormality and its context hinders the learning of clear abnormal event boundary. In this paper, we propose prompt-enhanced MIL to detect various abnormal events while ensuring clear event boundaries. Concretely, we design the abnormal-aware prompts by using abnormal class annotations together with learnable prompt, which can incorporate semantic priors into video features dynamically. The detector can utilize the semantic-rich features to capture diverse abnormal patterns. In addition, normal context prompt is introduced to amplify the distinction between abnormality and its context, facilitating the generation of clear boundary. With the mutual enhancement of abnormal-aware and normal context prompt, the model can construct discriminative representations to detect divergent anomalies without ambiguous event boundaries. Extensive experiments demonstrate our method achieves SOTA performance on three public benchmarks. The code is available at https://github.com/Junxi-Chen/PE-MIL.

1. Introduction

To identify anomaly at frame level in video, Video Anomaly Detection (VAD) has become vital in critical areas, *e.g.*, surveillance systems [19], medical imaging [33] and autonomous driving [1]. For great generalization ability across diverse scenes, researchers [6, 8, 29, 31, 32, 43, 45] turn to weakly-supervised VAD (wVAD) which only leverages video-level labels. Primarily, wVAD faces two key challenges: **1)** detecting complex anomalous patterns in varied scenarios where temporal relationship and visual appearance of anomalies exhibit substantial discrepancies; **2)** generating clear abnormal event boundaries in the absence of fine-grained boundary annotations.

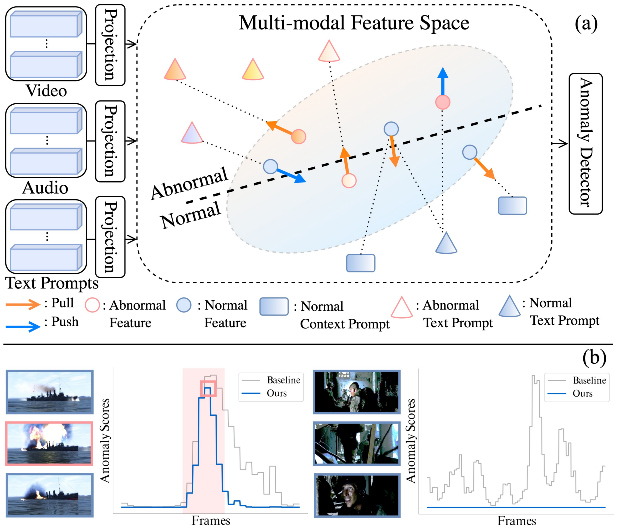


Figure 1. (a) Illustration of prompt-enhanced MIL. In multi-modal feature space, text prompts integrate abnormal-aware semantic priors into visual features. NCP incorporates normal semantic into ambiguous context feature. In such manner, our method learns a more discriminative representation to deliver a precise anomaly detection. (b) Anomaly detection examples of our method.

To tackle wVAD, prior works [52, 55] generate noisy frame-level abnormal labels and reduce noise subsequently, but such manner limits the generalization ability to unseen scenarios. Recently, Multi-Instance Learning (MIL) is leveraged by most methods [5, 6, 8, 23, 29, 45] to tackle wVAD due to its ability to model patterns under coarse-

*Corresponding authors

grained labels. To detect abnormal events by frame-level anomaly scores (1 for abnormal and 0 for normal events), MIL-based methods [5, 6, 24, 29, 42, 53] apply ranking loss to encourage the top scores in abnormal videos to be higher than that in normal videos. In this way, MIL can learn exclusive patterns in abnormal instances. However, MIL is notoriously known [18, 24, 25] to suffer from ambiguous event boundary and incompetence of modeling divergent abnormal patterns. To this end, some works modify the target [6, 25, 38, 54] or training strategies [18, 53] to facilitate learning of clear event boundary. Besides, several studies [6, 29, 45] introduce auxiliary information to facilitate modeling anomaly patterns in intricate scenarios.

Despite the progress, existing methods suffer from two major dilemmas. **1)** Binary labels are insufficient to capture intricate abnormal patterns, because they only indicate a general abnormal pattern. Besides, binary labels neglect the semantic relevance between abnormalities, which indicates the similar and unique patterns between abnormal events. **2)** Coupling of abnormality and context patterns impedes learning of clear boundary, *e.g.,* in Fig. 1b, an explosion (abnormality) is usually coupled with fire and smoke (context). But such abnormal context scene is rarely contained in normal videos. Consequently, the model learnt from coupled data is infeasible to decouple abnormality from its context, leading to ambiguous boundary. We notice that there exists strong semantic relevance between the contexts of abnormality and normality, under which the normal context can enrich the ambiguous boundary context to exhibit a discriminative pattern. Thus, modeling such relevance is able to improve the robustness of wVAD on highly coupled scenarios, resulting in clear event boundaries.

In this paper, we propose a novel prompt-enhanced MIL to capture diverse abnormalities with clear event boundaries, via introducing abnormal-aware and normal context prompt, as shown in Fig. 1a. Concretely, to construct abnormal-aware prompts, we obtain the embeddings of abnormal class annotations, augmented with learnable prompt. Prompt constraint loss is designed to ensure their semantic consistency, so as to obtain the abnormal-aware prompts with rich semantics. Next, we devise an event relevance reasoning module to dynamically guide the fine-grained alignment between the abnormal-aware prompts and video features. Through this manner, the video features incorporate accurate semantic prior to facilitate capturing complex and diverse abnormal patterns. For clear abnormal boundary, we further learn a normal context prompt through two-stage training, which serves as the comprehensive summary of the normal patterns. The normal context prompt can enrich boundary context feature for revealing their discriminative character. As such, the detector can better distinguish between context snippets and abnormal snippets, thereby enhancing the generation of clear event boundary.

