

Weakly-Supervised Video Anomaly Detection with Snippet Anomalous Attention

Yidan Fan, Yongxin Yu, Wenhuan Lu, Yahong Han

arXiv:2309.16309v1 [cs.CV] 28 Sep 2023

Abstract—With a focus on abnormal events contained within untrimmed videos, there is increasing interest among researchers in video anomaly detection. Among different video anomaly detection scenarios, weakly-supervised video anomaly detection poses a significant challenge as it lacks frame-wise labels during the training stage, only relying on video-level labels as coarse supervision. Previous methods have made attempts to either learn discriminative features in an end-to-end manner or employ a two-stage self-training strategy to generate snippet-level pseudo labels. However, both approaches have certain limitations. The former tends to overlook informative features at the snippet level, while the latter can be susceptible to noises. In this paper, we propose an Anomalous Attention mechanism for weakly-supervised anomaly detection to tackle the aforementioned problems. Our approach takes into account snippet-level encoded features without the supervision of pseudo labels. Specifically, our approach first generates snippet-level anomalous attention and then feeds it together with original anomaly scores into a Multi-branch Supervision Module. The module learns different areas of the video, including areas that are challenging to detect, and also assists the attention optimization. Experiments on benchmark datasets XD-Violence and UCF-Crime verify the effectiveness of our method. Besides, thanks to the proposed snippet-level attention, we obtain a more precise anomaly localization.

Index Terms—Weakly-supervised, Anomaly detection, Snippet anomalous attention, Multi-branch supervision.

I. INTRODUCTION

Video anomaly detection (VAD) is a crucial task in the analysis of activities in untrimmed videos, aiming to detect unusual events within video frames or snippets.

Unsupervised VAD has gained significant attention from researchers due to its ability to detect anomalies without requiring additional annotations [1]–[7], [33]. However, these approaches only have access to normal videos during the training phase, leading to a limited understanding of anomaly data. Consequently, unsupervised-VAD methods often exhibit a high false alarm rate for previously unseen normal events. In order to address the incorrect recognition of video anomalies in the unsupervised setting, a more practical scenario is considered where only the video level labels are available, i.e., weakly labeled abnormal or normal training videos. Addressing this scenario, the paper [8] first proposed weakly-supervised anomaly detection (WS-VAD). Compared to the unsupervised pipeline, the WS-VAD paradigm provides a better trade-off between detection performance and manual annotation cost.

This work was supported by the CAAI-Huawei MindSpore Open Fund. Yidan Fan, Yongxin Yu, Wenhuan Lu and Yahong Han are with the College of Intelligence and Computing, Tianjin University, Tianjin 300072, China. E-mail: {yidan_fan, yyx, wenhuan, yahong}@tju.edu.cn.

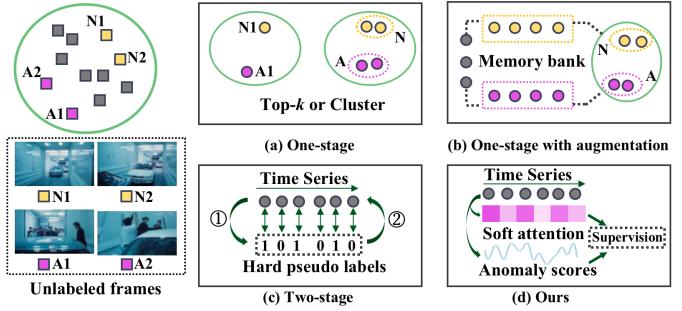


Fig. 1. Comparisons with the existing approaches. N refers to normal frames while A refers to abnormal. Frames “■” are firstly processed to snippet-level features “●”. In the one-stage method (a), normal and abnormal features are directly clustered or chosen with feature magnitude. In (b), additional memory banks are used to augment original features. While in (c), hard pseudo labels 0 or 1 are generated first and then used to direct snippet-level supervision. Labels are refined through the second stage. In our approach (d), snippet-level anomalous soft attention is generated first, and with general prediction scores, they are then fed into a multi-branch supervision module.

In the field of WS-VAD, a multitude of methods have been proposed to fully utilization of these weak annotations and existing approaches can be broadly categorized into two types based on the steps employed to generate final abnormal predictions: one-stage methods based on Multiple Instance Learning (MIL) [8]–[14], [27], [34], [35], and two-stage self-training methods [15]–[18]. In the case of one-stage methods, the key idea is to select representative abnormal and normal features, scores of these snippets are then used for final video-level classification. As shown in (a) of Figure 1, the top- k selection based on feature magnitude is applied in [11], [27] and in [22] a clustering distance-based loss is proposed to produce better anomaly representations. In [13], [14], additional memory modules are used to augment the original feature for the purpose of learning discriminative features, shown in (b). As for the two-stage approaches, in [15], [17], snippet-level pseudo labels are generated in stage 1 and then refined through the backpropagation process in step 2, which is shown in (c).

Although previous one-stage methods have achieved good performance, they still have limitations in snippet-level feature understanding. These methods tend to focus only on representative snippets, resulting in informative features being overlooked during the selection process which is biased and can then cause normal segments surrounding anomalies to be highly misclassified. Besides if the selected instance is wrong in the initial training stage, errors will accumulate and lead to poor predictions. Moreover, introducing a memory module

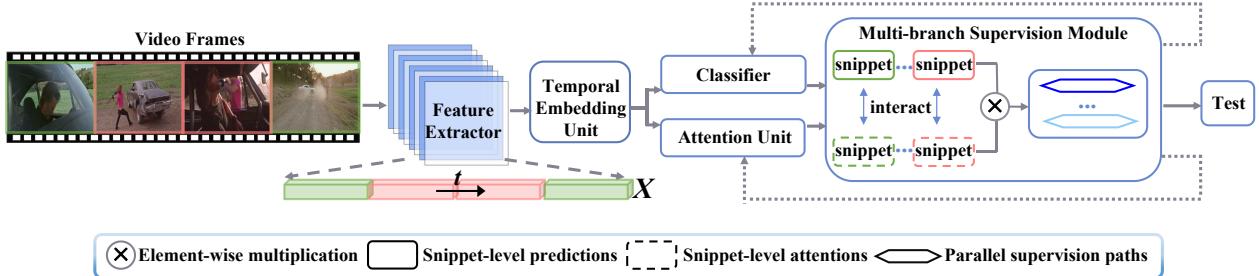


Fig. 2. The proposed method consists of three primary modules: Temporal Embedding Unit, Anomalous Attention Unit, and Multi-branch Supervision Module. The first module is responsible for encoding the feature, while the second module focuses on detecting snippet-level anomalies and generating attention. The third module aims to model the completeness of anomaly. To generate anomalous attention, an optimization process (dash line in the figure) is designed.

and updating it is inevitably time and resource-consuming. The two-stage methods try to tackle this problem with generated snippet-level pseudo labels. But pseudo labels are strong hints for supervision and contain much noise, thus also leading to unsatisfactory performance.

To address the aforementioned issues, our method first incorporates a temporal embedding unit to model the whole video, which aggregates both local and global information. Also, we adopt a soft attention mechanism to handle the WS-VAD task. Specifically, anomalous attention in the temporal dimension is generated to fully utilize intermediate snippet-level embeddings and guide the supervision process in a soft manner. Moreover, some anomalies are difficult to be distinguished due to their slight difference from normal events or only occupy a small portion of the frames, prompting us to propose a multi-branch supervision module, i.e., general supervision, attention-based supervision, general suppressed and attention-based suppressed supervision to explore the completeness of anomaly and detect difficult anomalous areas. As a result, a more robust anomaly-detected model can be obtained.

Previous methods for WS-VAD have incorporated attention mechanisms such as [9] and [10], but our approach is different from them in several ways. Specifically, both methods add attention branches with structures similar to the primary classifier branch while our attention unit is completely different from the classifier branch. In [9], the output of the attention branch tries to capture the total anomaly score of the entire video, while in [10], anomalous scores from both branches are simply averaged and used for final video classification. Unlike these methods, our attention mechanism is category-agnostic and anomaly-specific. Moreover, our attention measures anomalies in each snippet and is optimized by anomalous scores, rather than video-level labels.

