Text Prompt with Normality Guidance for Weakly Supervised Video Anomaly Detection

Zhiwei Yang[1], Jing Liu[1]*, Peng Wu[2]

[1]Guangzhou Institute of Technology, Xidian University, Guangzhou, China
[2]School of Computer Science, Northwestern Polytechnical University, Xi'an, China

{zwyang97, neouma}@163.com, xdwupeng@gmail.com

Abstract

*Weakly supervised video anomaly detection (WSVAD) is a challenging task. Generating fine-grained pseudo-labels based on weak-label and then self-training a classifier is currently a promising solution. However, since the existing methods use only RGB visual modality and the utilization of category text information is neglected, thus limiting the generation of more accurate pseudo-labels and affecting the performance of self-training. Inspired by the manual labeling process based on the event description, in this paper, we propose a novel pseudo-label generation and self-training framework based on **T**ext **P**rompt **w**ith Normality **G**uidance (TPWNG) for WSVAD. Our idea is to transfer the rich language-visual knowledge of the contrastive language-image pre-training (CLIP) model for aligning the video event description text and corresponding video frames to generate pseudo-labels. Specifically, We first fine-tune the CLIP for domain adaptation by designing two ranking losses and a distributional inconsistency loss. Further, we propose a learnable text prompt mechanism with the assist of a normality visual prompt to further improve the matching accuracy of video event description text and video frames. Then, we design a pseudo-label generation module based on the normality guidance to infer reliable frame-level pseudo-labels. Finally, we introduce a temporal context self-adaptive learning module to learn the temporal dependencies of different video events more flexibly and accurately. Extensive experiments show that our method achieves state-of-the-art performance on two benchmark datasets, UCF-Crime and XD-Violence, demonstrating the effectiveness of our proposed method.*

1. Introduction

Anomaly detection has been widely researched and applied in various fields, such as computer vision [23, 35, 40, 43,
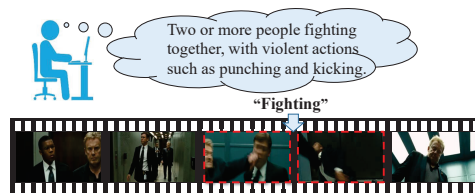

Figure 1. Illustration of the manual video frame labeling process.

49], natural language processing [1], and intelligent optimization [29]. One of the most important research issues is the video anomaly detection (VAD). The main purpose of VAD is to automatically identify events or behaviors in the video that are inconsistent with our expectations.

Due to the rarity of anomalous events and the difficulty of frame-level labeling, current VAD methods focus on semi-supervised [14, 16, 18] and weakly supervised [11, 26, 52] paradigms. Semi-supervised VAD methods aim to learn normality patterns from normal data, and deviations from this pattern are considered as anomalies. However, due to the lack of discriminative anomaly information in the training phase, these models are often prone to overfitting, leading to poor performance in complex scenarios. Subsequently, weakly supervised video anomaly detection (WSVAD) methods came into prominence. WSVAD involves both normal and abnormal videos with video-level labels in the training phase, but the exact location of abnormal frames is unknown. Current WSVAD methods mainly include one-stage methods based on multi-instance learning (MIL) [17, 26, 27] and two-stage methods based on pseudo-label self-training [6, 11, 51, 53]. While the one-stage methods based on MIL show promising results, this paradigm tends to focus on video snippets with prominent anomalous features and suboptimal attention to minor anomalies, thus limiting its further performance improvement.

In contrast to the one-stage methods mentioned above, two-stage methods based on pseudo-label self-training generally use an off-the-shelf classifier or MIL to obtain initial

*Corresponding authors.

pseudo-labels, and then train the classifier with further refined pseudo-labels. Because these methods train the classifier directly with the generated fine-grained pseudo-labels, they show great potential in performance. However, these methods still have two aspects that have not been considered: first, the generation of pseudo-labels is based only on visual modality and lacks the utilization of textual modality, which limits the accuracy and completeness of the generated pseudo-labels. Second, the mining of temporal dependencies among video frames is insufficient.

To further exploit the potential of pseudo-label-based self-training on WSVAD, we dedicate to investigating the two problems mentioned above in this paper. **Our motivation for the first question** is that we explore how the textual modal information can be effectively utilized to assist in generating pseudo-labels. Recalling our manual process of video frame labeling, we mainly based on textual definitions of anomalous events, i.e., prior knowledge about anomalous events, to accurately locate the video frames. As illustrated in Fig. 1, assuming that we need to annotate the abnormal video frames that contain "fighting" event, we will first associate the textual definition of "fighting" and then look for matching video frames, which is actually a process of text-image matching based on prior knowledge. Inspired by this process, we associate a highly popular and powerful contrastive language-image pre-training (CLIP) [19] model to assist us in achieving this goal. On the one hand, the CLIP learns a large number of image-text pairs on the web, and thus has a highly rich prior knowledge; on the other hand, the CLIP is trained by comparative learning, which empowers it with excellent image-text alignment capabilities. **For the second motivation**, because different video events have diverse durations, this leads to different ranges of temporal dependencies. Existing methods either do not consider temporal dependencies or only consider dependencies within a fixed temporal range, leading to inadequate modeling of temporal dependencies. Therefore, in order to achieve more flexible and adequate modeling of temporal dependencies, we should investigate methods that can adaptively learn temporal dependencies of different lengths.