The main contributions of this paper are as follows:

- We propose prompt-enhanced MIL for wVAD, which exploits abnormal-aware prompts to integrate semantic priors into video features for modeling divergent abnormal patterns precisely.
- We introduce the normal context prompt that serves as a summary of normal patterns, for enriching the boundary context feature. In doing so, a more discriminative character can be revealed to decouple anomaly and its context, resulting in a clear abnormal event boundary.
- Extensive experiments demonstrate that our method performs favourably against the state-of-the-art methods on three public datasets.

2. Related Work

2.1. Weakly-Supervised Video Anomaly Detection

In wVAD, video-level annotations are provided and frame-level anomaly scores are required. Most of the wVAD works apply MIL-based methods due to its ability to learn discriminative representation under weak labels. Sultani *et al*. [31] introduce the inaugural MIL approach featuring a ranking loss, in addition to presenting a large-scale VAD dataset. Later, Zhang *et al*. [54] explore an intra-bag loss, which is complementary to the MIL ranking loss. To catch a better sequential relationship for anomaly detection, Zhu *et al*. [57] use an attention module to model the temporal context of anomalies. As Graph Convolutional Network (GCN) prevailing [2], Cho *et al*. [6] model context and motion correlation by GCN to detect anomaly. Lv *et al*. [24] propose unbiased MIL by training model with both clustered confident and ambiguous sets to alleviate false alarm.

2.2. Prompt Learning in Video Understanding

Recent studies successfully extend prompt learning to video understanding tasks. For instance, Wang *et al*. [41] align video clips with text embeddings of category labels for action recognition task. Ju *et al*. [15] construct prompt templates by category labels and investigate the effect of label position within the templates. However, manual prompt design is time-consuming [10]. Some works [20, 29, 48] utilize the definitions provided by knowledge-base to create prompt templates. Wu *et al*. [46] learns semantic by predicting correct class labels with concatenated visual and text features. However, coarse-grained alignment learns confusing semantic, thereby leading to inaccuracy in anomaly detection. Different from them, we do a fine-grained alignment between visual and text features because wVAD is sensitive to noise. Furthermore, we utilize learnable prompt with proposed prompt constraint loss to integrate rich semantic features into visual features.

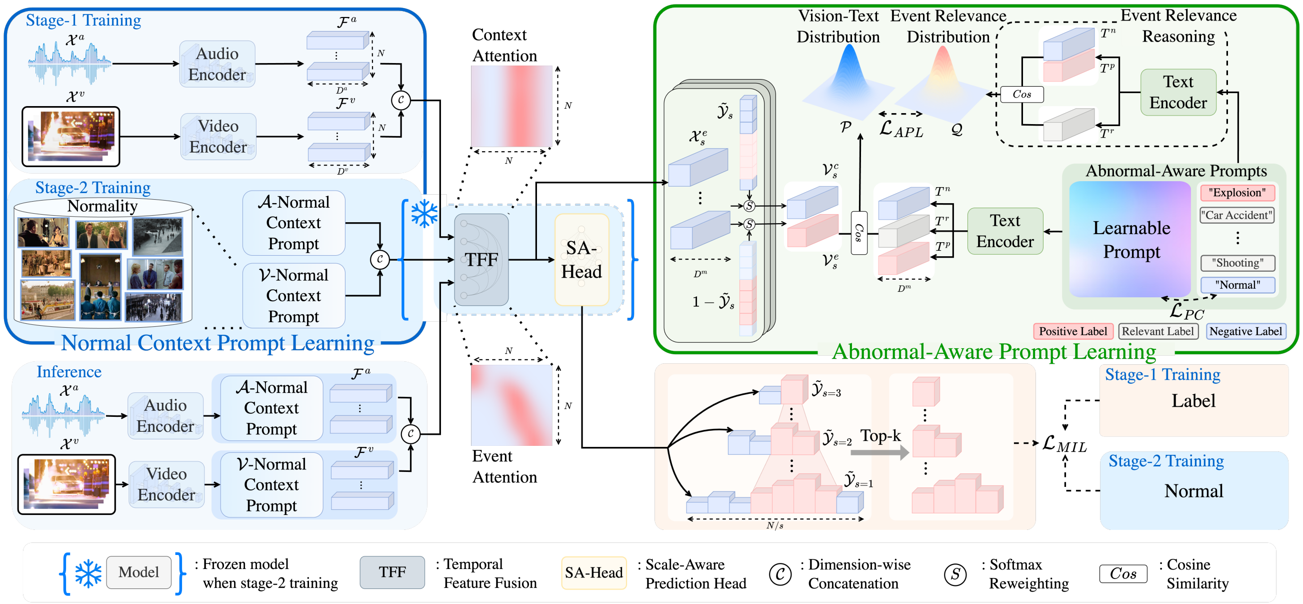


Figure 2. Pipeline of the proposed method: (1) Temporal Feature Fusion module (Sec. 3.2) and Scale-Aware Prediction Head (Sec. 3.3) are utilized to model the temporal relationships and generate multi-scale anomaly scores. (2) Abnormal-Aware Prompt Learning (Sec. 3.4) is applied to facilitate the intermediate feature incorporating a semantic priors by abnormal-aware prompts. (3) Normal Context Prompt (Sec. 3.5) is learned by two-stage training and enriches the boundary context feature to exhibit a discriminative pattern.

3. Method

3.1. Overview

The architecture of the overall framework is depicted in Fig 2. Specifically, given an untrimmed video $\mathcal{X}^v$ with corresponding audio $\mathcal{X}^a$, we employ pretrained backbone networks to extract video features $\mathcal{F}^a$ and audio features $\mathcal{F}^v$. To model the temporal relationship, these features are then passed to a transformer-based temporal feature fusion module, which leverages both context attention and event attention. Subsequently, the scale-aware prediction head is employed to predict the anomaly scores and facilitate modeling the abnormal patterns at different scales. To capture diverse abnormal patterns, abnormal-aware prompt learning is applied to incorporate rich semantic to intermediate features $\mathcal{X}^e$. Furthermore, we learn the normal context prompt guided by normal label with frozen model. The proposed normal context prompt can enrich the feature during model inference to generate clear event boundary.