In summary, the main contributions of this paper are as follows:

- We introduce a snippet-level attention mechanism using the intermediate embeddings from the consideration that they contain more semantic information and are beneficial for final frame-level¹ anomaly detection tasks. The attention is anomaly-specific and not optimized by video-level annotations but rather by anomaly predictions.

- With the assistance of anomalous attention in a soft manner, we propose a multi-branch supervision module to better explore the hard abnormal part of the whole video. Also, the completeness of the anomaly events and accuracy of the localization can be achieved.
- To validate our approach, we conducted experiments on two benchmark datasets: UCF-Crime and XD-Violence, and the state-of-the-art performance verify the effectiveness of our method.

II. RELATED WORK

A. Weakly-supervised Anomaly Detection

With limited labels, the objective of WS-VAD is to generate final frame-level anomalous scores. As mentioned before, previous approaches to WS-VAD can be broadly classified into two types: one-stage methods based on Multiple Instance Learning (MIL), and two-stage self-training methods.

Regarding one-stage methods, [8] is the first study to introduce the MIL frameworks to WS-VAD. Besides, in [8], hinge loss is employed, and anomaly scores of abnormal instances are enforced to be greater than the normal ones. Later, [11] notes that WS-VAD is biased by the dominant negative instances, especially when the abnormal events have subtle differences from normal events. Then they train a feature magnitude learning function to effectively recognize the positive instances. Then in [27], the author points out that feature magnitudes to represent the degree of anomalies typically ignore the effects of scene variations and thus propose a feature amplification mechanism and a magnitude contrastive loss to enhance the discriminativeness of features for detecting anomalies. Similar to the unsupervised anomaly detection methods, memory modules that can store the representative patterns are introduced in [13], [14]. The former encodes diverse normal features into prototypes and then constructs a similarity-based classifier. The latter uses two memory banks, one to store representative abnormal patterns and the other to store normal ones.

In the case of two-stage self-training methods, [15] introduced a multiple instance pseudo-label generator and a self-guided attention-boosted feature encoder to refine task-specific representations. In [16], a self-training strategy that gradually refines anomaly scores is proposed based on Multi-Sequence Learning (MSL). Furthermore, [17] presents a multi-

¹16 frames per snippet in our method like others.

head classification module and an iterative uncertainty pseudo-label refinement strategy.

Our approach to anomaly modeling differs from previous methods in two ways. Firstly, although our method also follows the one-stage MIL pipeline, it does not aim to choose representative features. Instead, the most discriminative part of the untrimmed video is suppressed in our approach. Secondly, from the perspective of modeling the anomaly completeness, although our approach has similar motivation with [17] which introduces a multi-head classification module, our multi-branch supervision module only utilizes one classifier and obtains diverse anomalous score sequences based on attention, thus effectively exploring anomaly completeness.

B. Weakly-supervised Temporal Action Localization

Weakly supervised temporal action localization is an efficient method for understanding human action [42] instances without overwhelming annotations. Several works [45], [46] have addressed this problem also using the multiple-instance learning (MIL) framework and primarily rely on aggregating class scores at the snippet level to generate video-level predictions, which is similar to the strategy generally used in the WS-VAD field. While others like [36]–[39], [41], [44] treat the background frames as an auxiliary class. Then utilizing the complementary learning scheme or filtering irrelevant information scheme at the snippet level to ensure precise positioning accuracy. Our work is mainly inspired by this snippet-level focus broadly used in WS-TAL tasks, but different from them in several ways. Firstly, any video in WS-TAL comprises both action frames and background frames, with background events typically regarded as an auxiliary class. Conversely, for WS-TAD, the normal subset only includes normal events. Secondly, WS-TAL is a multi-label classification task, the final result is the probability of each category, including the background class. On the other hand, WS-TAD is a regression task that generates precise anomaly scores as the final outcome. Lastly, in WS-TAL, the number of categories is known, and the given labels are exact action categories occurring in all videos, while for WS-TAD, the anomalies are varied and class-unknown. Therefore, despite our work drawing inspiration from WS-TAL, action temporal localization, and anomaly detection are significantly different tasks.

III. METHODS

In this paper, we propose a Snippet-level Anomalous Attention-based Multi-branch Supervision framework for WS-VAD task. The main structure is illustrated in Figure 2. The framework is composed of three core modules: the Temporal Embedding Unit for feature modeling, the Attention Unit for generating snippet-level anomalous attention, and the Multi-branch Supervision Module for learning anomaly completeness and improving localization accuracy. In the subsequent sections, we first give the formulation of the WS-VAD problem and then describe the three modules in detail. Finally, to generate precise anomaly attention, we present an optimization process and provide procedures for how to conduct the training and inference.

A. Problem Formulation

Following the MIL step, we formulate the WS-VAD problem as follows: let normal videos $V^n = \{v_i^n\}_{i=1}^N$ and abnormal videos $V^a = \{v_i^a\}_{i=1}^N$. Each anomaly video is a bag $Y_a = 1$, containing at least one abnormal instance, while normal videos are marked as $Y_n = 0$ with only normal instances. The objective of WS-VAD is to learn a function that can assign anomaly scores of snippets v_i for each video. To achieve this, we first extract features using pre-trained weights and then perform handling on the extracted features. In this paper, to ensure consistency with previous methods, we extract snippet-level appearance modality (RGB) features from non-overlapping video volumes containing 16 frames, using the I3D [29] network pre-trained on the Kinetics dataset [47] as the backbone. The features are 1024-dimensional for each snippet. For i -th video with T snippets, we represent the RGB features using matrix tensors $X_i^{RGB} \in R^{T*D}$ (abbr. $X \in R^{T*D}$), where D denotes the dimension of the feature vector.

B. Temporal Embedding Unit

Anomalies may occur in the short term or over a longer time, therefore both local and global temporal reliance should be considered in WS-VAD tasks. To address this issue, we introduce a temporal encoding unit with two branches: one for capturing local and the other for global dependencies.

Given the feature $F \in R^{T*D}$, in the global branch, we simply introduce the non-local block proposed in [19]:

$$F_g = \psi(F), \quad (1)$$

where ψ denotes the 1-D non-local operation and $F_g \in R^{T*\frac{D}{4}}$. As for the local branch, in order to acquire the different time scale local reliance, 1-D convolution operation with dilation (1, 2, 4) is separately used:

$$F_l = F_{l_2} = F_{l_3} = \phi(F), \quad (2)$$

where ϕ denotes the dilated convolution and $F_{l_3} \in R^{T*\frac{D}{4}}$.

Then F_l and F_g are concatenated in feature dimension and get $F^* \in R^{T*D}$. A temporal convolution layer is subsequently applied on F^* to aggregate features. Finally, with a residual connection, original feature F and F^* are simply fused by add operation and acquire enhanced feature F_e . Due to this temporal embedding unit is also widely used in other WS-VAD methods [11], [18], [40], thus we just briefly state and use it as our baseline in the ablation experiment.

C. Anomalous Attention Unit

The information surrounding a single snippet is crucial and can help to effectively detect anomalies at a more granular level. To address the problem of intermediate features not being fully utilized in the multiple instance learning (MIL) pipeline, we propose a snippet-level anomalous attention mechanism.

Specifically, after obtaining the enhanced feature, the Temporal Convolution layer TC is first adopted to fully capture channel-wise dependencies and infuse the local context from

the neighborhood snippets. Then to avoid some information not being activated in the whole training process and fully utilizing the semantic information, the LeakyRelu activated function LR is introduced for it can generate a negative value. Thus a basic attention unit can be formulated as:

$$F_e^{(l)} = (TC^{(l)}(F_e^{(l-1)}); LR), \quad (3)$$

where $F_e^{(l-1)}$ indicates the feature output from the $(l-1)^{th}$ basic unit and the whole attention unit is the stack of this basic unit. The feature dimension of the final TC layer is 1, which means $F \in R^{T \times 1}$. Then a sigmoid function is used to obtain normalized anomalous attention $A \in R^{T \times 1}$. This setting enables our method to use the attention normalization term to obtain highly confident snippets.