Based on the above two motivations, we propose a novel pseudo-label generation and self-training framework based on **T**ext **P**rompt **w**ith **N**ormality **G**uidance (TPWNG) for WSVAD. Our main idea is to utilize the CLIP model to match the textual descriptions of video events with the corresponding video frames, and then infer the pseudo-labels from match similarities. However, since the CLIP model is trained at the image-text level, it may suffer from domain bias and lacks the ability to learn temporal dependencies in videos. In order to better transfer the prior knowledge of CLIP to the WSVAD task, we first construct a contrastive learning framework by designing two ranking losses and a distributional inconsistency loss to fine-tune the CLIP

model for domain adaptation under the weakly-supervised setting. To further improve the accuracy of aligning the descriptive text of video events with video frames, we employ learnable textual prompts to facilitate the text encoder of CLIP to generate more generalized textual embedding features. On this basis, we propose a normality visual prompt (NVP) mechanism to aid this process. In addition, because abnormal videos contain normal video frames as well, we design a pseudo-label generation (PLG) module based on normality guidance, which can reduce the interference caused by individual normal video frames to the alignment of abnormal video frames, thus facilitating the obtaining of more accurate frame-level labels.

Furthermore, to compensate for the lack of temporal relationship modeling in CLIP as well as to more flexible and adequately mine the temporal dependencies between video frames, we introduce a temporal context self-adaptive learning (TCSAL) module for temporal dependency modeling, inspired by the work [25]. TCSAL allows the attention module in the Transformer to adaptively adjust the attention span according to the inputs by designing a temporal span adaptive learning mechanism. This can facilitate the model to capture the temporal dependencies of video events of different durations more accurately and flexibly.

Overall, our main contributions are summarized below:

- We propose a novel framework, i.e., TPWNG, to perform pseudo-label generation and self-training for WSVAD. TPWNG fine-tunes CLIP with the designed ranking loss and distributional inconsistency loss to transfer its strong text-image alignment capability to assist pseudo-label generation by means of the PLG module.
- We design a learnable text prompt and normality visual prompt mechanism to further improve the alignment accuracy of video events description text and video frames.
- We introduce a TCSAL module to learn the temporal dependencies of different video events more flexibly and accurately. To the best of our knowledge, we are the first to introduce the idea of self-adaptive learning of temporal context dependencies for VAD.
- Extensive experiments have been conducted on two benchmark datasets, UCF-Crime and XD-Violence, where the excellent performance demonstrates the effectiveness of our method.

2. Related Work

2.1. Video Anomaly Detection

The VAD task has been widely focused and researched, and many methods have been proposed to solve this problem. According to different supervision modes, these methods can be mainly categorized into semi-supervised-based and weakly supervised-based VAD.

Semi-supervised VAD. Early researchers mainly used

semi-supervised approaches to solve the VAD problem [2, 7, 8, 10, 14, 15, 20, 24, 31, 33, 41–44, 46, 50]. In the semi-supervised setting, only normal data can be acquired in the training phase, which aims to build a model that can characterize normal behavioral patterns by learning normal data. During the testing phase, data that contradict with the normal patterns are considered anomalies. Common semi-supervised VAD methods mainly include one-class classifier-based [21, 33, 37] and reconstruction [8, 38] or prediction errors-based methods [14, 42]. For example, Xu et al. [38] used multiple one-classifiers to predict anomaly scores based on appearance and motion features. Hasan et al. [8] built a fully convolutional auto-encoder to learn regular patterns in the video. Liu et al. in [14] proposed a novel video anomaly detection method that utilizes the U-Net architecture to predict future frames, where frames with large prediction errors are considered as anomalous.

Weakly Supervised VAD. Compared to semi-supervised VAD methods, WSVAD can utilize both normal and anomalous data with video-level labels in the training phase, but the exact frame location where the abnormal event occurred is unknown. In such a setting, the one-stage approaches based on MIL [3–5, 13, 17, 22, 26, 27, 32, 34, 45, 54] and the two-stage approaches based on pseudo-labels self-training [6, 11, 51, 53] are the two prevailing approaches. For example, Sultani et al. [26] first proposed a deep MIL ranking framework for VAD, where they considered anomalous and normal videos as positive and negative bags, respectively, and the snippets in the videos are considered as instances. Then a ranking loss is used to constrain the snippets with the highest anomaly scores in the positive and negative bags to stay away from each other. Later, many variants of the method were proposed on this basis. For example, Tian et al. [27] proposed a top-k MIL based VAD method with robust temporal feature magnitude learning.

However, these one-stage methods generally use a MIL framework, which leads to models that tend to focus only on the most significant anomalous snippets while ignoring nontrivial anomalous snippets. A two-stage approach based on pseudo-label self-training provides a relatively more promising solution. The two-stage approach first generates initial pseudo-labels using MIL or an off-the-shelf classifier and then refines the labels before using them for supervised training of the classifier. For example, Zhong et al. in [53] reformulated the WSVAD problem as a supervised learning task under noisy labels obtained by an off-the-shelf video classifier. Feng et al. in [6] introduced a multiple instance pseudo label generator that produces more reliable pseudo labels for fine-tuning a task-specific feature encoder with self-training mechanism. Zhang et al. in [51] exploited completeness and uncertainty properties to enhance pseudo labels for effective self-training. How-

ever, all these existing methods only generate pseudo-labels based on visual unimodal information and lack the utilization of textual modal. Therefore, in this paper, we endeavor to combine both visual and textual modal information in order to generate more accurate and complete pseudo-labels for self-training of the classifier.

