This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

TLMA: Mitigating the Impact of Weakly Labeled Information for Video Anomaly Detection

Rong Xu[1,2,3] Runqi Wang[2,3*] Yingjun Zhang[2,3] Tao Tao[1,2,3] Xiaomeng Li[1,2,3] Liping Jing[1,2,3]
[1]State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University
[2]Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing Jiaotong University
[3]School of Computer Science & Technology, Beijing Jiaotong University

Abstract

Weakly Supervised Video Anomaly Detection (WSVAD) aims to localize abnormal segments using only video-level labels during training. Although the paradigm significantly reduces annotation costs, the coarse-grained labels fail to precisely describe the full videos, resulting in the introduction of substantial Weakly Labeled Information (WLI) during training. The presence of WLI makes it difficult for the model to accurately learn the boundary between normal and abnormal behaviors, leading to misclassifications and compromising the precision of anomaly localization. To tackle the challenges posed by WLI, we propose a triplet learning strategy that selects hard segments from normal videos as anchors. By combining contrastive learning with Multiple Instance Learning (MIL) strategy, we increase the projection distance between abnormal segments and anchor samples, to reduce the interference of WLI in anomaly detection. Moreover, considering that anomalies typically occur in dynamic foreground regions, we further design a motion-aware feature enhancement module that extracts dynamic areas within each video segment to emphasize the representation of critical features. This not only improves the accuracy of anchors in triplets, but also enhances the discriminative power of instance features in MIL. Extensive experiments on UCF-Crime, XD-Violence, and MSAD datasets demonstrate the effectiveness of our approach.

1. Introduction

Video Anomaly Detection (VAD) aims to identify events or behaviors that deviate from normal patterns in videos, and plays a critical role in intelligent surveillance, public safety, and industrial monitoring [3, 10, 17]. In recent years, Weakly Supervised Video Anomaly Detection (WSVAD) has attracted increasing attention for its ability to perform effective anomaly detection using only video-level labels (i.e., a single "abnormal" or "normal" tag per video), thereby avoiding the need for expensive frame-level or pixel-level annotations. Most WSVAD approaches adopt the Multiple Instance Learning (MIL) framework [21], where a video is treated as a bag of segments: normal videos contain only normal segments, while abnormal videos contain at least one abnormal segment. A ranking loss is used in MIL to widen the score gap between high-anomaly-scoring segments from abnormal and normal videos, respectively, enabling segment-level detection. Building upon the MIL framework, subsequent studies have introduced various improvements [14, 29, 33, 35].

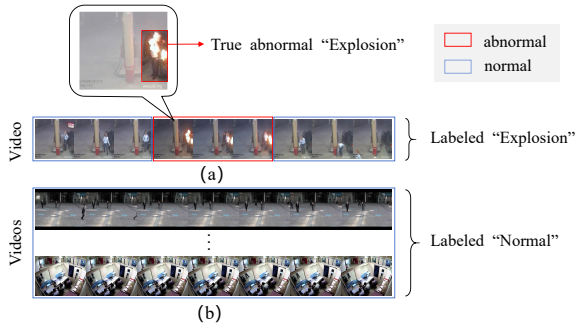
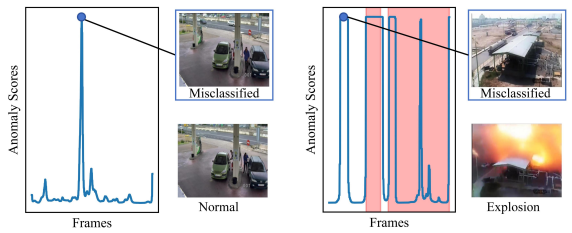


Figure 1. Typical forms of WLI. **(a)** An abnormal video uniformly labeled as "explosion", while the actual abnormal information only occurs in the foreground of partial segments (red box). Other segments and background are included in the label but are actually normal, which is known as WLI. **(b)** All of normal videos with diverse behaviors are labeled as "normal", where atypical segments (e.g., those with intense motion or uncommon patterns) may deviate from the norm and constitute WLI.

Although the WSVAD paradigm greatly reduces annotation costs by relying on video-level labels, it inevitably introduces substantial Weakly Labeled Information (WLI) that does not accurately match its labels. For example, it includes normal content within abnormal videos, as well as atypical content within normal videos. As illustrated in Figure 1(a), an abnormal video is assigned the label "explo-

*Corresponding author.