D. Multi-branch Supervision Module

Multi-Instance Learning is widely known [16], [17], [48] to suffer from numerous false alarms which are caused by the snippet-level detector to exhibit bias towards abnormal snippets with simple context. Thus intuitively, if reducing the focus on the most discriminative segments, we may effectively explore the completeness of anomalies and challenging snippets. Taking the account of the most discriminative segment, which may include crucial background information of the current video, our initial attempt is to assign lower attention to this discriminative segment and keep contextual information, however, it did not yield satisfactory results (as detailed in the experimental section). Then, we discovered that directly removing the most discriminative segment can satisfactorily enhance the overall detection performance. Thus with the enhanced features F_e and anomalous attention A , a multi-branch supervision module is designed, shown in Figure 3.

Original abnormal scores are directly obtained from the classifier with 3-layer MLP and the nodes are 512, 128, and 1 respectively. Also, each layer is followed by a ReLU and a dropout function. We denote the original anomalous scores as S^o . Then **Attention-based** abnormal scores S^a can be acquired by element-wise multiplication:

$$S^a = A * S^o, \quad (4)$$

with attention A utilized in S^a , only anomaly activity is considered and normal events have been suppressed.

Then for the purpose of avoiding the training process dominated by discriminative snippets and better learning the whole video, we calculate the **Suppressed original** abnormal scores S^{so} and **Suppressed Attention-based** abnormal scores S^{sa} . Concretely:

$$S_{ij}^{so} = \begin{cases} S_{ij}^o, & A_{ij} < \theta_i, \\ 0, & \text{otherwise}, \end{cases} \quad (5)$$

where $i \in (1, N)$ denotes the i^{th} video and $j \in (1, T)$ denotes the j^{th} snippets in current sequence. Besides, θ is a floating value based on the max and min value of the current A_i sequence:

$$\theta_i = [\max(A_i) - \min(A_i)] * \epsilon + \min(A_i), \quad (6)$$

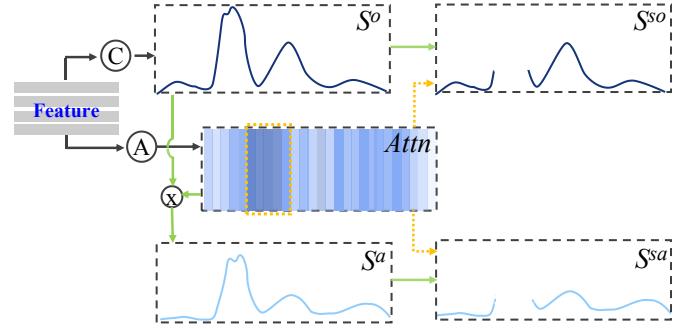


Fig. 3. Multi-branch Supervision Module. Features are fed into Classifier C and in the meantime directly utilized for snippet-level Anomalous Attention generation A . The most discriminative part will be suppressed (Orange) and hard abnormal snippets are given more focus.

and ϵ is a suppressed rate. The reason for processing θ in this manner is anomalies are diverse and some unobvious anomalies may gain a low anomalous attention. Thus using a fixed parameter as threshold is unreliable. Finally, for abnormal scores S^{sa} , it is handled in a similar manner to S^{so} :

$$S_{ij}^{sa} = \begin{cases} S_{ij}^a, & A_{ij} < \theta_i, \\ 0, & \text{otherwise}, \end{cases} \quad (7)$$

where the value of θ_i is equal to its value in Equation (6).

E. Optimizing Process

1) *Constraints on Attention*: We hope the attention is anomaly-specific, thus the distribution of the anomalous attention should be similar to the final anomalous scores. For negative bags with label $Y_n = 0$, the guide loss can be defined as:

$$L_{\text{guide}}^{\text{neg}} = \delta(A^{\text{neg}}, \{0 \dots 0\}), \quad (8)$$

where δ means a similarity metric function and we utilize mean square error (MSE). $\{0 \dots 0\}$ denotes a sequence that only consists of 0 and has the same size as A^{neg} .

Due to the lack of reliable predictions during the initial training phase, for the positive bag that the label is $Y_a = 1$ and contains both normal and abnormal instances, the guide loss is:

$$L_{\text{guide}}^{\text{pos}} = \begin{cases} \delta(A^{\text{pos}}, S_o^{\text{pos}}), & \text{if } \text{step} < M, \\ \delta(A^{\text{pos}}, \{0, 1, 1 \dots 1, 0\}), & \text{otherwise} \end{cases} \quad (9)$$

where M represents the number of training iterations and the sequence $\{0, 1, 1 \dots 1, 0\}$ can be obtained by:

$$1/0 = \begin{cases} 1, & \text{if } S_o^{\text{pos}} > 0.5, \\ 0, & \text{otherwise}, \end{cases} \quad (10)$$

Besides, due to anomaly is sparse, we also utilize a normalization loss L_{norm} to make the attention more polarized:

$$L_{\text{norm}} = \|A^{\text{pos}}\|_1, \quad (11)$$

where $\|\cdot\|_1$ is a L1-norm function.

2) *Constraints on Video-level Supervision*: We apply the widely used binary classification loss on S^o , S^a , S^{so} and S^{sa} :

$$L_c = \eta(S, Y), \quad (12)$$

where η is a binary cross-entropy loss and:

$$L_c^{all} = \alpha(L_c^o + L_c^a) + (1 - \alpha)(L_c^{so} + L_c^{sa}). \quad (13)$$

F. Network Training and Testing

1) *Training*.: The same amount of normal and abnormal videos are combined as a batch and fed into our model. Also, we incorporate the temporal smoothness and sparsity constraints term that are also commonly used in other WS-VAD methods and defined as:

$$\begin{aligned} L_{sm} &= \sum_{j=1}^{T-1} (S_j - S_{j+1})^2, \\ L_{sp} &= \sum_{j=1}^T S_j, \end{aligned} \quad (14)$$

where L_{sm} and L_{sp} are only applied on S^o and S^a branches. And the final loss for the training process is:

$$L = L_c^{all} + \gamma L_{norm} + L_{guide}^{neg} + L_{guide}^{pos} + \mu L_{sm} + L_{sp}. \quad (15)$$

2) *Testing*.: The testing videos are input into our network and the final predictions are calculated by:

$$S_{test} = S^a. \quad (16)$$

Finally, the snippet labels are assigned to the frame level.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

1) *Dataset*: Similar to previous approaches, we assess the performance of our method using two large datasets specifically designed for video anomaly detection: UCF-Crime [8] and XD-Violence [20].

UCF-Crime dataset is a large-scale dataset of 128 hours of videos. It consists of 1900 long and untrimmed real-world surveillance videos, with 13 realistic anomalies such as fighting, road accident, burglary, robbery, etc. as well as normal activities. The training set includes 800 normal and 810 abnormal videos with video-level labels, while the testing set has 140 normal and 150 abnormal videos, which are annotated at the frame-level.

XD-Violence. The XD-Violence is a more challenge dataset includes a total of 4754 videos collected from both movies and YouTube (in-the-wild scenes) with a variety of scenarios. There are 2405 violent videos and 2349 non-violent videos in the dataset. The training set has 3954 videos while the test set has 800 videos, consisting of 500 violent and 300 non-violent videos.

2) *Evaluation metrics*: Consistent with prior researches, we employ the area under the receiver operating characteristic curve (AUC) to evaluate the effectiveness of our method on the UCF-Crime dataset. Meanwhile, we adopt the precision-recall curve and the corresponding area under the curve (average precision, AP) to evaluate our approach on the XD-Violence dataset. Additionally, in the ablation study, we also report the AUC and AP results on the anomaly subset, following the approach outlined in [12], [15].