2.2. Large Vision-Language Models

Recently, there has been an emergence of large vision-language models that learn the interconnections between visual and textual modalities by pre-training on large-scale datasets. Among these methods, the CLIP demonstrates unprecedented performance in many visual-language downstream tasks, e.g. image classification [55], object detection [56], semantic segmentation [12] and so on. The CLIP model has recently been successfully extended to the video domain as well. VideoCLIP [39] is proposed to align video and textual representations by contrasting temporally overlapping video-text pairs with mined hard negatives. Action-CLIP [30] formulated the action recognition task as a multimodal learning problem rather than a traditional unimodal classification task. However, there are fewer attempts to utilize CLIP models to solve VAD tasks. Joo et al. in [9] simply utilizes CLIP's image encoder for extracting more discriminative visual features and does not use textual information. Wu et al. [36], Zanella et al. [48] mainly use textual features from CLIP to enhance the expressiveness of the overall features, followed by MIL-based anomaly classifier learning. The major difference with the above works is that our method is the first to utilize the textual features encoded by the CLIP text encoder in conjunction with the visual features to generate pseudo-labels, and then employ a supervised approach to train an anomaly classifier.

3. Method

In this section, we first present the definition of the WSVAD task, then introduce the overall architecture of our proposed method, and subsequently elaborate on the details of each module and the execution process.

3.1. Overall Architecture

Formally, we first define sets $D^a = \{(v_i^a, y_i)\}_{i=1}^{M}$ and $D^n = \{(v_i^n, y_i)\}_{i=1}^{M}$ containing M abnormal and normal videos with ground-truth labels, respectively. For each v_i^a, it is labeled $y_i = 1$, indicating that this video contains at least one anomalous video frame, but the exact location of the anomalous frame is unknown. For each v_i^n, it is labeled $y_i = 0$, indicating that this video consists entirely of normal frames. With this setting, WSVAD task is to utilize coarse-grained video-level labels to enable a classifier to learn to predict fine-grained frame-level anomaly scores.

Fig. 2 illustrates the overall pipeline of our approach. Normal and abnormal video along with learnable category
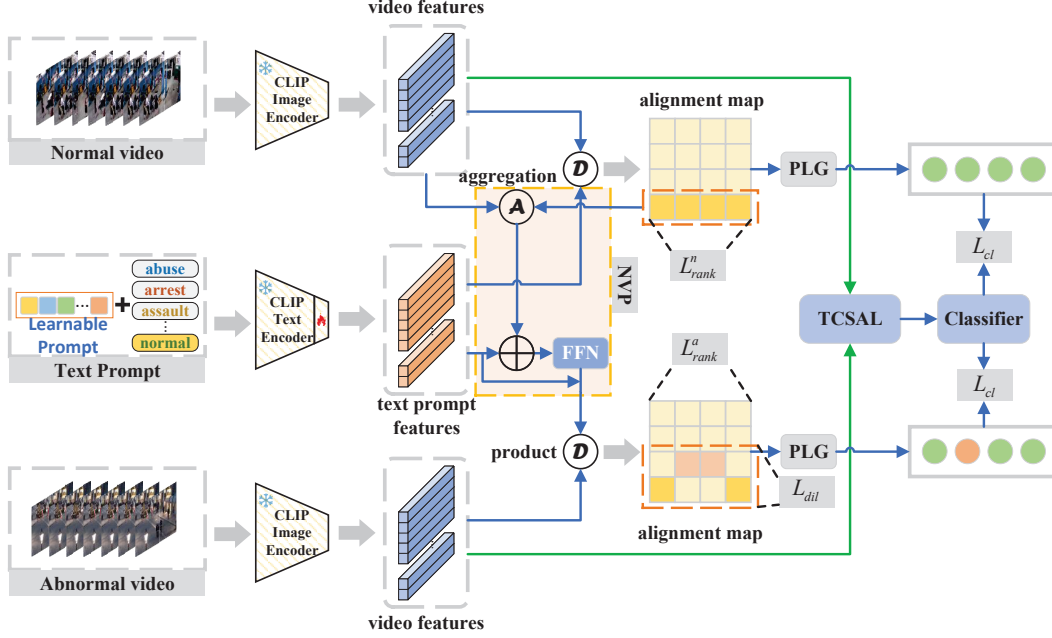
Figure 2. The overall architecture of our proposed TPWNG.

prompt text are encoded as feature embedding by the image encoder and text encoder of CLIP, respectively. Then, the text encoder of CLIP is encouraged by fine-tuning it to produce textual feature embedding of video event categories that accurately match anomalous or normal video frames, and the NVP assists in this process. Meanwhile, the image features feed the TCSAL module to perform self-adaptive learning of temporal dependencies. Finally, a video frame classifier is trained to predict anomaly scores under the supervision of pseudo-labels obtained by the PLG module.

3.2. Text and Normality Visual Prompt

Learnable Text Prompt. Constructing textual prompts that can accurately describe various video event categories is a prerequisite for realizing the alignment of text and corresponding video frames. However, it is impractical to manually define description texts that can completely characterize anomalous events in all different scenarios. Therefore, inspired by CoOp [55], we employ a learnable text prompt mechanism to adaptively learn representative video event text prompts to align the corresponding video frames. Specifically, we construct a learnable prompt template, which adds l learnable prompt vectors in front of the tokenized category name, as follows:

$$p_{label} = (\partial_1, \ ..., \ \partial_l, \ Tokenizer(label)), \qquad (1)$$

where ∂_l denotes the $l-$th prompt vector. Tokenizer is converting original category labels, i.e., "fighting", "accident", ..., "normal", etc., into class tokens by means of

CLIP tokenizer. Then, we add the corresponding location information pos to the learnable prompts and then feed it to the CLIP text encoder ζ_{text} to get the feature embedding $T_{label} \in \mathbb{R}^D$ of the video event description text as follows:

$$T_{label} = \zeta_{text}(p_{label} \oplus \ pos), \qquad (2)$$

Finally, we compute all video event categories according to Eqs. (1) and (2) to obtain the video event description text embedding set $E = \{T_1^a, \ T_2^a, \ ..., \ T_{k-1}^a, \ T_k^n\}$, where $\{T_i^a\}_{i=1}^{k-1}$ denotes the description text embedding of preceding $k-1$ abnormal events and T_k^n denotes the description text embedding of normal events.