(a) Misclassification in normal video (b) Misclassification in abnormal video

Figure 2. Illustration of misclassification cases caused by WLI. The anomaly scores (blue curve) are produced by UR-DMU [33], with pink areas indicating ground-truth anomalies. In (a), a normal refueling scene at a gas station is misclassified as anomalous, in (b), the normal scene preceding an explosion in an abnormal video is also misclassified.

sion" based solely on the presence of anomalous foreground objects in a specific segment. However, the remaining information, which includes normal transitional segments and background regions, is also described by the video-level "explosion" label. The information that is abnormally-labeled yet actually normal constitutes a critical part of WLI, which distracts the model from true anomalies. On the other hand, normal videos, which account for nearly half of the training data, are uniformly labeled as "normal" despite their high diversity in content and behavior, as shown in Figure 1(b). This single label fails to represent the wide range of normal patterns. In particular, atypical normal segments (e.g., those with intense motion or uncommon patterns) deviate significantly from typical normal features in the embedding space and are easily misclassified as anomalous. Such segments further contribute to WLI.

A more practical issue is that the cost of increasing high-quality, fine-grained annotations is extremely high, making it infeasible. Therefore, we have to **mitigate the impact of WLI and enhance anomaly detection performance under weak supervision, without relying on additional frame-level or pixel-level annotations**. We observe that misclassifications most often occur in WLI, where the distinction between normal and abnormal behaviors is unclear. Such segments lack precise supervision due to coarse video-level labels and disrupt the model's learning process. As a result, normal parts of abnormal videos may be misclassified as anomalies, while atypical normal segments are also mistaken for anomalies, as shown in Figure 2. Effectively distinguishing WLI from true abnormal cues is therefore crucial to mitigate misclassification in WSVAD.

To address this problem, we propose a **triplet learning** strategy with anchor, positive, and negative samples for WSVAD. It explicitly increases the distance between WLI and truly abnormal content in the feature space. In the absence of fine-grained labels, the triplets are constructed dynamically based on the model's predictions. The core of our strategy is to select anchors that accurately represent WLI. Conceptually, WLI consists of two components: (1) atypical segments in normal videos and (2) normal segments in abnormal videos. However, the latter lack reliable annotations, making them infeasible for direct selection. Therefore, we focus on the former, namely atypical segments in normal videos, and use the high-anomaly-scoring, "anomalous-looking" ones as **anchors**. Although these segments are labeled as normal, they are highly prone to misclassification, making them a practical and representative proxy for WLI. The **positives** are low-anomaly-scoring segments from abnormal videos, typically containing no true anomalies and also belonging to WLI. The **negatives** are high-anomaly-scoring segments from abnormal videos, which most likely reflect true abnormal content that is most consistent with the video-level labels. The triplet loss encourages the model to pull the anchor and positive (both WLI) closer, while pushing them away from the negative (true anomaly), thereby learning a more structured feature space that separates WLI from true anomalies. Our strategy cooperates with the MIL ranking loss to deliver more precise anomaly localization. The MIL loss encourages scores in normal videos to approach zero (thereby suppressing anchors and positives) while amplifying significant anomalies.

In addition, anomalies in videos typically occur only within dynamic foreground regions. However, most existing WSVAD methods extract features from entire frames by default, where redundant background information may introduce interference. This not only degrades localization accuracy but also corrupts the feature quality for triplet construction. To address this, we introduce a **motion-aware feature enhancement** module, which identifies motion-salient regions by capturing inter-frame edge variations. This module enhances the representation of dynamic foreground regions while suppressing irrelevant background content, thereby improving the model's ability to perceive abnormal targets. The resulting motion-focused features possess stronger discriminability, providing more precise features for the triplet loss to operate on. Ultimately, this module mitigates the impact of WLI, without requiring additional fine-grained annotations.

To tackle the challenges posed by WLI as analyzed above, we propose a novel framework named **TLMA** (**T**riplet **L**earning with **M**otion-**A**ware enhancement) to mitigate its impact and enhance anomaly detection under weak supervision. Our main contributions are summarized as follows:

- To explicitly separate WLI from truly anomalous content in the feature space, we propose a triplet learning strategy that constructs triplets from model outputs. This strategy encourages tighter representation of WLI segments and greater separation from true anomalies.

- We introduce a motion-aware feature enhancement module that focuses on dynamic foreground regions by leveraging inter-frame edge changes, effectively suppressing background interference and improving anomaly discriminability.
- Extensive experiments on three WSVAD benchmarks show that our method TLMA achieves competitive performance, consistently validating the effectiveness of both the triplet learning and motion-aware components in mitigating WLI interference.