TABLE I
COMPARISON OF THE FRAME-LEVEL PERFORMANCE AUC ON UCF-CRIME TESTING SET WITH PREVIOUS METHODS. \dagger MEANS WE RE-TRAIN THE CODE WITH THE OPEN-SOURCE FEATURES.

Un.	Method		Feature (RGB)	AUC (%)
	GCL [2] S3R* [18]	CVPR 2022 ECCV 2022		
Weakly	GCN [25]	CVPR 2019	TSN [30]	82.12
	MIST [15]	CVPR 2021	I3D	82.30
	MSL [16]	AAAI 2022	I3D	85.30
	MSL [16]	AAAI 2022	VideoSwin [24]	85.62
	CU-Net [17]	CVPR 2023	I3D	86.22
	Sultani et al. [8]	CVPR 2018	I3D	76.21
	CLAWS [22]	ECCV 2020	C3D [31]	83.03
	RTFM [11]	ICCV 2021	I3D	84.30
	DAR [26]	TIFS 2022	I3D	85.18
	NG-MIL [14]	WACV 2023	I3D	85.63
One-stage	S3R [18]	ECCV 2022	I3D	85.99
	UR-DMU \dagger [13]	AAAI 2023	I3D	86.34
	UR-DMU [13]	AAAI 2023	I3D	86.97
	Ours (Pytorch)		I3D	86.19
	Ours (MindSpore)		I3D	84.94

B. Implementation Details

The whole model can run on a single RTX 2080Ti GPU. During the training process, we utilize the Adam optimizer [21] with a learning rate of 0.0001 and a weight decay of 5e-4. Our batch size is set at 32 with 4000 iterations. In equation (9), M is set to 400. The number of basic units in the Anomalous Attention Unit is 2. And the dropout rate of the MLP part is set as 0.7. We set the hyperparameters $(\epsilon, \alpha, \gamma, \mu)$ to $(0.2, 0.8, 0.8, 0.01)$ in our paper. For γ and μ which are the weights of the loss item, their values are set by a simple balance of the overall loss function without elaborately fine-tuning. The ablation of ϵ and α will be shown later. In addition to utilizing PyTorch, we also infer our method using Huawei's AI framework, Mindspore [50], on the UCF-Crime dataset.

C. Result on UCF-Crime

We evaluate the AUC performance of our method on the UCF-Crime dataset. Specifically, RGB features with a 10-crop augmentation are applied which is consistent with the existing approaches. Our method has shown a 1.89% improvement compared to the one-stage MIL-based method RTFM [11], and a 0.57% improvement compared to the two-stage self-training method MSL [16] based on 3D-transformer [24]. That is, our approach demonstrates superior performance due to its ability to preserve informative snippet-level features and employ a soft attention mechanism, surpassing both regular MIL-based models and even self-training methods. As for a typical UR-DMU [13] method, which introduces two additional memory modules to better store representative patterns, we re-train the code released by the author with the suggested feature (open source and utilized in our method) and get an 86.34% AUC value. It is important to note that AUC typically demonstrates optimistic results when dealing with class-imbalanced data, such as cases with numerous negative samples. In other words, in terms of AUC performance, on the UCF-Crime dataset which normal data accounts for a significant proportion of test videos, our method achieves a comparable result, with only a 0.15% difference. But when it comes to anomaly detection and localization, no matter whether the distribution of anomaly is dispersed or occupies a large portion of the

TABLE II
COMPARISON OF FRAME-LEVEL AP PERFORMANCE ON THE XD-VIOLENCE VALIDATION SET.

Method	Feature	AP(%)
Sultani et al. [8]	CVPR 2018	73.20
HL-Net [20]	ECCV 2020	73.67
HL-Net [20]	ECCV 2020	78.64
RTFM [11]	ICCV 2021	77.81
MSL [16]	AAAI 2022	78.28
DAR [26]	TIFS 2022	78.94
DAR [26]	TIFS 2022	79.32
NG-MIL [14]	WACV 2023	78.51
S3R [18]	ECCV 2022	80.26
CU-Net [17]	CVPR 2023	78.74
CU-Net [17]	CVPR 2023	81.43
Pang et al. [28]	ICASSP 2021	81.69
UR-DMU [13]	AAAI 2023	81.66
UR-DMU [13]	AAAI 2023	81.77
Ours	RGB	83.59
Ours	RGB+Audio (concat)	83.77
Ours	RGB+Audio (with TC)	84.23

whole video, our method outperforms UR-DMU [13] due to our emphasis on identifying snippet-level abnormalities, and will be demonstrated in the subsequent subsection IV-F.

D. Result on XD-Violence

Table II displays the AP scores of state-of-the-art methods on the XD-Violence dataset. The feature we used is in RGB modality from the I3D network with 5-crop augmentation provided by [20]. Compared to the latest work UR-DMU [13], which employs additional normal and abnormal memory blocks to acquire more discriminative features, our method achieves a significant improvement of 1.83%. Furthermore, to test the robustness of our method when using multi-modal features such as RGB and Audio (to be consistent with previous works, we also concatenate them), resulting in an AP score of 83.77% for our technique which exceeds all the existing works. Additionally, we have implemented a straightforward fusion procedure where the RGB modality is first fed into a temporal convolution (TC) module and then concatenated with the original audio feature in the feature dimension. We adopt this approach because the extracted RGB feature is deemed more significant, and there is a domain gap between the final detection task and the extracted network. Introducing a temporal embedding layer aids in aligning the feature to be more task-oriented for the detection task. The new state-of-the-art detection performance of 84.23% verify our consideration. All these results demonstrate the efficacy of our method in identifying anomalous events, particularly those with disturbances.

E. Ablation Study

1) *Impact of the Suppressed manner:* As mentioned earlier, our goal is to suppress the most discriminative parts of the entire video while retaining some contextual information when using the MIL pipeline. However, after conducting experiments, we observed that thoroughly removing discriminative snippets actually resulted in greater improvements. The results are presented in Table III, where we examined different rates

TABLE III
ABLATION STUDY OF THE EXTENT OF SUPPRESSION OF DISCRIMINATIVE SNIPPETS. THE SMALLER β IS, THE GREATER LEVEL OF SUPPRESSION.

β	0	0.025	0.05	0.10	0.15	0.20
XD-Violence	83.59	82.29	81.08	81.84	81.21	80.28

TABLE IV
ABLATION STUDY ON THE EFFECTIVENESS OF THE DIFFERENT COMPONENTS IN THE OPTIMIZING PROCESS ON BOTH DATASET.

	L_c^o	L_c^a	L_c^{so}	L_c^{sa}	L_{guide}	L_{norm}	Result (%)
							UCF(AUC) XD(AP)
1	✓	-	-	-	-	-	81.96 79.28
2	✓	✓	-	-	-	-	84.34 79.60
3	✓	✓	-	-	-	✓	83.95 78.15
4	✓	✓	-	-	✓	-	84.73 80.20
5	✓	✓	-	-	✓	✓	84.98 80.05
6	✓	✓	✓	✓	-	-	84.45 82.76
7	✓	✓	✓	✓	✓	-	85.06 81.69
8	✓	✓	✓	✓	-	✓	85.07 81.07
9	✓	✓	-	✓	✓	✓	85.26 81.57
10	✓	✓	✓	-	✓	✓	84.35 80.92
11	✓	✓	✓	✓	✓	✓	86.19 83.59

of dropped snippets on the XD-Violence dataset. To provide more details, we conducted experiments using $\beta * S^{so}$ or $\beta * S^{sa}$ in place of 0 for the “otherwise” case in equations 5 and 7. We find that as the value of β decreases, the performance increases. The experimental results are not consistent with our anticipated outcomes, and we attribute this discrepancy to the background information already better modeled in less discriminative segments.

2) *Effectiveness of the Optimization Items:* In this part, we not only conduct an ablation study to examine the effectiveness of the various components in the optimizing process (as shown in Table IV), but we also present the AUC_sub and AP_sub values on the anomaly subset to analyze the function of the Multi-branch Supervision Module (as presented in Table V).