3.2. Temporal Feature Fusion

Temporal feature fusion module is proposed to model long-range context and short-range event temporal dependencies by utilizing self-attention mechanism [37]. Considering the computational overhead, we divided input video $\mathcal{X}^v$ and audio $\mathcal{X}^a$ into 16-frame non-overlapping snippets. Pretrained frozen backbones are utilized to extract video and audio features, formulating snippet-level feature sequence $\mathcal{F}^v \in \mathbb{R}^{N\times D_v}$ and $\mathcal{F}^a \in \mathbb{R}^{N\times D_a}$ where N is the number of snippets and D_v, D_a are the feature dimension of video and audio respectively. Then video and audio features are concatenated along dimension to create multi-modal features $\mathcal{F} \in \mathbb{R}^{N\times(D_v+D_a)}$.

Inspired by the success of attention-based methods [7, 17, 34–36, 39] in multi-modal data, we adopt attention mechanism to model temporal relationship.The similarity matrix is computed and dynamic position encoding [29] $\mathcal{E} \in \mathbb{R}^{N\times N}$ is add to similarity matrix $\mathcal{M} \in \mathbb{R}^{N\times N}$ to incorporate position prior information:

$$\begin{aligned}\mathcal{M} &= f_q(\mathcal{F}) \cdot f_k(\mathcal{F})^\top + \mathcal{E},\\ \mathcal{E}_{j,k} &= \exp\left(-\left|\gamma(j-k)^2+\beta\right|\right),\end{aligned} \tag{1}$$

where $f(\cdot)$ corresponds to linear layers and the symbol $\top$ denotes the transpose operation. The variables $j \in [1, N]$ and $k \in [1, N]$ refer to the number of two snippets, while γ and β represent learnable weight and bias terms. Then compute context attention map and feature $\mathcal{F}^g \in \mathbb{R}^{N\times D_h}$ as follows:

$$\mathcal{F}^g = \text{softmax}\left(\frac{\mathcal{M}}{\sqrt{D_h}}\right) \cdot f_v(\mathcal{F}), \tag{2}$$

where D_h indicates the hidden dimension.

To focus on consecutive event snippets and solve the long-range noise, the similarity matrix is masked to capture short-range event dependencies :

$$\tilde{\mathcal{M}}_{ij} = \begin{cases} \mathcal{M}_{ij}, & j \in \left[\max\left(0, i - \lfloor \frac{w}{2} \rfloor\right), \min\left(i + \lfloor \frac{w}{2} \rfloor, N\right)\right] \\ -\infty, & \text{otherwise} \end{cases} \tag{3}$$

where $\tilde{\mathcal{M}} \in \mathbb{R}^{N \times N}$ refers to event similarity matrix and w is the event size of the mask. Afterwards,we compute the event context feature $\mathcal{F}^e$ following equation 2.

A learnable gate weight α is introduced to fuse context and event features. Subsequently, a residual connection is utilize followed by layer normalization to derive feature $\mathcal{F}^c \in \mathbb{R}^{N \times (D_v + D_a)}$ which can be formulated as:

$$\begin{aligned} \mathcal{F}^o &= \alpha \cdot \mathcal{F}^g + (1 - \alpha) \cdot \mathcal{F}^e \\ \mathcal{F}^c &= \text{LayerNorm}\left(\mathcal{F} + f_o\left(\text{Norm}\left(\mathcal{F}^o\right)\right)\right) \end{aligned} \tag{4}$$

where $\text{Norm}(\cdot)$ denotes a composite of power normalization [50] and L2 normalization.

3.3. Scale-Aware Prediction Head

To magnify varied-scale abnormal events, we propose a scale-aware prediction head, as shown in Fig 2. To obtain high-level semantic feature $X_s^e \in \mathbb{R}^{\lfloor N/s \rfloor \times D_m}$, a multilayer perceptron is applied:

$$\mathcal{X}_s^e = \text{Dropout}\left(\text{GELU}\left(\text{Conv}_s\left(\mathcal{F}^c\right)\right)\right) \tag{5}$$

where D_m is the semantic feature dimension and Conv_s refers to one-dimension convolutional layer with a stride of s. The module is followed by GELU [13] activation and dropout operation. Subsequently, the anomaly scores are generated from semantic feature, which can be denoted as:

$$\tilde{\mathcal{Y}}_s = \sigma\left(f_t\left(\text{Dropout}\left(\text{GELU}\left(\text{Conv}_s\left(\mathcal{X}^e\right)\right)\right)\right)\right) \tag{6}$$

where $f_t(\cdot)$ refers to causal convolution layer and $\sigma(\cdot)$ is the sigmoid activation function. $\mathcal{X}^e$ is extracted with stride 1 and $\tilde{\mathcal{Y}}_s \in \mathbb{R}^{\lfloor N/s \rfloor}$ indicates the predicted frame-level anomaly scores.

Following [43], we apply the MIL-based loss as the fundamental objective function. For abnormal videos, top-k anomaly scores are selected to reinforce the abnormal features and for normal videos, the maximum score is sampled to decrease the prominent anomaly score in normal video. Parameter k is set as follows:

$$k = \begin{cases} \lfloor \frac{N}{16 \times s} \rfloor + 1, & \mathcal{Y} = 1 \\ 1, & \mathcal{Y} = 0 \end{cases} \tag{7}$$

where $\mathcal{Y} \in \mathbb{R}$ refers to video-level ground-truth. $\mathcal{Y}$ equals 1 if it is an abnormal video and 0 if it is a normal video.