2. Related Work

Weakly Supervised Video Anomaly Detection (WSVAD). WSVAD trains the model using video-level labels from normal and anomalous videos, without requiring frame-level annotations. The pioneering work by Sultani et al. [21] first proposed a Multiple Instance Learning (MIL) framework for video anomaly detection, treating videos as bags of instances and employing a ranking loss for model training. Subsequent WSVAD approaches have introduced numerous improvements based on the MIL framework [1, 5, 6, 9, 11, 20, 22]. Tian et al. proposed RTFM [22], which learns a feature magnitude function to suppress noise and capture multi-scale temporal dependencies for better anomaly detection. Wu et al. [27] aggregate the scores of top-k high-scoring snippets during training. NG-MIL [14] learns prototypes of normal patterns and employs a triplet loss to enhance discriminability. UR-DMU [33] also employs a triplet loss, introducing a dual-memory structure to model both normal and abnormal patterns. Some works, such as MIST [7] and CUPL [31], adopt pseudo-label-based two-stage training schemes. CUPL [31] introduced MC Dropout to estimate epistemic uncertainty and selected confident pseudo-labels for a two-stage training process. Other works [12, 15, 28, 29] have explored multi-modal settings. VadCLIP [29] leverages CLIP and utilizes text prompts to align video-text semantics. MajhiD et al. [12] enhances RGB representations by introducing five additional modalities. Although these methods achieve promising results on various benchmarks, they still suffer from the effects of Weakly Labeled Information (WLI), which often leads to misclassifications. To mitigate this issue, we propose a triplet-based training strategy that explicitly separates WLI from true anomalies in the feature space, thereby enhancing anomaly localization under weak supervision.

Feature Extraction. Most WSVAD approaches adopt pretrained networks as feature extractors. Early methods typically used models such as C3D [24] and I3D [2] trained on video classification datasets. With the rise of multimodal learning, CLIP [16] has been widely adopted for its ability to align visual and textual representations. More recently, transformer-based models [23, 25, 26] pretrained via self-supervised video representation learning, such as VideoMAE [23], have been employed to leverage their strong spatiotemporal modeling capacity [20]. While these pretrained models improve the overall feature quality, redundant background information in videos still interferes with anomaly detection. To address this, we design a motion-aware feature enhancement module on top of the ViT-based encoder. Our approach highlights motion foregrounds and suppresses background redundancy by simply computing inter-frame edge differences, requiring neither additional networks nor extra training.

3. Method

To mitigate the adverse effects of WLI on anomaly detection under weak supervision, we propose a unified framework named **TLMA**, which integrates two key components: (1) a **T**riplet **L**earning strategy to separate WLI from true anomalies, and (2) a **M**otion-**A**ware feature enhancement module to amplify motion-salient foreground cues. The overall architecture of TLMA is illustrated in Figure 3.

3.1. MIL Formulation

In the WSVAD setting, only video-level labels $Y \in \{0, 1\}$ are available, where $Y = 1$ indicates the presence of anomaly in a video, and $Y = 0$ indicates the video is entirely normal. Following the MIL paradigm, a video is divided into L non-overlapping segments. Each segment is encoded into a feature representation and assigned an anomaly score $s_j \in [0, 1]$ by a classifier. To aggregate segment-level predictions into a video-level decision, we adopt a top-k pooling strategy [27], where the final video-level anomaly score S_V is computed as the average of the top-k segment scores:

$$S_V = \text{mean}(\text{Top-}k(s_1, s_2, \ldots, s_L)). \tag{1}$$

This formulation encourages the model to assign high scores to at least a few abnormal segments in abnormal videos, while maintaining low scores across all segments in normal videos. The video-level classification loss is defined as:

$$\mathcal{L}_{\text{MIL}} = -\mathbb{E}\left[Y \log S_V + (1 - Y) \log(1 - S_V)\right]. \tag{2}$$

3.2. Triplet Learning Strategy

WLI, introduced by coarse video-level supervision in WSVAD, inevitably injects interference into the learning process and often causes misclassifications. To mitigate this issue, we propose a **triplet learning** strategy that helps the model explicitly separate WLI from truly abnormal patterns in the feature space.

Since fine-grained annotations are unavailable, we construct triplets from model outputs. The strategy hinges on selecting anchors that accurately represent WLI. Although