Normality Visual Prompt. For an anomalous video, which contains both anomalous and normal frames, our core task is to infer pseudo-labels from the match similarities between the description text of the anomalous events and the video frames. However, this process is susceptible to interference from normal frames in the anomalous video because they have a similar background to the anomalous frames. To minimize this interference, we propose a NVP mechanism. NVP is used to assist the normal event description text to more accurately align normal frames in the abnormal video, and thus indirectly assist the description text of abnormal event to align abnormal video frames in the abnormal video by means of the distribution inconsistency loss that will be introduced in Sec. 3.5. Specifically, we first compute the match similarities $S_{i, k}^{nn} \in \mathbb{R}^F$ between the description text embedding of normal event and the video frame features in the normal video. Then, the match similarities after softmax operation are used as weights to aggre-

gate normal video frame features to obtain NVP $Q_i \in \mathbb{R}^D$. The formulas are represented as follows:

$$S_{i,k}^{nn} = X_i^n (T_k^n)^\top, Q_i = softmax((S_{i,k}^{nn})^\top) X_i^n, \quad (3)$$

where $X_i^n \in \mathbb{R}^{F \times D}$ denotes the visual features of the normal video v_i^n obtained by the CLIP image encoder, where F and D denote the number of video frames and feature dimensions, respectively. Then, we concatenate Q_i and T_k^n in the feature dimension and feed an FFN layer with skip connections to obtain the enhanced description text embedding $\dot{T}_k^n$ of normal events. The formula is represented as follows:

$$\dot{T}_k^n = FFN((T_k^n \cup Q_i)) + T_k^n. \quad (4)$$

3.3. Pseudo Label Generation Module

In this subsection, we detail how to generate frame-level pseudo labels. For a normal video, we can directly get the frame-level pseudo-labels, i.e., for a $v_i^n = \{I_j\}_{j=1}^F$ containing F normal frames, it corresponds to a label set $\{\gamma_{i,j}^n = 0\}_{j=1}^F$. Our main goal is to infer the pseudo-labels for anomalous videos that contain both anomalous and normal frames. To this end, we propose a PLG module for inferring accurate pseudo-labels based on the normality guidance. PLG module infers frame-level pseudo-labels by incorporating the match similarities between the description text of the normal event and the abnormal video as a guide into the match similarities between the description text of the corresponding abnormal event and the abnormal video.

Specifically, we first compute the match similarities $S_{i,k}^{an} = X_i^a (\dot{T}_k^n)^\top$ between normal event description text embedding enhanced with NVP and anomalous video features, where $X_i^a \in \mathbb{R}^{F \times D}$ denotes the visual features of the anomalous video v_i^a obtained by the CLIP image encoder. Similarly, we compute the match similarities $S_{i,\tau}^{aa} = X_i^a (T_\tau^a)^\top$ between the description text embedding T_τ^a of the corresponding τ-th ($1 \leqslant \tau \leqslant k-1$) real anomaly category and the anomaly video features X_i^a.

Theoretically, for $S_{i,\tau}^{aa}$, it should have high match similarities corresponding to abnormal frames and low match similarities for normal frames. But it may be interfered by normal frames from the same video having the same background. To reduce the interference of normal frames, we infer pseudo-labels by incorporating the matching similarity corresponding to the description text of normal events with certain weights as a guide into the matching similarity of the description text of corresponding real abnormal events. Specifically, we first perform a normalization and fusion operation on $S_{i,\tau}^{aa}$ and $S_{i,k}^{an}$ as follows:

$$\psi_i = \alpha \tilde{S}_{i,k}^{an} + (1-\alpha)(1 - \tilde{S}_{i,\tau}^{aa}), \quad (5)$$

where $\tilde{*}$ denotes the normalization operation and α denotes the guidance weight. After obtaining ψ_i, we similarly perform a normalization operation on it to obtain $\tilde{\psi}_i$. Then,

we set a threshold θ on $\tilde{\psi}_i$ to obtain the frame-level pseudo-labels in the anomalous video as follows:

$$\gamma_{i,j}^a = \begin{cases} 1, & \tilde{\psi}_{i,j} \geq \theta; \\ 0, & \tilde{\psi}_{i,j} < \theta, \end{cases} \quad i = 1,2,...,M; \; j = 1,2,...,F \quad (6)$$

where $\gamma_{i,j}^a$ denotes the pseudo-label of the j-th frame in the i-th anomaly video. Finally, we combine the frame-level pseudo-labels $\gamma_{i,j}^n$ and $\gamma_{i,j}^a$ of normal and anomalous videos to get the total pseudo-label set $\{\gamma_{i,j}\}_{j=1}^F$.

3.4. Temporal Context Self-adaptive Learning

To adaptively adjust the learning range of temporal relationship based on the input video data, inspired by the work [25], we introduce a TCSAL module. The backbone of TCSAL is the transformer-encoder, but unlike the original transformer, the spanning range of attention is controlled by a soft mask function χ_z for each self-attention head at each layer. χ_z is a piecewise function mapping a distance to a value between [0, 1] as follows:

$$\chi_z(h) = \min\left[\max\left[\frac{1}{R}(R + z - h), 0\right], 1\right], \quad (7)$$

where h represents the distance between the current t-th frame in a video and the $r-$th ($r \in [1, t-1]$) frame in the past temporal range. R is a hyperparameter used to control the softness. z is a learnable parameter that is adaptively tuned with the input as follows:

$$z = F\sigma(C^\top X + b), \quad (8)$$

here σ represents the sigmoid operation, C and b are learnable parameters during model training. With the soft mask function χ_z, the corresponding attention weights $\omega_{t,r}$ is computed within this mask, i.e.,

$$\omega_{t,r} = \frac{\chi_z(t-r) \exp(\beta_{t,r})}{\sum_{q=1}^{t-1} \chi_z(t-q) \exp(\beta_{t,q})}, \quad (9)$$

here $\beta_{t,r}$ denotes the dot product output of the *Query* corresponding to the t-th frame in a video with the *Key* corresponding to the $r-$th frame in the past. Under the control of χ_z, the self-attention heads will be able to adaptively adjust the self-attention span range according to the input.