With the optimization items of guide and norm combined, in 2 and 5, and lines 6 and 11, the performance has achieved 0.64% and 1.74% improvement on the UCF dataset. For XD, guide and norm loss provide 0.45% and 0.83% AP gain, indicating the usefulness of these two items and they are more important when the “suppress” S^{so} and S^{sa} are utilized. Besides, using the guide and norm separately does not increase the performance significantly or even harm the result (3, 4 and 7, 8), especially in the XD dataset, where the abnormal snippets containing are more evenly distributed.

The introduced attention scores S^a lead to significant performance improvements, particularly on the UCF dataset with gains from 81.96% to 84.34% for lines 1 and 2 in Table IV, and improvements in AUC_sub and AP_sub from 62.20% to 66.67% and 22.66% to 31.43% respectively, as shown in line 6 and 7 of Table V. Results from the ablation study suggest that detecting abnormalities with BinaryCrossEntropy alone is insufficient. After the attention-based scores are utilized, the UCF-Crime dataset in which the anomaly is more dispersed can obtain a better improvement. But for XD, because WS-VAD is only a two-class problem, the gain is not obvious.

Besides, performance improvement also occurs on both datasets when the attention-based suppressed abnormal branch S^{sa} is introduced, line 9 of Table IV, items 4 and 9 of Table

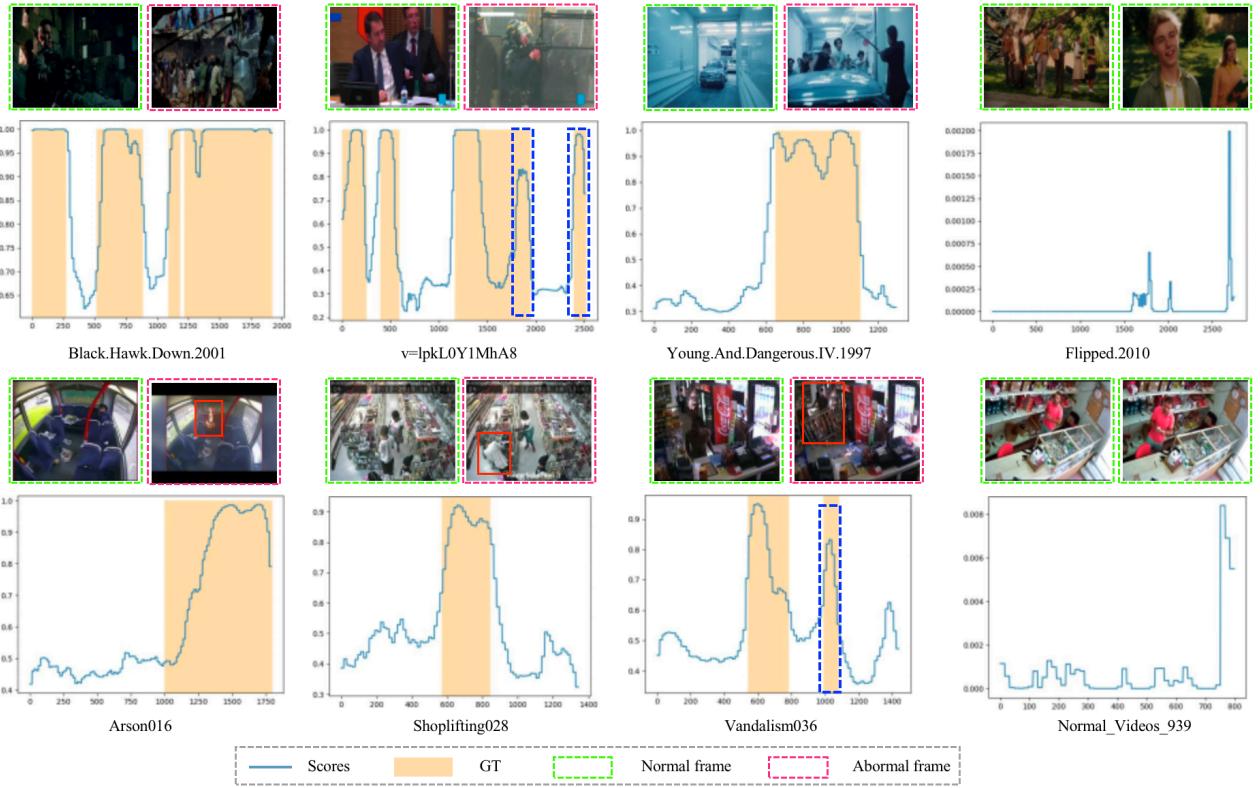


Fig. 4. Qualitative results of anomaly detection performance on XD-Violence (Two rows above) and UCF-Crime dataset (Two rows below).

TABLE V

ABLATION STUDY ON THE FUNCTION OF THE COMPONENTS IN MULTI-BRANCH SUPERVISION MODULE. AUC_{SUB} AND AP_{SUB} ARE THE SCORES ONLY USING ABNORMAL DATA WHICH CAN BE USED TO EVALUATE THE ANOMALY LOCALIZATION PERFORMANCE.

Datasets	Setting	AUC	AP	AUC _{SUB}	AP _{SUB}
XD	1 S^o	92.37	79.28	81.86	81.63
	2 $S^o + S^a$	92.88	79.60	81.20	81.19
	3 $S^o + S^a + S^{so}$	93.68	80.92	81.16	81.84
	4 $S^o + S^a + S^{sa}$	93.95	81.57	81.92	82.26
	5 ALL	94.47	83.59	83.02	84.19
UCF	6 S^o	81.96	19.61	62.20	22.66
	7 $S^o + S^a$	84.34	29.45	66.67	31.43
	8 $S^o + S^a + S^{so}$	84.35	29.11	64.52	30.63
	9 $S^o + S^a + S^{sa}$	85.26	30.65	67.35	32.83
	10 ALL	86.19	31.11	68.77	33.44

V. The phenomenon is predictable that suppressing the most discriminative snippets would aid the learning of hard snippets, thereby improving the result of anomaly detection. But when the original suppressed scores S^{so} are applied individually, the performance will be harmed, line 10 of Table IV, items 3 and 8 of Table V, due to overly severe penalties for the snippets which have low anomalous attention. Finally, the combination of all supervisions leads to our method achieving state-of-the-art performance.

3) *Effectiveness of the Attention Branch:* In Table VI, we compare the results of different scores used in the testing process. Our results show that attention greatly improves performance, no matter in the whole testing sets or only anomaly subsets. Furthermore, we demonstrate the effectiveness of the snippet-level anomaly-specific attention in Figure 5, where we

TABLE VI

ABLATION STUDY ON THE PERFORMANCE THAT USE DIFFERENT VALUES OF SCORES DURING TESTING.

Datasets		UCF				XD			
		AUC	AP	AUC _{SUB}	AP _{SUB}	AUC	AP	AUC _{SUB}	AP _{SUB}
W/O		85.03	29.34	66.58	31.20	94.08	81.84	82.10	82.65
With		86.19	31.11	68.77	33.44	94.47	83.59	83.02	84.19

compare testing video scores with S^a and without attention S^o . Our results indicate that the separate attention module allows the network to identify abnormal frames more accurately, resulting in a more polarized score curve, and can suppress false positives, shown in the *violet* rectangle of Figure 5. In summary, an attention branch is an effective tool for improving detection performance in differentiating between abnormal and normal classes from the snippet level.

4) *Impact of the number of divided snippets:* In the WS-VAD task, each video is typically divided into non-overlapping snippets during the training stage. In previous researches, T = 32 is set in [11], [14] and in [23] is 150, and in [13] is set with 200. Our approach takes snippet-level semantic information into consideration, which suggests that the number of snippets used in the training stage may impact performance. Therefore, we conduct an ablation experiment under different numbers of snippets. Table VIII demonstrates that the best results are obtained on the XD-Violence dataset when T = 100 with an AP of 83.59, while for the UCF-Crime dataset, the best performance 86.19 is achieved when T = 320. Furthermore, to ensure fairness in the comparison and provide substantial evidence of the effectiveness of our method, we reproduce

TABLE VII
COMPASION WITH UR-DMU UNDER VARYING NUMBERS OF DIVIDED SEGMENTS.