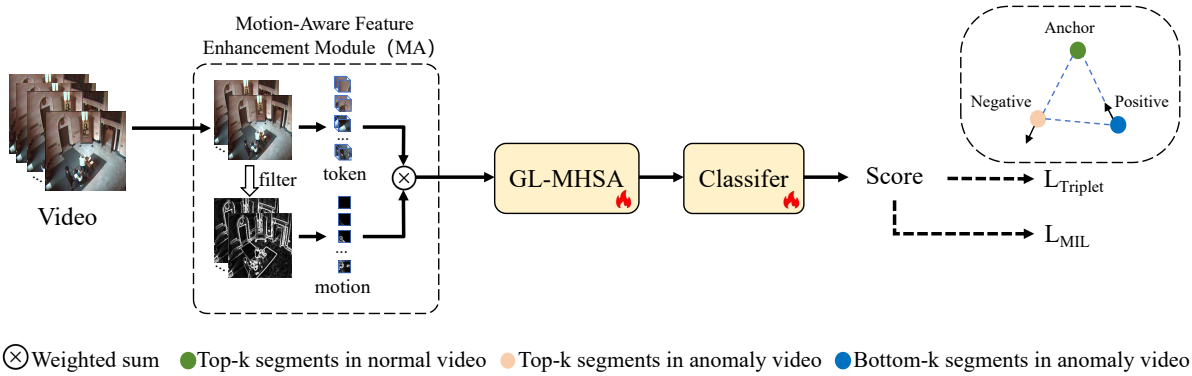


Figure 3. Overview of the proposed **TLMA framework**. It consists of two key components: a **T**riplet **L**earning strategy, which explicitly separates WLI from true anomalies in the feature space, and a **M**otion-**A**ware feature enhancement module, which emphasizes motion-salient foreground cues while suppressing background interference. The overall model is then optimized under weak supervision to improve the accuracy of anomaly localization.

WLI conceptually consists of atypical segments in normal videos and normal segments in abnormal videos, the latter cannot be reliably identified due to the lack of segment-level annotations. In contrast, the atypical segments in normal videos are both accessible and visually "anomalous-looking" (i.e., assigned high anomaly scores by the model), making them highly prone to misclassification. Therefore, we designate these segments as **anchors** in our triplet design, leveraging them as a practical and effective proxy for WLI. We then select **positives** from low-anomaly-scoring segments in abnormal videos. These segments are highly likely to not contain true anomalies and are also considered WLI. However, they are not certain to be WLI and thus cannot be used as anchors. Finally, **negatives** are chosen from high-anomaly-scoring segments from abnormal videos, which most likely correspond to truly anomalous regions. As illustrated in Figure 4, rely on the predicted anomaly scores, we construct triplets (F_a, F_p, F_n) as follows:

- **Anchor** (F_a): average feature of the top-k_a high-scoring segments from normal videos;
- **Positive** (F_p): average feature of the k lowest-scoring segments extracted from abnormal videos;
- **Negative** (F_n): average feature of the k highest-scoring segments extracted from abnormal videos.

We adopt a margin-based triplet loss:

$$\mathcal{L}_{\text{Triplet}} = \sum_{(a,p,n)} \max\left(\|F_a - F_p\|_2 - \|F_a - F_n\|_2 + m,\ 0\right), \tag{3}$$

where the margin m is set to 1.

The triplet loss encourages the model to pull the anchor and positive samples (highly likely to be WLI) closer together, while pushing them away from the negative sample (true anomaly). This promotes a more structured feature space that explicitly separates WLI from truly anomalous patterns. In parallel, a MIL-based ranking loss operates at the score level. It suppresses anomaly scores across normal videos, thereby influencing anchors and positives, while amplifying high scores in abnormal segments. Together, these complementary objectives enable more precise anomaly localization.

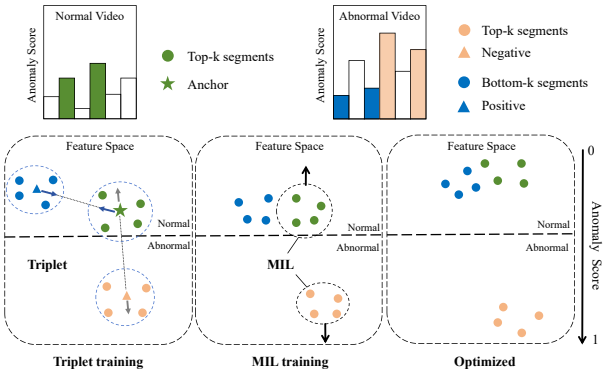


Figure 4. Illustration of the triplet learning strategy and its synergy with MIL optimization. Triplets are constructed based on predicted anomaly scores: top-k_a segments from normal videos form anchor, bottom-k segments from abnormal videos as positive, and top-k segments from abnormal videos as negative. The triplet loss enlarges the distance between WLI (anchor and positive) and true anomalies (negative) in the feature space, while the MIL loss increases anomaly score separation.

3.3. Motion-Aware Feature Enhancement

In video anomaly detection, abnormal behaviors typically occur in dynamic foreground regions. However, existing WSVAD methods often extract features from full frames, introducing irrelevant background information. This background interference not only degrades the precision of anomaly scoring but also impairs the selection of reliable contrastive samples for the triplet learning strategy. To al-

leviate these issues, we propose a motion-aware feature enhancement module that suppresses static background interference and enhances motion-salient foreground representations.