Finally, the video features after temporal context adaptive learning are fed into a classifier to predict the frame-level abnormality scores $\{\eta_{i,j}\}_{j=1}^F$.

3.5. Objective Function

First, we fine-tune the CLIP text encoder. For a normal video, we further compute the match similarities set $\varphi_i^{na} = \{S_{i,\tau}^{na} = X_i^n (T_\tau^a)^\top \mid 1 \leqslant \tau \leqslant k-1\}$ between the description texts of the other $k-1$ anomalous events and the normal frames. We expect that the maximum in the similarity set

φ_i^{na} should be as small as possible while the maximum in $S_{i,k}^{nn}$ should be as large as possible. Thus, we design the following ranking loss for constraints:

$$L_{rank}^n = \max(0,\ 1-\max(S_{i,k}^{nn})+\max(\max(\varphi_i^{na})). \quad (10)$$

For an anomalous video, we first calculate the similarities $S_{i,k}^{an} = X_i^a(\dot{T}_k^n)^\top$ between the description text embedding of normal event and the anomalous video features, the similarity $S_{i,\tau}^{aa} = X_i^a(T_\tau^a)^\top$ between the description text embedding of the τ-th $(1 \leqslant \tau \leqslant k-1)$ real anomalous event category and the anomalous video features, and the similarity set $\varphi_i^{aa} = \{S_{i,g}^{aa} = X_i^a(T_g^a)^\top \,|1 \leqslant g \leqslant k-1,\ g \neq \tau\ \}$ between the description text embedding of other $k-2$ anomalous event categories and the anomalous video features, respectively. We expect that the maximum value in $S_{i,k}^{an}$ should be greater than the maximum value in φ_i^{aa}. Similarly, the maximum value in $S_{i,\tau}^{aa}$ should be greater than the maximum value in φ_i^{aa}. In short, it means that we expect that the description texts of real abnormal and normal events should match the abnormal and normal frames in the abnormal video with the highest possible similarity, respectively. Thus, the ranking loss for anomalous videos is designed as follows:

$$L_{rank}^a = \max(0,\ 1 - \max(S_{i,k}^{an}) + \max(\max(\varphi_i^{aa}))) + \\ \max(0,\ 1 - \max(S_{i,\tau}^{aa}) + \max(\max(\varphi_i^{aa}))). \quad (11)$$

In addition, to further ensure that the description texts of real abnormal events and normal events can accurately align the abnormal and normal video frames in the abnormal video, respectively, we design a distribution inconsistency loss (DIL). DIL is used to constrain the similarities between the description text of the real abnormal event and the video frames to be inconsistent with the similarity distribution between the description text of the normal event and the video frames. We use cosine similarity to perform this loss:

$$L_{dil} = \frac{1}{MF} \sum_{i=1}^{M} \sum_{j=1}^{F} \frac{\tilde{S}_{i,j,\tau}^{aa} \cdot \tilde{S}_{i,j,k}^{an}}{\left\| \tilde{S}_{i,j,\tau}^{aa} \right\|_2 \cdot \left\| \tilde{S}_{i,j,k}^{an} \right\|_2}. \quad (12)$$

Then, following the work [26], in order to make the generated pseudo-labels satisfy sparsity and smoothing in temporal order, we impose sparsity and smoothing constraints, $L_{sp} = \sum_{j=1}^{F}(\tilde{S}_{i,j,\tau}^{aa} - \tilde{S}_{i,j+1,\tau}^{aa})^2$, $L_{sm} = \sum_{j=1}^{F} \tilde{S}_{i,j,\tau}^{aa}$, on the similarity vectors $\tilde{S}_{i,\tau}^{aa}$.

Then, we calculate the binary cross-entropy between the anomaly score $\eta_{i,j}$ predicted by the classifier and the pseudo-label $\gamma_{i,j}$ as the classification loss:

$$L_{cl} = -\frac{1}{MF} \sum_{i=1}^{M} \sum_{j=1}^{F} [\eta_{i,j} \log(\gamma_{i,j}) \\ + (1 - \eta_{i,j}) \log(1 - \gamma_{i,j})]. \quad (13)$$

The final overall objective function balanced by λ_1 and λ_2 is designed as follows:

$$L_{all} = L_{rank}^n + L_{rank}^a + L_{dil} + L_{cl} + \lambda_1 L_{sp} + \lambda_2 L_{sm}. \quad (14)$$

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. We conduct extensive experiments on two benchmark datasets, UCF-Crime [26] and XD-Violence [34]. **UCF-Crime** is a large-scale real scene dataset for WSVAD. UCF-Crime duration is 128 hours in total and contains 1900 surveillance videos covering 13 anomaly event categories, of which 1610 videos with video-level labels are used for training and 290 videos with frame-level labels are used for testing. **XD-Violence** is a large-scale violence detection dataset collected from movies, online videos, surveillance videos, CCTVS, etc. XD-Violence lasts 217 hours and contains 4754 videos covering 6 anomaly event categories, of which 3954 training videos with video-level labels and 800 test videos with frame-level labels.