Segments	Setting	32				64				100				200				320				Avg			
		AUC	AP	AUC _{sub}	AP _{sub}	AUC	AP	AUC _{sub}	AP _{sub}	AUC	AP	AUC _{sub}	AP _{sub}	AUC	AP	AUC _{sub}	AP _{sub}	AUC	AP	AUC _{sub}	AP _{sub}	AUC	AP	AUC _{sub}	AP _{sub}
XD	UR-DMU	92.73	76.99	79.22	78.36	90.55	78.55	79.87	80.92	92.99	78.00	80.87	79.64	94.02	81.66	82.36	82.85	93.53	80.86	80.78	81.73	92.76	79.21	80.62	80.70
	OURS	92.27	78.91	80.64	81.29	93.71	80.68	82.14	82.02	94.47	83.59	83.02	84.19	94.05	83.31	81.44	83.83	94.35	83.17	82.50	83.75	93.77	81.93	81.95	83.02
	Margin	0.46 ↓	1.92↑	1.42↑	2.93↑	3.16↑	2.13↑	2.27↑	1.10↑	1.48↑	5.59↑	2.15↑	4.55↑	0.03↑	1.65↑	0.92↓	0.98↑	0.82↑	2.31↑	1.72↑	2.02↑	1.01↑	2.72↑	1.33↑	2.32↑

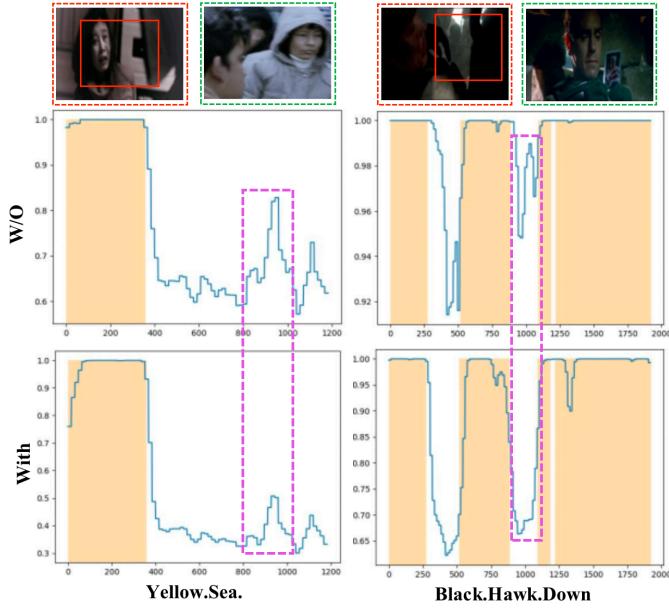


Fig. 5. Visualization of obtained curves when using different values of anomaly scores (W/O refers to S^o and With refers to S^a) during the testing phase.

TABLE VIII

PERFORMANCE OF THE NUMBER OF DIVIDED SNIPPETS DURING THE TRAINING STAGE.

Dataset	Segments					
	32	64	100	200	320	400
UCF	85.24	84.93	84.40	84.75	86.19	85.49
XD	78.91	80.68	83.59	83.31	83.17	81.53

the UR-DMU [13] method, which is openly available, under varying settings for the number of divided snippets on the XD-Violence dataset. The presented results, as shown in Table VII, establish that our method consistently outperforms UR-DMU. Furthermore, we performed a comparison of our approach against other studies while maintaining the same number of segmented snippets (i.e., T=32). The results are shown in Table IX, and our method continues to demonstrate superior performance.

5) *Effect of iterations M in abnormal guide loss:* The accuracy of abnormal video predictions during the initial training stage is unreliable. Therefore, we employ a soft approach to processing the guide loss for abnormal videos when the iteration is less than M . Once the iteration count exceeds M , we utilize hard labels with values of 1 and 0 for guiding backpropagation. The effect of parameter M on the final results of both datasets is presented in Table XI. Our method shows a rapid convergence, thus we only set the maximum value of the ablation as 700. It is evident that using hard labels to monitor at the beginning will result in

TABLE IX
COMPARISON WITH OTHER METHODS AT T = 32 ON XD-VIOLENCE DATASET.

Result (%)	XD-Violence				
	RTFM [11]	MSL [16]	NG-MIL [14]	CU-Net [17]	OURS
77.81	78.28	78.51	78.74	78.91	

TABLE X
IMPACT OF THE HYPERPARAMETER ϵ IN EQUATION (6).

ϵ	0.0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
UCF	85.45	84.44	86.01	85.46	86.19	85.37	85.38	85.94	85.49	84.78	
XD	79.92	81.82	82.13	82.07	83.59	82.24	82.35	82.35	82.30	82.63	82.60

a decrease in performance. And when the best parameter settings are reached, subsequent loss settings will also damage performance.

6) *Effect of hyperparameter α in the suppressed branch:* The hyperparameter α in equation (13) regulates the proportion of suppressed information during the optimization process. It also reflects the degree of emphasis placed on challenging anomalies. Ablation of α is shown in Figure 6. It is evident that excessively focusing on non-discriminative components during the optimization procedure will result in a decline in performance. Nevertheless, an interesting improvement in both the XD and UCF datasets was observed when alpha was set to 0.55, pink area in the figure.

7) *Effect of hyperparameter ϵ in the suppressed branch:* The hyperparameter ϵ in equation (6) determines which segments of the video will be identified as discriminative. Firstly, we would like to state that a smaller value of Ellison indicates a threshold closer to the minimum value of the entire video attention. This means that more video segments are considered discriminative. In the XD dataset, we observed that setting a lower value for Ellison leads to lower final detection results. Once the suppression limit is reached, there is no significant fluctuation in performance. We believe that this situation arises

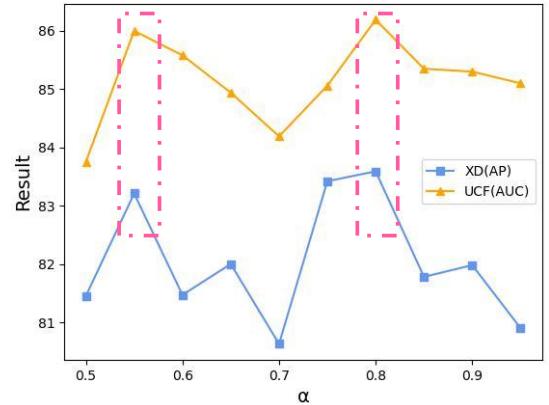


Fig. 6. Impact of the hyperparameter α in the total loss function.