Spatiotemporal Feature Tokenization. Following [20], we adopt a ViT-based backbone pretrained with VideoMAEv2 [26] as our base feature extractor, leveraging its strong capacity for long-range spatiotemporal modeling. The masked reconstruction strategy of VideoMAEv2 encourages temporal structure awareness, making the extracted features more generalizable. Given a video sequence with T frames $V = \{f_1, f_2, \ldots, f_T\}$, we divide it into non-overlapping 16-frame segments and resize frames to 224×224. Each segment is partitioned into spatiotemporal cubes of size $2 \times 14 \times 14$ (temporal × spatial), which are projected through 3D convolutions into token embeddings $\{z_1, z_2, \ldots, z_N\}$. These embeddings are then passed through multi-layer transformer blocks ϕ to obtain contextualized token-level features: $\tilde{Z} = \phi(\{z_1, z_2, \ldots, z_N\}) \in \mathbb{R}^{N \times D}$, where N is the number of spatiotemporal tokens in a segment, and D is the feature dimension.

Motion-Aware Token Attention. To further enhance the sensitivity of feature representations to foreground dynamics, we introduce a novel motion-aware token attention mechanism that emphasizes dynamic regions aligned with the token structure, as illustrated in Figure 5. Specifically, we apply Sobel filters [19] to each frame to extract spatial edge responses, as they offer a simple yet effective way to highlight structural contours that are sensitive to motion. Compared to raw intensity values, Sobel responses are less affected by illumination changes and better preserve motion-relevant boundaries, making them ideal for motion-aware representation. The horizontal and vertical Sobel kernels are defined as:

$$S_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, \quad S_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}. \tag{4}$$

For each frame f_t, the edge magnitude map is computed as:

$$E_t(x, y) = \sqrt{(S_x * f_t)^2(x, y) + (S_y * f_t)^2(x, y) + \epsilon}. \tag{5}$$

To capture temporal motion, we compute inter-frame differences of edge maps:

$$M_k(x, y) = \left| E_{2k}(x, y) - E_{2k-1}(x, y) \right|, \quad k = 1, \ldots, 8. \tag{6}$$

Notably, the temporal resolution of motion computation aligns with the 2-frame temporal window used in the ViT's spatiotemporal tokenization, ensuring that each motion score corresponds directly to a specific 3D cube in the feature space.

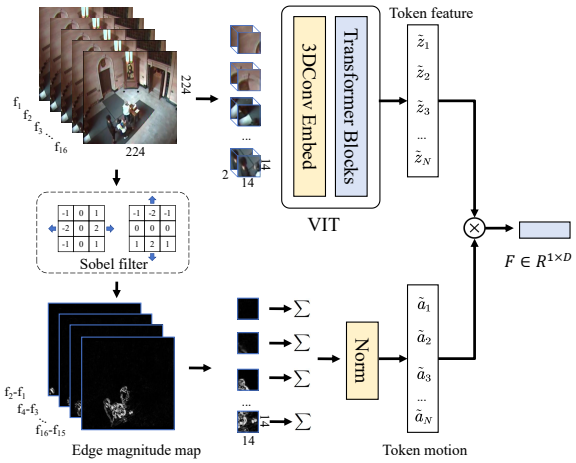


Figure 5. Overview of the proposed motion-aware feature enhancement module. The module extracts dynamic regions by computing inter-frame edge variations, guiding the model to emphasize motion-related anomaly cues while suppressing redundant background information.

These motion magnitude maps are then aligned with the token grid. For each token's spatial region Ω_i, its motion importance score $\tilde{a}_i$ is calculated as:

$$a_i = \sum_{(x,y)\in\Omega_i} M(x, y), \quad \tilde{a}_i = \frac{a_i}{\max_j a_j + \epsilon}. \tag{7}$$

Finally, the motion-aware enhanced feature is computed as a weighted sum of token features:

$$F_{\text{enhanced}} = \sum_{i=1}^{N} \tilde{a}_i \cdot \tilde{z}_i. \tag{8}$$

This process adaptively emphasizes tokens corresponding to regions with stronger motion, allowing the model to focus on foreground dynamics that are more indicative of anomalies. The resulting motion-sensitive features enhance the model's ability to produce more reliable anomaly score predictions, thus benefiting downstream triplet-based contrastive learning. Building upon these motion-aware features, we further incorporate the GL-MHSA module [33] to enhance temporal modeling and capture long-range inter-segment dependencies.

3.4. Joint Optimization Objective

The overall training objective of the TLMA framework integrates the MIL loss and the proposed triplet loss. The total loss is defined as:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{MIL}} + \lambda \cdot \mathcal{L}_{\text{Triplet}}, \tag{9}$$

where λ is a balancing hyperparameter that controls the contribution of the triplet loss. This joint optimization encourages consistency between video-level supervision and

feature-space structuring. The MIL loss guides the model to assign discriminative anomaly scores, while the triplet loss explicitly separates WLI from true anomalies. Together, they reinforce each other to enhance anomaly localization performance under weak supervision.