The video-level prediction $\hat{\mathcal{Y}}_{s2} \in \mathbb{R}$ can be computed as the mean of the top-k anomaly scores:

$$\hat{\mathcal{Y}}_s = \frac{1}{k} \sum_{i \in \text{top-}k} \tilde{\mathcal{Y}}_s^i \tag{8}$$

The MIL-based loss function is computed by binary cross-entropy as follows:

$$\mathcal{L}_{MIL} = -\mathcal{Y} \log\left(\hat{\mathcal{Y}}_s\right) - (1 - \mathcal{Y}) \log\left(1 - \hat{\mathcal{Y}}_s\right) \tag{9}$$

3.4. Abnormal-Aware Prompt Learning

Abnormal-aware Prompt Learning (APL) is proposed to facilitate modeling diverse anomalous patterns with semantic-rich visual features. Through APL, semantic prior from text prompts is incorporated into visual feature. The process of APL includes three steps which are event-context separation, abnormal-aware prompts construction and dynamic cross-modal alignment.

Firstly, we separate event and context feature to enable a fine-grained semantic learning. Since the videos may contain event and context instances, aligning all snippets with same prompt will confuse the model with unclear semantic. We leverage the scaled anomaly scores as activations to separate video-level event and context features as illustrated:

$$\begin{aligned} \mathcal{V}_s^e &= \frac{\exp(\mu \tilde{\mathcal{Y}}_s) - 1}{\sum_t (\exp(\mu \tilde{\mathcal{Y}}_s^t) - 1)} \cdot \mathcal{X}_s^e \\ \mathcal{V}_s^c &= \frac{\exp(\mu (1 - \tilde{\mathcal{Y}}_s)) - 1}{\sum_t (\exp(\mu (1 - \tilde{\mathcal{Y}}_s^t)) - 1)} \cdot \mathcal{X}_s^e \\ \mathcal{V}_s &= \{\mathcal{V}_s^e, \mathcal{V}_s^c\} \end{aligned} \tag{10}$$

where $\mathcal{V}_s^e \in \mathbb{R}^{D_m}$ and $\mathcal{V}_s^c \in \mathbb{R}^{D_m}$ refer to the event and context feature respectively and t indicates the number of the snippet. Predetermined scaling factor is denoted by μ that works in conjunction with the $\exp(\cdot)$ operation to amplify activations with high confidence. Then the event feature $\mathcal{V}_s^e \in \mathbb{R}^{D_m}$ and context feature $\mathcal{V}_s^c \in \mathbb{R}^{D_m}$ are concatenated together to form overall visual feature $\mathcal{V}_s$. For normal videos, only event visual feature is sampled as $\mathcal{V}_s = \mathcal{V}_s^e$.

Secondly, we composite abnormal-aware prompts as semantic cues. To determine the precise semantic relationship for capturing various abnormal patterns, the annotations are divided into three sub-classes which are positive, relevant and negative class. For an abnormal video, the positive annotation indicates the abnormal classes presented within the video, while the relevant annotation refers to collection of nonexistent abnormal class labels. The negative label denotes normal label. For normal videos, the positive and relevant annotations are the normal label and the negative annotation corresponds to all abnormal labels. The original abnormal text labels are transformed into embedding tensors through the tokenizer and embedding layer of the text backbone. For instance, the initial embedding T_{init}^p of positive label can be derived as:

$$T_{init}^p = \text{Embed}(\text{Token}(\text{'Positive Label'})) \tag{11}$$

Considering the original class annotations are too succinct to summarize complex events and scarce of rich semantic, we introduce learnable prompt to the original text embedding to increase the generalization capability and derive semantic-rich text feature. We concatenate the learnable prompt with the embedding tensor:

$$T^p_{embed} = \{T^l, T^p_{init}\} \tag{12}$$

where $T^l \in \mathbb{R}^{L \times D_m}$ denotes learnable prompt and L is the length of learnable prompt. Subsequently, the semantic-rich label embedding is passed to the text encoder to obtain text feature of sub-classes, denoted as T^p, T^r and T^n. They are then concatenated to form the text feature of each label:

$$T = \{T^p, T^r, T^n\} \tag{13}$$

Finally, by dynamic cross-modal alignment through event relevance reasoning, we enrich the visual features with semantic prior to learn class-specific abnormal patterns. The visual-text correlation distribution P is computed as:

$$\begin{aligned} \psi(\mathcal{V}_s, T) &= \frac{\mathcal{V}_s \cdot T^\top}{\|\mathcal{V}_s\| \|T\|} \\ P(\mathcal{V}_s) &= \frac{\exp(\psi(\mathcal{V}_s, T)/\tau)}{\sum_{k=1}^{C+1} \exp\left(\psi\left(\mathcal{V}_s, T_k\right)/\tau\right)} \end{aligned} \tag{14}$$

where C refers to the number of abnormal class and τ corresponds to temperature factor. We propose event relevance reasoning module to dynamically calculate the semantic relevance, and formulte the alignment objective $\mathcal{O}$. The process can be denoted as follows:

$$\mathcal{O} = \begin{bmatrix} \mathcal{O}^e_p & \mathcal{O}^e_r & \mathcal{O}^e_n \\ \mathcal{O}^c_p & \mathcal{O}^c_r & \mathcal{O}^c_n \end{bmatrix} = \begin{bmatrix} 1 & c \cdot \psi(T^p, T^r) & 0 \\ 0 & c \cdot \psi(T^n, T^r) & 1 \end{bmatrix} \tag{15}$$

where c is a scaling factor which equals 1 if it is a normal video. e, c in superscript correspond to the event and context features. p, r, n in subscript refer to the positive, relevant and negative text features. The target distribution can be computed as:

$$Q(\mathcal{V}_s) = \frac{\exp\left(\mathcal{O}^v_t\right)}{\sum_{k=1}^{C+1} \exp\left(\mathcal{O}^v_{t,k}\right)} \tag{16}$$

The abnormal-aware prompt learning loss $\mathcal{L}_{APL}$ can be computed by Kullback-Leibler divergence, as follows:

$$\mathcal{L}_{APL} = \mathbb{E}_{P \sim P(v)} \left[\log P(\mathcal{V}_s) - \log Q(\mathcal{V}_s)\right] \tag{17}$$

To ensure the consistency between learnable prompt and class annotation. Prompt constraint loss $\mathcal{L}_{PC}$ is introduced:

$$\mathcal{L}_{PC} = 1 - \frac{T^l \cdot T^\top_{init}}{\|T^l\| \, \|T_{init}\|} \tag{18}$$

In the first training phase, the overall objective function can be denoted as:

$$\mathcal{L} = \mathcal{L}_{MIL} + \lambda \mathcal{L}_{APL} + \beta \mathcal{L}_{PC} \tag{19}$$

where hyper-parameter λ and β are used to balance the loss. By optimizing the objective function, the model can leverage semantic-rich feature to generate a more discriminate representation. Consequently, our method can detect various anomaly patterns precisely.