Evaluation Metrics. Following the previous methods [6, 26], for the UCF-Crime dataset, we measure the performance of our method using the area under the curve (AUC) of the frame-level receiver operating characteristics (ROC). Similarly, for the XD-Violence dataset, we follow the evaluation criterion of average precision (AP) suggested by the work [34] to measure the effectiveness of our method.

4.2. Implementation Details

The image and text encoders in our method use a pre-trained CLIP (VIT-B/16), in which both the image and text encoders are kept frozen, except for the text encoder where the final projection layer is unfrozen for fine-tuning. The feature dimension D is 512. FFN is a standard block from Transformer. The length l of the learnable sequence in the text prompt is set to 8. The normality guidance weight α is set to 0.2 for both the UCF-Crime and XD-Violence datasets. The pseudo-labels generation threshold θ is set to 0.55 and 0.35 for the UCF-Crime and XD-Violence datasets, respectively. The parameter R used to control the softness of the soft mask function is set to 256. The sparse loss and smoothing loss weights are set to $\lambda_1 = 0.1$ and $\lambda_2 = 0.01$. Please refer to the supplementary materials for more details on implementation.

4.3. Comparison with State-of-the-art Methods

We compare the performance on the UCF-Crime and XD-Violence datasets with the current state-of-the-art (SOTA) methods in Tab. 1. As can be observed from the table, our method achieves a new SOTA on both the UCF-Crime and XD-Violence datasets. Specifically, for the UCF-Crime dataset, our method outperforms the current SOTA method

	Methods	UCF (AUC)	XD (AP)
	Sultani et al.[26]	77.92%	73.20%
	GCN [53]	82.12%	-
	HL-Net [34]	82.44%	73.67%
	CLAWS [45]	82.30%	-
	MIST [6]	82.30%	-
	RTFM [27]	84.30%	77.81%
Weakly	CRFD [32]	84.89%	75.90%
	GCL [47]	79.84%	-
	MSL [11]	85.62%	78.58%
	MGFN [3]	86.67%	80.11%
	Zhang et al.[51]	86.22%	78.74%
	UR-DMU [54]	86.97%	81.66%
	CLIP-TSA [9]	87.58%	82.17%
	Ours	**87.79%**	**83.68%**

Table 1. AUC and AP on UCF-Crime and XD-Violence dataset.

	UCF-Crime (AUC)	XD-Violence (AP)
w/o NVP	85.25%	81.58%
w NVP-FA	87.32%	83.13%
w NVP-AS	87.79%	83.68%

Table 2. The AUC and AP of our method on the UCF and XD datasets without NVP, with NVP-FA, and with NVP-AS.

	UCF-Crime (AUC)	XD-Violence (AP)
w/o NG	85.83%	81.32%
w NG	87.79%	83.68%

Table 3. The AUC and AP of our method on the UCF and XD datasets with NG and without NG.

CLIP-TSA [9] by 0.21%, which is not a trivial improvement for the challenging WSVAD task. Most importantly, compared to methods MIST [6] and Zhang et al. [51] similar to ours that also use pseudo-label-based self-training, our method significantly outperforms them by 5.49% and 1.57%, respectively. This fully demonstrates that our proposed pseudo-label generation and self-training framework is vastly superior to the above two approaches. This also indicates that transferring visual language multimodal associations through CLIP is conducive to generating more accurate pseudo-labels compared to merely utilizing unimodal visual information. For the XD-Violence dataset, our method also surpasses the current optimal method CLIP-TSA [9] by 1.52%. Compared to a similar pseudo-label-based self-training method Zhang et al. [51], our method also outperforms it by 4.94%. The consistent superior performance on two large-scale real datasets strongly demonstrates the effectiveness of our method. This also shows the extraordinary potential of the pseudo-label based self-training scheme, if accurate pseudo-labels can be generated utilizing multiple modality information.

4.4. Ablation Studies

We conduct ablation experiments in this subsection to analyze the effectiveness of each component of our framework.

Effectiveness of Normal Visual Prompt. To verify the validity of NVP, we execute three comparison experiments: without NVP, with NVP based on frame averaging (NVP-FA), and with NVP based on match similarities aggregation (NVP-AS). As can be seen from the results in Tab. 2, in the absence of NVP, the performance of our method on the UCF-Crime and XD-Violence datasets decreases by 2.54% and 2.10% compared to with an NVP-AS, respectively. NVP-AS boosts the performance of the method by 0.47% and 0.55% more compared to NVP-FA on UCF-Crime and

XD-Violence datasets, respectively. This reveals two facts: first, NVP can help the text embedding to better match normal frames in anomalous videos, which indirectly aids in generating more accurate pseudo-labels in cooperation with the DIL and the normality guidance mechanism. Second, the NVP-AS can effectively reduce the interference of some noise snippets (e.g., prologue, perspective switching, etc.) in normal videos compared to the NVP-FA approach, thus obtaining a purer NVP.

Effectiveness of the Normality Guidance. In the pseudo-label generation module, instead of inferring pseudo-labels directly based on the similarity between the corresponding abnormal event description text and the abnormal video, we incorporate guidance from the match similarities of the normal event description text counterparts, aiming to reduce the interference of partially noisy video frames and generate more accurate pseudo-labels. To verify the contribution of the normality guidance, we compare the impact of the pseudo-label generation module on the performance of our method with and without normal guidance (NG), respectively. As can be observed from Tab. 3, when our method is equipped with normal guidance, the performance rises by 1.96% and 2.36% on the UCF-Crime and XD-Violence datasets, respectively. This validates the effectiveness of the normality guidance.