4. Experiments

4.1. Datasets and Implementation Details

Datasets. We evaluate our method on three widely used public WSVAD datasets: UCF-Crime [21], XD-Violence [27], and MSAD [34]. UCF-Crime contains 1,900 real-world surveillance video clips with a total duration of 128 hours, spanning 13 types of anomalous behaviors. XD-Violence contains 4,754 untrimmed videos covering six different categories of abnormal or violent behaviors. MSAD contains 720 video clips from 14 distinct scenes, annotated with 11 types of anomalous events.

Evaluation Metrics. We use the area under the receiver operating characteristic (AUC) curve on UCF-Crime and MSAD to evaluate performance, as it effectively reflects the balance between true positive rate and false positive rate. For XD-Violence, we employ average precision (AP) to evaluate the balance between precision and recall. We further compute AUC_A and AP_A exclusively on abnormal videos, which better reflect the model's ability to localize anomalies. In addition, we report the false alarm rate (FAR) on normal videos, which measures the proportion of normal frames misclassified as anomalies when the anomaly score threshold is fixed at 0.5, as in [21]. All metrics are reported in percentages (%).

Implementation Details. We implement our model in PyTorch and train it on an NVIDIA L20 GPU using the AdamW optimizer [8] with a learning rate of 1×10^{-3}. The batch size is set to 64 for UCF-Crime and XD-Violence, and 32 for MSAD. All models are trained for 3,000 iterations. Input frames are resized to 224×224 pixels, and we adopt a pre-trained VideoMAEv2 (ViT-G/14) as the backbone token feature encoder. For both the MIL loss $\mathcal{L}_{\text{MIL}}$ and the selection of positive/negative samples in the triplet learning strategy, we set $k = \lfloor L/16 \rfloor + 1$, where L is the number of video segments [33]. In contrast, the anchor selection uses a dataset-specific top-k_a scheme: $k_a = 3$ for UCF-Crime, $k_a = 11$ for XD-Violence, and $k_a = 1$ for MSAD. The loss balancing weight λ is set to 0.1 for all three datasets.

4.2. State-of-the-art Comparison

We evaluate the performance of our TLMA method on three public benchmarks: UCF-Crime, XD-Violence, and MSAD, comparing it against a wide range of state-of-the-art WSVAD approaches. As shown in Table 1, TLMA achieves strong results on both the UCF-Crime and XD-Violence datasets. Specifically, it obtains an AUC of 89.47% and AUC_A of 76.16% on UCF-Crime, outperforming VadCLIP by 1.45% and 5.93%, respectively. These results demonstrate its effectiveness in recognizing diverse anomalous behaviors under complex surveillance conditions. On XD-Violence, TLMA achieves an AP of 86.78% and an AP_A of 86.23%, surpassing the previous best AP (PEL4VAD: 85.59%) by 1.19% and the highest AP_A (UR-DMU: 83.94%) by 2.29%, further validating its strong capability in modeling complex violence under weak supervision. On the recently introduced MSAD dataset, which contains diverse scene contexts and anomaly categories, TLMA exhibits even greater performance advantages. As shown in Table 2, our method achieves an AUC of 93.68% and an AP of 81.3%, showing substantial improvements over recent methods such as PI-VAD and GS-MOE. More importantly, our method yields the highest metrics computed on abnormal videos, with AUC_A of 79.52% and AP_A of 84.42%,

Table 1. Comparison with state-of-the-art WSVAD methods on the UCF-Crime and XD-Violence datasets. The best results are written in **bold**.

Method	Encoder	UCF-Crime		XD-Violence	
		AUC	AUC_A	AP	AP_A
DeepMIL [21]	C3D	75.40	54.25	73.20	78.61
GCN [32]	TSN	82.12	59.02	–	–
HL-Net [27]	I3D	82.44	–	75.41	–
MIST [7]	I3D	82.30	–	–	–
RTFM [22]	I3D	84.30	62.96	77.81	78.57
MSL [9]	V-Swin	85.62	–	78.59	–
MGFN [5]	V-Swin	86.67	–	80.11	–
CUPL [31]	I3D	86.22	–	81.43	–
UMIL [11]	CLIP	86.75	68.68	–	–
NG-MIL [14]	I3D	85.63	–	78.51	–
UR-DMU [33]	I3D	86.97	70.81	81.66	83.94
PEL4VAD [15]	I3D	86.76	–	85.59	–
TPWNG [30]	CLIP	87.79	–	83.86	–
VadCLIP [29]	CLIP	88.02	70.23	84.51	–
TLMA(ours)	MAE	**89.47**	**76.16**	**86.78**	**86.23**

Table 2. Comparison with state-of-the-art WSVAD methods on the MSAD dataset. The best results are written in **bold**.