3.5. Normal Context Prompt

Normal Context Prompt (NCP) is proposed to generate a clear abnormal event boundary by enriching the ambiguous context feature. NCP is devised to summary the latent normal event distribution of the trained model. NCP $\mathcal{V}_{NCP} \in \mathbb{R}^{K \times D_v}$ can be interpreted as a normal visual feature sequence, where K refers to the NCP length.

We apply a two-stage training strategy to learn NCP as illustrated in Fig. 2. In first stage, the model captures abnormal and normal patterns. In the second stage, we freeze the model for NCP to fit captured normal distribution. NCP is passed to the model as input and the provided ground-truth label is 0. We compute the mean square error loss for NCP to learn normal distribution. The loss can be denoted as:

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i=1}^{N} (\mathcal{Y} - \tilde{\mathcal{Y}}^i)^2 \tag{20}$$

During inference phase, we concatenate the normal context prompt with multi-modal feature sequence. The enriched feature is passed to temporal feature fusion module which can fuse the multi-model feature with NCP by attention mechanism. It amplifies the feature discrepancy by dynamically integrating abnormal context with enrichment from normal domain, taking advantage of the higher feature relevance. The detector can leverage the distinction to determine accurate event boundary.

4. Experimental Results

4.1. Datasets

UCF-Crime [31] encompasses 13 distinct anomaly categories, originating from diverse scenes, including streets, family rooms, and shopping malls. The dataset comprises 1610 training videos and 290 test videos.

ShanghaiTech [19] comprises 13 campus scenes from surveillance system with fixed point of view. It contains 238 videos in training set and 199 videos in the test set.

XD-Violence [45] is the most extensive dataset currently available for wVAD. It comprises videos gathered from various sources, including movies, games, and car cameras. The dataset is challenging because it contains rich artistic expressions such as changing perspective and dynamic camera movements. This dataset includes 3954 training videos with video-level annotations, 800 test videos with frame-level labels, and encompasses six distinct anomaly categories. In addition, it offers video with audio, facilitating anomaly detection by harnessing multi-modal cues.

Supervision	Methods	Feature	AP(%)
Semi-Supervised	SVM Baseline	I3D+VGGish	50.78
	OCSVM [30]	I3D+VGGish	27.25
	Conv-AE [12]	I3D+VGGish	30.77
Weakly-Supervised	MIL-Rank [31]	C3D RGB	73.20
	CA-VAD [4]	I3D RGB	76.90
	RTFM [32]	I3D RGB	77.81
	CRFD [44]	I3D RGB	75.90
	DDL [28]	I3D RGB	80.72
	MSL [18]	VideoSwin-RGB	78.59
	S3R [42]	I3D RGB	80.26
	MGFN [5]	VideoSwin-RGB	80.11
	UR-DMU [56]	I3D RGB	81.66
	Zhang *et al.* [53]	I3D+VGGish	81.43
	CMA-LA [27]	I3D+VGGish	83.54
	MACIL-SD [49]	I3D+VGGish	83.40
	CoMo [6]	I3D RGB	81.30
	PEL4VAD [29]	I3D RGB	85.59
	HyperVD [26]	I3D+VGGish	85.67
	Ours	I3D RGB	**88.05**
	Ours	I3D+VGGish	**88.21**(+2.54)

Table 1. Comparison with other methods on XD-Violence.

4.2. Evaluation Metrics

Following previous works [8, 31, 54], we opt the Area Under the Curve (AUC) of the frame-level Receiver Operating Characteristic (ROC) curve as the evaluation metric, for assessing the performance of our method on UCF-Crime and ShanghaiTech datasets. For XD-Violence, following [18, 45, 53], we use Average Precision (AP) as the metric.

4.3. Implementation Details

Consistent with existing methods [29, 44], we encode the videos into 1024-dimension video features by RGB-stream I3D [3] video encoder, which is pretrained on Kinetics [16] dataset. For audio features, we leverage VGGish [14] audio encoder pretrained on YouTube [14] dataset. Each snippet consists of 16 frames. The batch size is 128 and the learning rate is 5×10^{-4} with a cosine decay strategy. The window size w is 9. The NCP length K is 35 for XD-Violence and UCF-Crime, and 5 for ShanghaiTech. λ and β are 1 and 8 respectively for model training. The scales of s are 2 and 3. λ with 0.001 is applied to balance multi-scale loss. In comparison, we reproduce other methods by released codes. More implementation details are in the supplementary.

4.4. Comparisons with SOTA methods

Results on XD-Violence. The proposed method is compared with following SOTA methods, which can be categorized as semi-supervised methods [12, 30] and weakly supervised methods [4–6, 18, 26–29, 31, 32, 42, 44, 49, 53, 56]. The results are shown in Table 1. Our method surpasses all previous semi-supervised methods and weakly-supervised methods. Notably, when utilizing the same I3D-RGB video features and VGGish audio features, our method achieves an absolute gain of 2.54% in terms of the AP compared to the best previous method [26]. This superiority results from precise modeling of divergent abnormal patterns by APL. NCP also contributes to the deliver a more precise anomaly detection results by generating clear event anomaly boundary and eliminating false alarm.