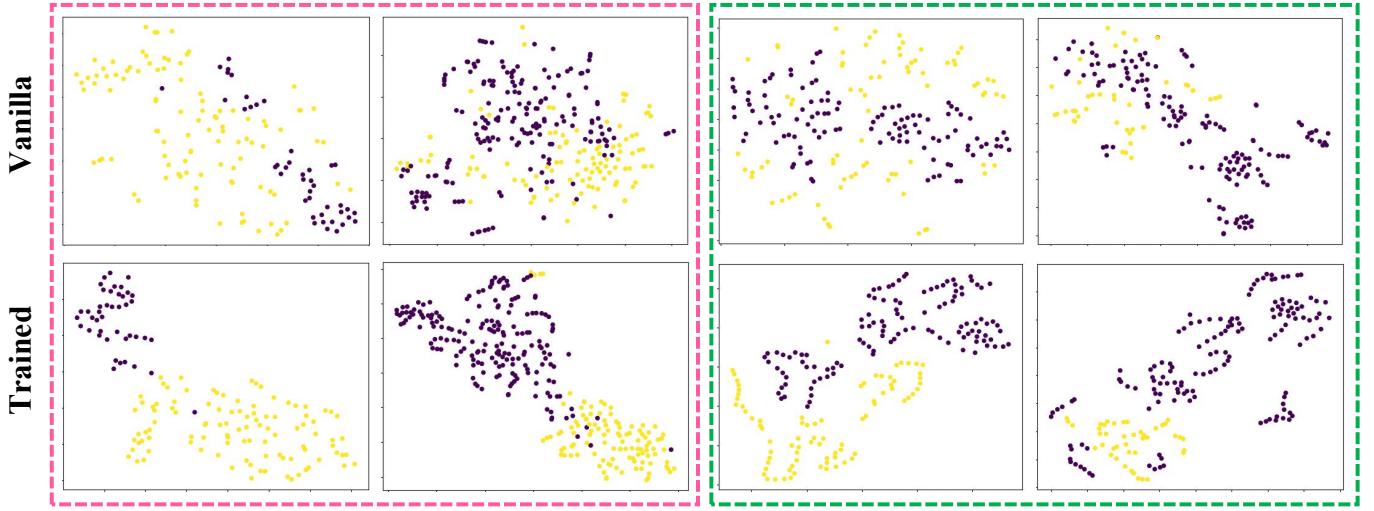


Fig. 7. Visualizations of the vanilla features and the output of our model on testing videos (XD-Violence and UCF-Crime).

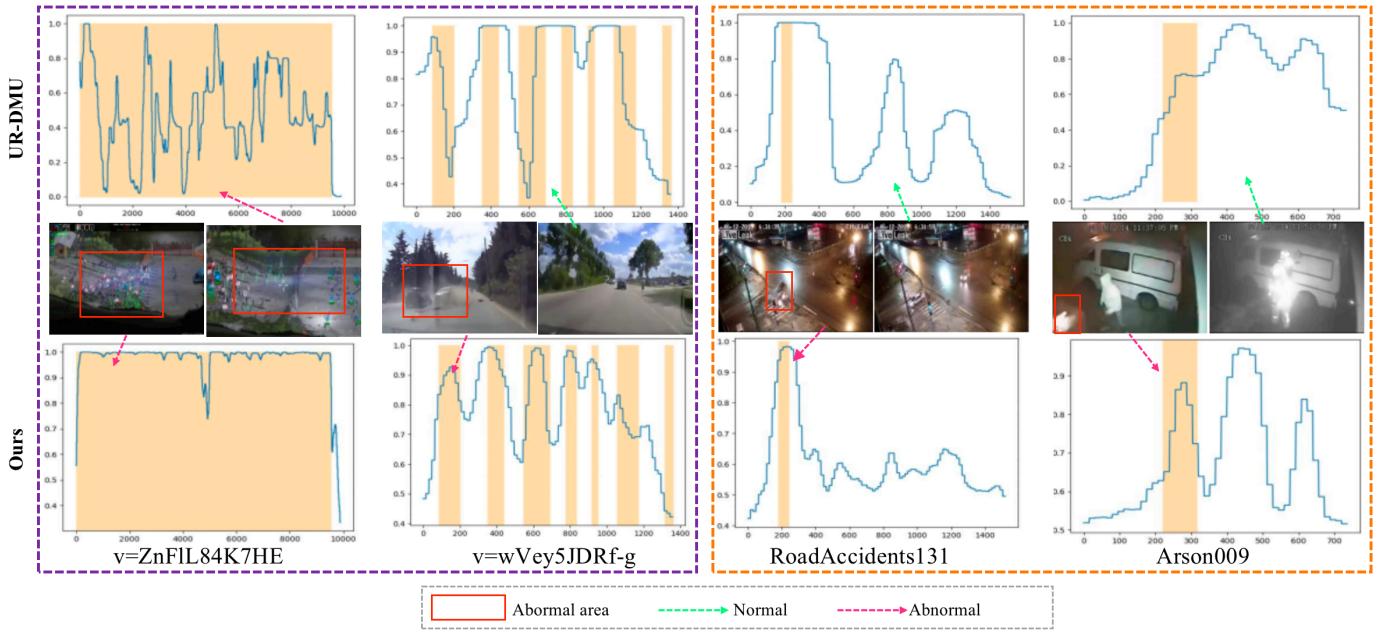


Fig. 8. Qualitative results of anomaly detection and localization performance on XD-Violence and UCF-Crime when compare with UR-DMU [13] method.

TABLE XI
PERFORMANCE OF THE M ITERATIONS IN EQUATION 9.

Dataset	Iteration							
	0	100	200	300	400	500	600	700
UCF	84.66	84.49	85.01	84.68	86.19	84.94	85.42	85.98
XD	81.82	81.99	81.75	82.16	83.59	82.24	81.86	81.46

because the number of discriminative segments in a video is limited, and they are truly easy to be recognized. Therefore, as the threshold increases, the actual variation in the recognized discriminative segments is not significant.

F. Qualitative Results

To further assess the effectiveness of our approach, we present qualitative results on two datasets in Figure 4. Eight

videos are shown with frame-level anomaly predictions, where the six videos on the left are positive, and the two on the far right are normal. The orange area indicates the ground truth, while cyan lines represent our predictions. Additionally, we highlight normal and abnormal frames in the videos with green and red boxes, respectively. The figure demonstrates that our method has effectively achieved good performance in terms of anomaly detection and event localization. Notably, our method can also effectively detect short-duration abnormal snippets, the blue rectangle shown in the figure.

We also present t-SNE [49] visualizations depicting the feature distributions on both benchmark test sets. Figure 7 displays the results, where abnormal segments are represented by yellow dots and normal features are represented by purple dots. It is evident that the normal and abnormal features



Fig. 9. Comparison of anomaly detection results on the XD-Violence dataset using either RGB-only features or RGB-Audio fusion features.

are distinctly clustered, and the distance between unrelated features is widened after the training process. This observation demonstrates that with the assistance of our proposed network, instances are effectively differentiated, thereby further validating the efficacy of our framework. Further, we compare our method with UR-DMU [13] on two datasets from the perspective of anomaly detection and localization. The result of UR-DMU is drawn through the best checkpoint released by the author, shown in Figure 8. Our method demonstrates greater robustness to noise from object changing and scene transforming (even if computer game scenes). When anomalies account for a large proportion of the whole video, our method effectively integrates local information surrounding the anomaly while concurrently restraining the influence of discriminative segments, resulting in a relatively smooth anomaly score curve with small fluctuations (column 1). Additionally, owing to the snippet-level focus, our approach demonstrates improved accuracy in localizing anomalies, particularly when the abnormal snippets within a video are short or the anomaly distribution is dispersed. This is evident from the result presented in columns 2 and 3. In column 4, the “Arson” event actually is of short duration, however, the camera videos are collected at night, thus the firefighters wearing reflective vests also are recognized as arson event. Although neither method effectively distinguishes normal snippets, our approach provides a more polarized result and proves that our method can better detect anomalies.

Finally, we present anomaly detection results utilizing vari-

ous modalities, as depicted in Figure 9. Normal frames are represented by the **green** box, while abnormal frames are indicated by the **red** box. The fusion of modalities, represented by the **violet** box, provides an enhanced final detection outcome. Conversely, the result represented by the **cyan** color is detrimental. In the first row, the introduction of audio features enables accurate localization of abnormal events, despite flame elements seemly existing in the **blue** portion of the video clip, with the background sound suggesting that it is actually driving. In the second row, even though there are no obvious cues such as guns or other visual information, the abnormal frame is still detected, due to the presence of gunfire sounds throughout the entire video. The subsequent two rows display normal videos, with the last row revealing an increase in abnormal scores, which may be attributed to the presence of noisy background sound. To sum up, the visualization showcases the benefits of combining visual and acoustic information, highlighting the importance of using a reasonable fusion approach to avoid introducing noise that could potentially compromise the overall performance.