Method	AUC	AUC_A	AP	AP_A
RTFM [22]	86.65	–	–	–
MGFN [5]	84.96	–	–	–
TEVAD [4]	86.82	–	–	–
UR-DMU [33]	85.02	–	–	–
GS-MOE [1]	87.72	69.54	68.26	76.68
PI-VAD [12]	88.68	71.25	71.26	77.86
TLMA(ours)	**93.68**	**79.52**	**81.3**	**84.42**

Table 3. Ablation study of the contributions of key components on the UCF-Crime and XD-Violence datasets.

Baseline	MA	$\mathcal{L}_{\text{Triplet}}$	UCF-Crime		XD-Violence	
			AUC	FAR ↓	AP	FAR ↓
✓			85.69	3.3	83.61	1.07
✓	✓		86.84	2.62	85.35	1.11
✓		✓	88.37	1.93	84.64	0.96
✓	✓	✓	**89.47**	**1.56**	**86.78**	**0.43**

Table 4. Impact of different anchor-k_a values on performance across datasets.

Anchor-k_a	UCF(AUC)	XD(AP)	MSAD(AUC)
1	89.38	85.08	**93.68**
3	**89.47**	86.13	93.30
5	89.17	86.64	93.32
7	89.26	86.52	93.25
9	89.17	86.50	93.28
11	89.12	**86.78**	93.32
13	88.92	86.74	93.33

demonstrating its superior capability in precise anomaly localization. Overall, TLMA consistently outperforms existing methods across all three benchmarks, confirming the effectiveness and generalizability of the proposed framework. See Appendix for per-category metrics.

4.3. Ablation Experiments

Effect of proposed module. We conduct an ablation study on UCF-Crime and XD-Violence to evaluate the effectiveness of each proposed component, as shown in Table 3. Our baseline MIL model uses a pre-trained VideoMAEv2 (ViT-G/14) as the feature extractor. Segment-level features are obtained by average-pooling the token features within a segment, followed by optimization with the standard MIL ranking loss. Building upon this baseline, incorporating the Motion-Aware (MA) module consistently improves performance. On XD-Violence, the AP increases from 83.61% to 85.35%, indicating that emphasizing motion-related cues aids in distinguishing true anomalies from background interference. The introduction of the Triplet Learning (TL) strategy ($\mathcal{L}_{\text{Triplet}}$) further boosts performance: on UCF-Crime, AUC increases from 85.69% to 88.37%, and on XD-Violence the AP also improves compared to the baseline. This demonstrates that explicitly separating WLI from true anomalies via triplet learning enhances anomaly localization precision. By integrating both MA and the triplet learning, our full TLMA model achieves the best overall results, with the highest AUC on UCF-Crime (89.47%) and the highest AP on XD-Violence (86.78%). Moreover, the FAR steadily decreases as modules are added, dropping from 3.3% to 1.56% on UCF-Crime and from 1.07% to 0.43% on XD-Violence. This trend indicates that our method effectively suppresses misclassifications by mitigating the impact of WLI.

Table 5. Ablation study of anchor selection strategies on UCF-Crime dataset.

	Random	Low-score	High-score
AUC	89.04	89.06	**89.47**

Table 6. Impact of the loss weight λ in the joint objective on detection performance.

λ	UCF (AUC)	XD (AP)	MSAD (AUC)
1	88.42	83.27	93.13
0.1	**89.47**	**86.78**	**93.68**
0.01	88.57	86.01	93.33
0.001	86.88	85.42	93.25

Table 7. Ablation study of edge filters in the Motion-Aware (MA) module on the MSAD dataset.

Metric	Sobel	Scharr	DoG	No filter
AUC	**93.68**	93.31	93.27	93.20
AP	**81.30**	79.82	79.82	80.12

Effect of Anchor Selection. Table 4 presents an analysis of the k_a value, which governs the selection of anchors for the triplet construction. The results demonstrate the robustness of our approach, with stable performance across a range of k_a values on all data—measured by AUC for UCF-Crime and MSAD, and AP for XD-Violence. On UCF-Crime, the highest AUC (89.47%) is achieved when $k_a = 3$, while on XD-Violence, the best AP (86.78%) is obtained with $k_a = 11$. For MSAD, $k = 1$ yields the highest AUC (93.68%). These results indicate that selecting a suitable k_a allows us to better represent diverse WLI segments in the normal videos. Smaller k_a tends to be more effective in simple scenes (e.g., MSAD), while complex datasets (e.g., XD-Violence) may benefit from a larger k_a. To further validate the importance of anchor selection, we conduct an ablation study comparing different strategies on the UCF-Crime dataset, as shown in Table 5. Using high-anomaly-scoring segments from normal videos as anchors yields the best AUC (89.47%), outperforming random (89.04%) and low-anomaly-scoring selection (89.06%). This confirms that anomalous-looking segments in normal videos, despite being labeled as normal, serve as more representative proxies of WLI. They are more likely to be misclassified and thus provide stronger supervision signals for triplet-based learning to distinguish WLI from true anomalies.