Supervision	Methods	Feature	AUC(%)
Semi-Supervised	Mem-AE [11]	-	71.20
	HF^2-VAD [21]	-	76.20
	DLAN-AC [47]	-	74.70
Weakly-Supervised	MIL-Rank [31]	C3D RGB	86.30
	GCN [55]	TSN RGB	84.44
	CLAWS [51]	C3D RGB	89.67
	AR-Net [38]	RGB+Flow	91.24
	MIST [9]	I3D RGB	94.83
	CRFD [44]	I3D RGB	97.48
	RTFM [32]	I3D RGB	97.21
	MSL [18]	VideoSwin-RGB	97.32
	NL-MIL [25]	I3D RGB	97.43
	S3R [42]	I3D RGB	97.48
	UMIL [24]	X-CLIP RGB	96.78
	Ours	I3D RGB	**98.35**(+0.87)

Table 2. Comparison with other methods on ShanghaiTech.

Supervision	Methods	Feature	AUC(%)
Semi-Supervised	Conv-AE [12]	-	50.60
	Lu *et al.* [22]	-	76.20
	GODS [40]	BoW+TCN	70.46
Weakly-Supervised	MIL-Rank [31]	C3D RGB	75.41
	GCN [55]	TSN RGB	82.12
	MIST [9]	I3D RGB	82.30
	CRFD [43]	I3D RGB	84.89
	RTFM [32]	I3D RGB	84.30
	MSL [18]	VideoSwin-RGB	85.62
	PEL4VAD [29]	I3D RGB	85.62
	NL-MIL [25]	I3D RGB	85.63
	S3R [42]	I3D RGB	85.99
	CoMo [6]	I3D RGB	86.10
	UMIL [24]	X-CLIP RGB	86.75
	Ours	I3D RGB	**86.83** (+0.73)

Table 3. Comparison with other methods on UCF-Crime.

Results on ShanghaiTech. Table 2 presents the performance comparisons on the ShanghaiTech dataset. Our method demonstrates superior performance in terms of AUC when compared to previous semi-supervised approaches [11, 21, 47] and weakly-supervised methods [9, 18, 24, 25, 29, 31, 32, 38, 42, 44, 51, 55]. Specifically, when using the same I3D features, our method outperforms the state-of-the-art methods. APL facilitates modeling of divergent abnormal patterns and NCP helps to generate of clear event anomaly scores. They contribute to more accurate and reliable anomaly detection.

Results on UCF-Crime. Table 3 presents the performance comparisons on UCF-Crime dataset. Our method demonstrates favorable performance comparing to current methods [6, 9, 12, 18, 22, 24, 25, 29, 31, 32, 40, 42, 43, 55]. It is noted that compared to the other datasets, our method does not bring much gains. Our conjecture is that different abnormal events in UCF-Crime dataset exhibit a high degree

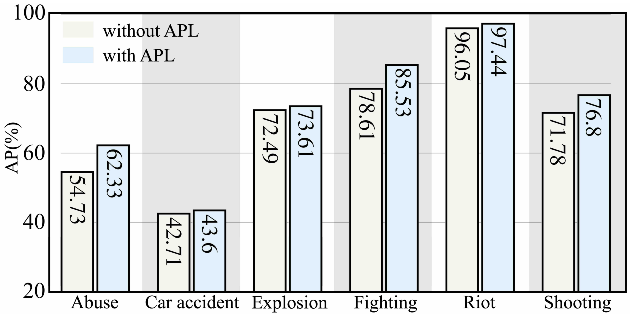


Figure 3. AP results w.r.t. sub-classes on XD-Violence

of homogeneity and fixed-view surveillance video results in a low level of coupling between abnormal events and context. The attribute of the abnormality diminishes the impact of our method.

Baseline	SA-Head	APL	NCP	AP(%)
✓				78.73
✓	✓			80.82
✓	✓	✓		86.62
✓	✓	✓	✓	88.21

Table 4. Ablation studies of proposed modules on XD-Violence.

ERR	LP	$\mathcal{L}_{PC}$	AP(%)
			85.61
✓			87.78
✓	✓		87.45
✓	✓	✓	88.21

Table 5. Ablation studies of modules in APL on XD-Violence.

K	0	1	5	10	35	50
AP(%)	86.62	87.47	87.87	88.09	88.21	88.20

Table 6. Ablation studies of NCP length on XD-Violence.

4.5. Ablation Studies

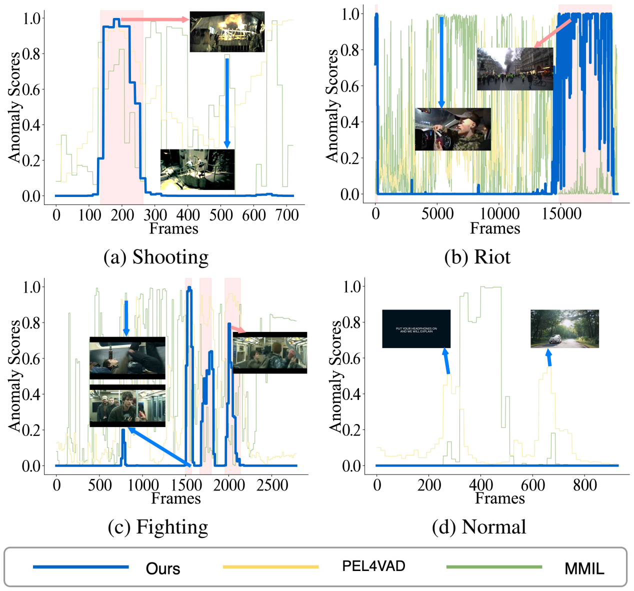
Effect of proposed module. We conduct ablation studies of proposed modules on XD-Violence dataset as shown in Table 4. To prove the necessity of capturing multi-scale abnormality, we compare our method with Scale-Aware prediction Head (SA-Head) with baseline. The results show a 2.9% improvement in terms of AP, which illustrate that learning multi-scale abnormal events can capture the temporal abnormal pattern better, thus improving detection performance. Besides, to illustrate the effect of Abnormal-aware Prompt Learning (APL), we compare the baseline with/without APL. From the results, we note that APL can bring 5.8% performance gain in AP, which indicates the importance of semantic prior in modeling diverse abnormal patterns. In addition, we illustrate the impact of NCP which can lead to 1.59% improvement in AP. This result demonstrates that NCP can decouple the abnormal context and event effectively, reducing boundary enlargement and false alarm to boost detection performance. To evaluate the ability of capturing diverse abnormalities, we compare AP of sub-classes of the model without APL (green bar) and the model with APL (blur bar) in Fig. 3. The model with APL outperforms the one without APL, which proves that APL can facilitate detecting diverse abnormal events.