Effectiveness of TCSAL. To analyze the effectiveness of TCSAL module, we conduct comparative experiments with the Transformer-encoder (TF-encoder) module in [28], MTN module in [27], and GL-MHSA module in [54] by replacing the temporal learning module in our framework with each of these three modules. From Tab. 4, it can be observed that the TF-encoder module has the lowest performance, which is understandable since the global self-attention computation way makes it neglect to pay attention to the local temporal information. Both MTN and GL-MHSA outperform TF-encoder with comparable performance. Our introduced TCSAL module achieved the best performance on both datasets. This indicates that adopting
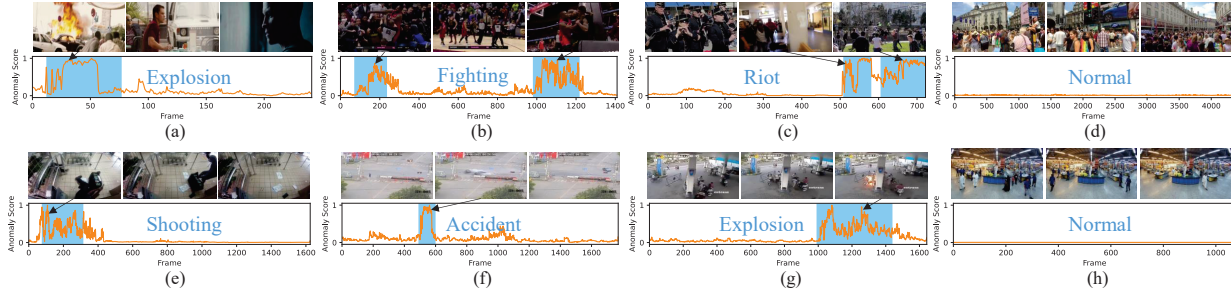
Figure 3. Anomaly score curves of several test samples on the UCF-Crime and XD-Violence dataset.

	UCF (AUC)	XD (AP)
w TF-encoder	85.12%	80.02%
w MTN	86.22%	81.02%
w GL-MHSA	86.43%	81.23%
w TCSAL	87.79%	83.68%

Table 4. The AUC and AP of our method on the UCF and XD datasets with different temporal modules.

Loss term				Dataset	
bs	L_{rank}^n	L_{rank}^a	L_{dil}	UCF (AUC)	XD (AP)
✓				77.12%	73.32%
✓	✓			81.34%	78.67%
✓		✓		84.45%	81.56%
✓			✓	82.47%	79.96%
✓	✓	✓	✓	87.79%	83.68%

Table 5. Comparison of the AUC and AP of our method with different loss terms on the UCF-Crime and XD-Violence datasets. "bs" indicates that L_{cl}, L_{sp}, L_{sm} three loss functions are used.

the mechanism of self-attention span range adaptive learning enables the temporal learning module to self-adapt to the inputs of videos with different event lengths, achieving more accurate modeling of temporal dependencies while weakening the interference of other non-relevant temporal information in the non-event span range.

4.5. Qualitative Results

We show the anomalous scores of our method on several test videos in Fig. 3. It can be obviously noticed that there is a steep rise in the anomaly scores when various anomalous events occur, and as the anomalous events end, the anomaly scores fall back to the lower range rapidly. For normal events, our method gives a lower abnormal score. This intuitively demonstrates that our method has good sensitivity to abnormal events and can accurately and timely detect the occurrence of abnormal events while maintaining a low abnormal score prediction for normal events.

4.6. Analysis of Losses

To analyze the impact of the three loss functions L_{rank}^n, L_{rank}^a, and L_{dil}, we perform ablation experiments on the UCF-Crime and XD-Violence datasets. As shown in Tab. 5, when all three loss functions are absent, the performance of our method is unsatisfactory. This reveals that the original CLIP suffers from domain bias and is not directly applicable to the VAD domain. When three loss functions are available individually, the performance of our method is clearly improved, where the L_{rank}^a gives the biggest boost to the performance. When all three losses are combined and cooperate with each other, our method achieves the best performance. This demonstrates the effectiveness of the three loss functions we have designed, and they can effectively assist CLIP in domain adaptation for WSVAD.

5. Conclusions

In this paper, we propose a novel framework, TPWNG, to perform pseudo-label generation and self-training for WS-VAD. TPWNG finetunes CLIP with the designed ranking loss and distributional inconsistency loss to transfer its text-image alignment capability to assist pseudo-label generation with the PLG module. Further, we design a learnable text prompt and normality visual prompt mechanisms to further improve the alignment accuracy of video events description text and video frames. Finally, we introduce a TC-SAL module to learn the temporal dependencies of different video events more flexibly and accurately. We perform extensive experiments on the UCF-Crime and XD-Violence datasets, and the superior performance compared to existing methods demonstrates the effectiveness of our method.

6. Acknowledgments

References

[1] Christophe Bertero, Matthieu Roy, Carla Sauvanaud, and Gilles Trédan. Experience report: Log mining using natural language processing and application to anomaly detection. In *ISSRE*, pages 351–360, 2017. 1

[2] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. Appearance-motion memory consistency network for video anomaly detection. In *AAAI*, pages 938–946, 2021. 3

[3] 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 *AAAI*, pages 387–395, 2023. 3, 7

[4] 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 *CVPR*, pages 12137–12146, 2023.