V. FURTHER CONSIDERATION

During the experimental process, we discovered two limitations in our methods. Firstly, the MSE loss function poses an unsatisfactory backpropagation when the attention value approaches 0.5. However, 0.5 is a crucial threshold for distinguishing between normal and abnormal events. Secondly, we explored other new forms of attention modules, such as a fusion of saliency and contextual information, which initially showed promising results during training but ultimately yielded dissatisfactory outcomes. In essence, these situations highlight the need for further improvement in the optimization process designed by us.

VI. CONCLUSION

In this paper, we propose a method that takes the snippet-level encoded features into consideration. Concretely, after modeling the original features at global and local levels, an attention mechanism is introduced. Then together with the snippet anomalous attention, a multi-branch supervision module is proposed, in which not only the general predicted scores but also attention-based predictions are utilized. Besides, we also suppress the most discriminative snippets, so the hard portion of the video can be learned and then explored the anomaly completeness. Finally, for better generating anomalous attention, an optimizing process that contains norm and guide items is given. With the combination of components mentioned above, our method achieves state-of-the-art performance on two large benchmark datasets.

REFERENCES

- [1] Y. Lai, Y. Han, and Y. Wang, “Anomaly detection with prototype-guided discriminative latent embeddings,” in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 300–309.
- [2] M. Z. Zaheer, A. Mahmood, M. H. Khan, M. Segu, F. Yu, and S.-I. Lee, “Generative cooperative learning for unsupervised video anomaly detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14744–14754.

- [3] J. T. Zhou, L. Zhang, Z. Fang, J. Du, X. Peng, and Y. Xiao, "Attention-driven loss for anomaly detection in video surveillance," *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 12, pp. 4639–4647, 2019.
- [4] Y. Lu, C. Cao, Y. Zhang, and Y. Zhang, "Learnable locality-sensitive hashing for video anomaly detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 2, pp. 963–976, 2022.
- [5] Y. Zhong, X. Chen, Y. Hu, P. Tang, and F. Ren, "Bidirectional spatio-temporal feature learning with multiscale evaluation for video anomaly detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8285–8296, 2022.
- [6] Y. Zhang, X. Nie, R. He, M. Chen, and Y. Yin, "Normality learning in multispace for video anomaly detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 9, pp. 3694–3706, 2020.
- [7] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. 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*, 2021, pp. 13588–13597.
- [8] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.
- [9] Y. Zhu and S. Newsam, "Motion-aware feature for improved video anomaly detection," *arXiv preprint arXiv:1907.10211*, 2019.
- [10] J. Zhang, L. Qing, and J. Miao, "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 4030–4034.
- [11] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4975–4986.
- [12] P. Wu and J. Liu, "Learning causal temporal relation and feature discrimination for anomaly detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3513–3527, 2021.
- [13] H. Zhou, J. Yu, and W. Yang, "Dual memory units with uncertainty regulation for weakly supervised video anomaly detection," *arXiv preprint arXiv:2302.05160*, 2023.
- [14] S. Park, H. Kim, M. Kim, D. Kim, and K. 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*, 2023, pp. 2665–2674.
- [15] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, "Mist: Multiple instance self-training framework for video anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14009–14018.
- [16] S. Li, F. Liu, and L. Jiao, "Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 1395–1403.
- [17] C. Zhang, G. Li, Y. Qi, S. Wang, L. Qing, Q. Huang, and M.-H. Yang, "Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection," *arXiv preprint arXiv:2212.04090*, 2022.
- [18] J.-C. Wu, H.-Y. Hsieh, D.-J. Chen, C.-S. Fuh, and T.-L. Liu, "Self-supervised sparse representation for video anomaly detection," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*. Springer, 2022, pp. 729–745.
- [19] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
- [20] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXXII 16*. Springer, 2020, pp. 322–339.
- [21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [22] M. Z. Zaheer, A. Mahmood, M. Astrid, and S.-I. 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 XXXII 16*. Springer, 2020, pp. 358–376.
- [23] C. Cao, X. Zhang, S. Zhang, P. Wang, and Y. Zhang, "Adaptive graph convolutional networks for weakly supervised anomaly detection in videos," *IEEE Signal Processing Letters*, vol. 29, pp. 2497–2501, 2022.
- [24] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.
- [25] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. 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*, 2019, pp. 1237–1246.
- [26] T. Liu, C. Zhang, K.-M. Lam, and J. Kong, "Decouple and resolve: Transformer-based models for online anomaly detection from weakly labeled videos," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 15–28, 2022.
- [27] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, and Y.-C. Wu, "Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection," *arXiv preprint arXiv:2211.15098*, 2022.
- [28] W.-F. Pang, Q.-H. He, Y.-j. Hu, and Y.-X. Li, "Violence detection in videos based on fusing visual and audio information," in *ICASSP 2021–2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2021, pp. 2260–2264.
- [29] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.
- [30] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2740–2755, 2018.
- [31] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.
- [32] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.
- [33] N.-C. Risteau, N. Madan, R. T. Ionescu, K. Nasrollahi, F. S. Khan, T. B. Moeslund, and M. Shah, "Self-supervised predictive convolutional attentive block for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13576–13586.
- [34] D. Purwanto, Y.-T. Chen, and W.-H. Fang, "Dance with self-attention: A new look of conditional random fields on anomaly detection in videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 173–183.
- [35] J. Wu, W. Zhang, G. Li, W. Wu, X. Tan, Y. Li, E. Ding, and L. Lin, "Weakly-supervised spatio-temporal anomaly detection in surveillance video," *arXiv preprint arXiv:2108.03825*, 2021.
- [36] A. Islam, C. Long, and R. Radke, "A hybrid attention mechanism for weakly-supervised temporal action localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1637–1645.
- [37] P. Lee, Y. Uh, and H. Byun, "Background suppression network for weakly-supervised temporal action localization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11320–11327.
- [38] F.-T. Hong, J.-C. Feng, D. Xu, Y. Shan, and W.-S. Zheng, "Cross-modal consensus network for weakly supervised temporal action localization," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1591–1599.
- [39] S. Qu, G. Chen, Z. Li, L. Zhang, F. Lu, and A. Knoll, "Acnn-net: Action context modeling network for weakly-supervised temporal action localization," *arXiv preprint arXiv:2104.02967*, 2021.
- [40] M. Liu, X. Li, Y. Liu, and Y. Han, "Weakly supervised anomaly detection with multi-level contextual modeling," *Multimedia Systems*, pp. 1–12, 2023.
- [41] Y. Zhao, H. Zhang, Z. Gao, W. Gao, M. Wang, and S. Chen, "A novel action saliency and context-aware network for weakly-supervised temporal action localization," *IEEE Transactions on Multimedia*, 2023.
- [42] N. A. Tu, T. Huynh-The, K. U. Khan, and Y.-K. Lee, "Mi-hdp: A hierarchical bayesian nonparametric model for recognizing human actions in video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 3, pp. 800–814, 2018.
- [43] L. Huang, L. Wang, and H. Li, "Foreground-action consistency network for weakly supervised temporal action localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8002–8011.
- [44] Z. Luo, D. Guillory, B. Shi, W. Ke, F. Wan, T. Darrell, and H. Xu, "Weakly-supervised action localization with expectation-maximization multi-instance learning," in *Computer Vision–ECCV 2020: 16th European*

- Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX*
16. Springer, 2020, pp. 729–745.
- [45] S. Paul, S. Roy, and A. K. Roy-Chowdhury, “W-talc: Weakly-supervised temporal activity localization and classification,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 563–579.
- [46] A. Islam and R. Radke, “Weakly supervised temporal action localization using deep metric learning,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 547–556.
- [47] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, “The kinetics human action video dataset,” *arXiv preprint arXiv:1705.06950*, 2017.
- [48] H. Lv, Z. Yue, Q. Sun, B. Luo, Z. Cui, and H. Zhang, “Unbiased multiple instance learning for weakly supervised video anomaly detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8022–8031.
- [49] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.” *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [50] Huawei. Mindspore. <https://www.mindspore.cn/>, 2020.