Effect of modules in APL. To demonstrate effectiveness of each module and loss function in APL, we conduct ablation studies in APL, as shown in Table 5. Intending to demonstrate the necessity of doing fine-grained alignment based on event relevance, we compare our method using Event Relevance Reasoning (ERR) with the baseline using constant factors as target. The result shows the ERR can gain 2.17% increment in AP, which reveals the event relevance plays a vital role in modeling complex relationship of abnormal patterns and essence in diverse abnormalities. In addition, to verify the effect of learnable prompt, we conduct experiments between raw text embedding and embedding with learnable prompt. Notably, the performance drops by 0.34 % in term of AP without the prompt constraint loss. This illustrates the significance of prompt constraint loss for generating semantic related abnormal-aware prompts. Together with learnable prompt and prompt constraint loss, our method can boost performance by 0.43%, which verifies the effectiveness of semantic-rich abnormal aware prompts in capturing diverse abnormal patterns.

Effect of NCP length. To illustrate effectiveness of NCP, we conduct ablation studies with varied lengths K of NCP, as shown in Table 6. Notably, the NCP can bring 0.8% gain in AP with length of 1 and maximum 1.59% gain in AP with length of 35, which proves the effect of NCP in detecting anomaly precisely.

4.6. Qualitative Results

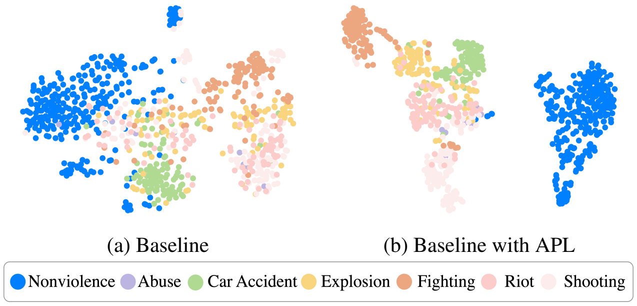
Anomaly Scores. To intuitively substantiate the effectiveness of our approach, the anomaly scores predicted by our method are visualized on the most challenging XD-Violence dataset in Fig. 4 compared to other methods [29, 31]. As demonstrated in Fig. 4a, our method effectively predict precise anomaly scores while significantly mitigating the occurrence of ambiguous boundary and false alarm, compared to the SOTA method [29]. Fig. 4b and Fig. 4c exemplify the proficiency of our method in predicting precise anomaly scores for long-term anomaly videos with multi-segment variable types and subtle intervals between anomalies. Fig. 4d further illustrates the efficacy of mitigating false alarm in challenging normal videos. The capability of detecting diverse abnormal patterns demonstrates that effectiveness of semantic prior in capturing various abnormalities. The decoupling of abnormal event and



(a) Shooting (b) Riot

(c) Fighting (d) Normal

Figure 4. Qualitative results of our method on XD-Violence test video. The pink square indicates the section where abnormal events occur. The Y-axis represents anomaly scores, while the X-axis represents the frame number of videos.



(a) Baseline (b) Baseline with APL

Figure 5. Feature distribution of the semantic feauture without and with APL using t-SNE

context further prove that NCP can enrich the ambiguous context feature to exhibit a more discriminative character.

Feature Distribution. For better comprehension of the APL module, we employ t-SNE for visualizing the features $\mathcal{X}^e$ from the intermediate layer. As depicted in Fig. 5a, the disparity between normal and abnormal features is minimal without APL, where some abnormal features are mingled with the normal cluster. With APL, a clear gap emerges between the abnormal and normal clusters. Moreover, all the abnormal features are exclusively grouped within the correct cluster, as demonstrated in Fig. 5b. This illustrates APL can facilitate to increase discrepancy between normal and abnormal events, leading to accurate anomaly detection.

Attention Map. To demonstrate the effect of NCP, we visualize the anomaly scores and corresponding attention maps with and without NCP. In Fig. 6a, Fig. 6b and Fig. 6c, we illustrate the anomaly scores with/without NCP by blue and gray lines, respectively. By introducing NCP, the model effectively establishes clear abnormal event boundaries. From Fig. 6d, Fig. 6e and Fig. 6f, we find that the attention maps without NCP are scattered. Instead, with the enrichment of NCP, our method can highlight the abnormal features and decoupled the context features, as shown in Fig. 6g, Fig. 6h and Fig. 6i. By employing NCP, the distinctiveness of the context is enhanced, leading to precise anomaly detection.

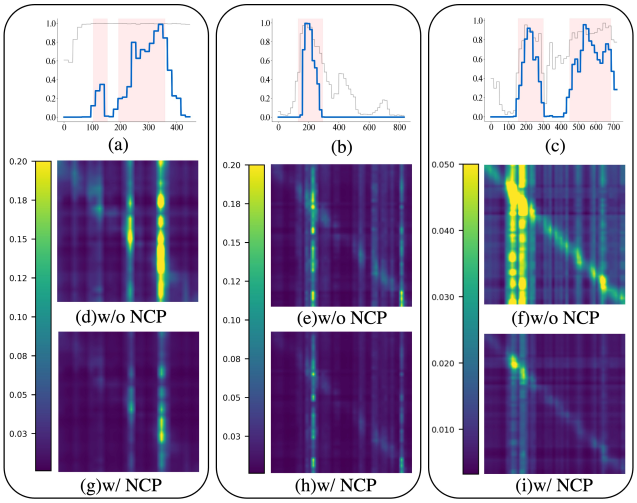


Figure 6. (a), (b) and (c) are visualization of anomaly scores on three hard cases with coupled boundary. (d), (e) and (f) are the attention maps of above cases without NCP. (g), (h) and (i) are the attention maps with NCP (only maps of visual features are shown).