[5] 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 *CVPR*, pages 12137–12146, 2023. 3

[6] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *CVPR*, pages 14009–14018, 2021. 1, 3, 6, 7

[7] 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 *CVPR*, pages 1705–1714, 2019. 3

[8] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *CVPR*, pages 733–742, 2016. 3

[9] Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *ICIP*, pages 3230–3234, 2023. 3, 7

[10] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Bman: Bidirectional multi-scale aggregation networks for abnormal event detection. *IEEE TIP*, 29:2395–2408, 2019. 3

[11] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *AAAI*, pages 1395–1403, 2022. 1, 3, 7

[12] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *CVPR*, pages 15305–15314, 2023. 3

[13] Tianshan Liu, Kin-Man Lam, and Jun Kong. Distilling privileged knowledge for anomalous event detection from weakly labeled videos. *IEEE TNNLS*, pages 1–15, 2023. 3

[14] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *CVPR*, pages 6536–6545, 2018. 1, 3

[15] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *ICCV*, pages 341–349, 2017. 3

[16] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *CVPR*, pages 15425–15434, 2021. 1

[17] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE TIP*, 30:4505–4515, 2021. 1, 3

[18] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *CVPR*, pages 14372–14381, 2020. 1

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2

[20] Mohammad Sabokrou, Mahmood Fathy, Mojtaba Hoseini, and Reinhard Klette. Real-time anomaly detection and localization in crowded scenes. In *CVPR*, pages 56–62, 2015. 3

[21] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *CVPR*, pages 3379–3388, 2018. 3

[22] Hitesh Sapkota and Qi Yu. Bayesian nonparametric submodular video partition for robust anomaly detection. In *CVPR*, pages 3212–3221, 2022. 3

[23] Fangtao Shao, Jing Liu, Peng Wu, Zhiwei Yang, and Zhaoyang Wu. Exploiting foreground and background separation for prohibited item detection in overlapping x-ray images. *PR*, 122:108261, 2022. 1

[24] Giulia Slavic, Abrham Shiferaw Alemaw, Lucio Marcenaro, David Martin Gomez, and Carlo Regazzoni. A kalman variational autoencoder model assisted by odometric clustering for video frame prediction and anomaly detection. *IEEE TIP*, 32:415–429, 2022. 3

[25] Sainbayar Sukhbaatar, Édouard Grave, Piotr Bojanowski, and Armand Joulin. Adaptive attention span in transformers. In *ACL*, pages 331–335, 2019. 2, 5

[26] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, pages 6479–6488, 2018. 1, 3, 6, 7

[27] 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 *ICCV*, pages 4975–4986, 2021. 1, 3, 7

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 7

[29] Chao Wang, Jing Liu, Kai Wu, and Zhaoyang Wu. Solving multitask optimization problems with adaptive knowledge transfer via anomaly detection. *IEEE TEC*, 26(2):304–318, 2021. 1

[30] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 3

[31] Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. Robust unsupervised video anomaly detection by multipath frame prediction. *IEEE TNNLS*, 33(6):2301–2312, 2021. 3

[32] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE TIP*, 30:3513–3527, 2021. 3, 7

[33] Peng Wu, Jing Liu, and Fang Shen. A deep one-class neural network for anomalous event detection in complex scenes. *IEEE TNNLS*, 31(7):2609–2622, 2019. 3

[34] 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 *ECCV*, pages 322–339, 2020. 3, 6, 7

[35] Peng Wu, Jing Liu, Xiangteng He, Yuxin Peng, Peng Wang, and Yanning Zhang. Towards video anomaly retrieval from video anomaly detection: New benchmarks and model. *arXiv preprint arXiv:2307.12545*, 2023. 1

[36] 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. In *AAAI*, pages 6074–6082, 2024. 3

[37] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015. 3

[38] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *CVIU*, 156:117–127, 2017. 3

[39] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 3

[40] Minghui Yang, Jing Liu, Zhiwei Yang, and Zhaoyang Wu. Slsg: Industrial image anomaly detection by learning better feature embeddings and one-class classification. *arXiv preprint arXiv:2305.00398*, 2023. 1

[41] Zhiwei Yang, Jing Liu, and Peng Wu. Bidirectional retrospective generation adversarial network for anomaly detection in videos. *IEEE Access*, 9:107842–107857, 2021. 3

[42] Zhiwei Yang, Peng Wu, Jing Liu, and Xiaotao Liu. Dynamic local aggregation network with adaptive clusterer for anomaly detection. In *ECCV*, pages 404–421, 2022. 3

[43] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *CVPR*, pages 14592–14601, 2023. 1

[44] Muchao Ye, Xiaojiang Peng, Weihao Gan, Wei Wu, and Yu Qiao. Anopcn: Video anomaly detection via deep predictive coding network. In *ACM MM*, pages 1805–1813, 2019. 3

[45] 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 *ECCV*, pages 358–376, 2020. 3, 7

[46] Muhammad Zaigham Zaheer, Jin-Ha Lee, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Stabilizing adversarially learned one-class novelty detection using pseudo anomalies. *IEEE TIP*, 31:5963–5975, 2022. 3

[47] Muhammad Zaigham Zaheer, Arif Mahmood, Muhammad Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *CVPR*, pages 14744–14754, 2022. 7

[48] Luca Zanella, Benedetta Liberatori, Willi Menapace, Fabio Poiesi, Yiming Wang, and Elisa Ricci. Delving into clip latent space for video anomaly recognition. *arXiv preprint arXiv:2310.02835*, 2023. 3

[49] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draema discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, pages 8330–8339, 2021. 1

[50] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. In *ICML*, pages 1100–1109, 2016. 3

[51] 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 *CVPR*, pages 16271–16280, 2023. 1, 3, 7

[52] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *ICIP*, pages 4030–4034, 2019. 1

[53] 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 *CVPR*, pages 1237–1246, 2019. 1, 3, 7

[54] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *AAAI*, pages 3769–3777, 2023. 3, 7

[55] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 3, 4

[56] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, pages 350–368, 2022. 3