5. Conclusions

This paper proposes the prompt-enhanced MIL to capture diverse abnormal patterns with a clear abnormal event boundary for wVAD. Given the embeddings of abnormal class annotations, we first introduce learnable prompt to augment them and design a constraint loss to guarantee their semantic consistency, thus gaining abnormal-aware prompt. Next, we incorporate the semantic prior into video features by aligning them with the learned prompt. In this way, the model can use semantic-rich features to capture diverse abnormal patterns. Further, normal context prompt is introduced as a summary of normal patterns to amplify the distinction of abnormality and abnormal context. Ambiguous context feature is enriched to generate a clear event boundary. Extensive experiments show our method achieves SOTA performance on three public benchmarks.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China: 62322211, 61931008, 62236008, 62336008, U21B2038, 62225207, the Key R&D Plan Project of Zhejiang Province (No.2024C01023), and Fundamental Research Funds for the Central Universities (E2ET1104).

References

[1] Daniel Bogdoll, Maximilian Nitsche, and J Marius Zöllner. Anomaly detection in autonomous driving: A survey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4488–4499, 2022. 1

[2] Shaofei Cai, Liang Li, Xinzhe Han, Shan Huang, Qi Tian, and Qingming Huang. Semantic and correlation disentangled graph convolutions for multilabel image recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[4] Shuning Chang, Yanchao Li, Shengmei Shen, Jiashi Feng, and Zhiying Zhou. Contrastive attention for video anomaly detection. *IEEE Transactions on Multimedia*, 24:4067–4076, 2021. 6

[5] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 387–395, 2023. 1, 2, 6

[6] MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun Lee. Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12137–12146, 2023. 1, 2, 6

[7] Gaoxiang Cong, Liang Li, Yuankai Qi, Zheng-Jun Zha, Qi Wu, Wenyu Wang, Bin Jiang, Ming-Hsuan Yang, and Qingming Huang. Learning to dub movies via hierarchical prosody models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2023. 3

[8] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 6

[9] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14009–14018, 2021. 6

[10] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. 2

[11] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019. 6

[12] 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

[13] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[14] 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

[15] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer, 2022. 2

[16] 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

[17] Liang Li, Xingyu Gao, Jincan Deng, Yunbin Tu, Zheng-Jun Zha, and Qingming Huang. Long short-term relation transformer with global gating for video captioning. *IEEE Transactions on Image Processing*, 31:2726–2738, 2022. 3

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

[19] W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 5

[20] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3003–3018, 2022. 2

[21] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13588–13597, 2021. 6

[22] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 6

[23] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE Transactions on Image Processing*, page 4505–4515, 2021. 1

[24] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8022–8031, 2023. 2, 6

[25] Seongheon Park, Hanjae Kim, Minsu Kim, Dahye Kim, and Kwanghoon Sohn. Normality guided multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2665–2674, 2023. 2, 6

[26] 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. 6

[27] Yujiang Pu and Xiaoyu Wu. Audio-guided attention network for weakly supervised violence detection. In *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 219–223. IEEE, 2022. 6

[28] Yujiang Pu and Xiaoyu Wu. Locality-aware attention network with discriminative dynamics learning for weakly supervised anomaly detection. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 6

[29] Yujiang Pu, Xiaoyu Wu, and Shengjin Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *arXiv preprint arXiv:2306.14451*, 2023. 1, 2, 3, 6, 7

[30] Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *Advances in neural information processing systems*, 12, 1999. 6

[31] 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, 7

[32] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, JohanW. Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. *Cornell University - arXiv,Cornell University - arXiv*, 2021. 1, 6

[33] Yu Tian, Guansong Pang, Fengbei Liu, Yuyuan Liu, Chong Wang, Yuanhong Chen, Johan Verjans, and Gustavo Carneiro. Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2022. 1

[34] Yunbin Tu, Liang Li, Li Su, Shengxiang Gao, Chenggang Yan, Zheng-Jun Zha, Zhengtao Yu, and Qingming Huang. I2 transformer: Intra-and inter-relation embedding transformer for tv show captioning. *IEEE Transactions on Image Processing*, 31:3565–3577, 2022. 3

[35] Yunbin Tu, Chang Zhou, Junjun Guo, Huafeng Li, Shengxiang Gao, and Zhengtao Yu. Relation-aware attention for video captioning via graph learning. *Pattern Recognition*, 136:109204, 2023.

[36] Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. Smart: Syntax-calibrated multi-aspect relation transformer for change captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[38] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2020. 2, 6

[39] Hao Wang, Zheng-Jun Zha, Liang Li, Xuejin Chen, and Jiebo Luo. Semantic and relation modulation for audio-visual event localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[40] Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8201–8211, 2019. 6

[41] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2

[42] 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. 2, 6

[43] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, page 3513–3527, 2021. 1, 4, 6

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

[45] 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*, page 322–339. 2020. 1, 2, 5, 6

[46] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. *arXiv preprint arXiv:2308.11681*, 2023. 2

[47] Zhiwei Yang, Peng Wu, Jing Liu, and Xiaotao Liu. Dynamic local aggregation network with adaptive clusterer for anomaly detection. In *European Conference on Computer Vision*, pages 404–421. Springer, 2022. 6

[48] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022. 2

[49] 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. 6

[50] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017. 4

[51] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly

supervised learning with normalcy suppression for anomalous event detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 358–376. Springer, 2020. 6

[52] Muhammad Zaigham Zaheer, Arif Mahmood, Hochul Shin, and Seung-Ik Lee. A self-reasoning framework for anomaly detection using video-level labels. *IEEE Signal Processing Letters*, 27:1705–1709, 2020. 1

[53] 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. 2, 6

[54] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4030–4034. IEEE, 2019. 2, 6

[55] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1237–1246, 2019. 1, 6

[56] 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. 6

[57] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211*, 2019. 2