Effect of Loss Weight λ. We evaluate the impact of

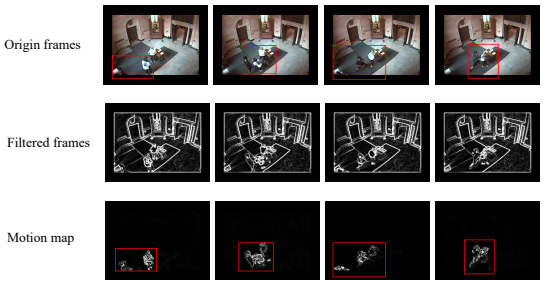


Figure 6. Visualization of the proposed motion-aware feature enhancement module. From top to bottom: originaTLMAl video frames, Sobel-filtered edge responses, and the resulting motion maps. The motion map highlights dynamic foreground regions while suppressing static background content.

the loss weight λ in the joint objective, as shown in Table 6. Setting $\lambda = 0.1$ yields the best results across all datasets (89.47% AUC on UCF-Crime, 86.78% AP on XD-Violence, 93.68% AUC on MSAD). A large λ overemphasizes triplet loss and harms MIL optimization, while a small λ weakens the separation of WLI and true anomalies. Thus, we adopt $\lambda = 0.1$ as the default.

Effect of Edge Filters in MA. To assess the sensitivity of the MA module to different edge extraction strategies, we conduct experiments on the MSAD dataset by replacing the Sobel filter with two alternatives: Scharr [18] and Difference of Gaussian (DoG) [13], and also test a variant without edge filtering. As shown in Table 7, Sobel yields the best results (93.68% AUC, 81.30% AP), outperforming Scharr, DoG, and the no-filter baseline. This indicates that Sobel provides more discriminative motion cues.

4.4. Qualitative Results

Motion Map. To better illustrate the effect of the proposed motion-aware feature enhancement module, we visualize its intermediate outputs in Figure 6. The top row displays the original video frames, the middle row shows Sobel-filtered edge responses, and the bottom row presents the resulting motion maps. As shown, the motion maps clearly emphasize dynamic foreground regions while suppressing static background content. This selective enhancement enables the model to better focus on motion-related cues that are more indicative of abnormal behavior, thereby improving the quality of feature representations used for downstream anomaly detection.

Anomaly Score. Figure 7 presents a qualitative comparison of anomaly score predictions on test videos from the UCF-Crime dataset. The blue curve represents the anomaly scores generated by our proposed TLMA method, while the orange curve corresponds to the baseline method. As shown in Figure 7(c), after the fighting incident ends, people in the video resume normal activity. However, the baseline continues to assign high anomaly scores, resulting in misclassifications caused by WLI. Notably, TLMA detects high

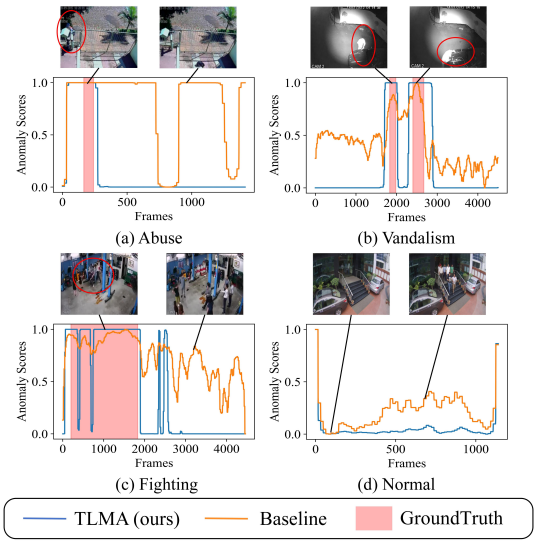


Figure 7. Qualitative results of our method on UCF-Crime test video. Anomaly scores from our TLMA (blue) and the baseline (orange) are shown. The pink region indicates the ground-truth anomaly interval, and the red circles highlight specific occurrences of anomalies.

anomaly scores outside the provided ground-truth. These segments were confirmed to be actual, unannotated fighting events, which further confirms its superior detection sensitivity. Overall, our method produces significantly fewer false positives and aligns more closely with the ground-truth boundaries. It is evident that TLMA yields a sharper and more temporally accurate response within the abnormal interval compared to the baseline. These results demonstrate that TLMA not only mitigates the effect of WLI but also achieves higher anomaly sensitivity and improved temporal localization. Additional qualitative results are provided in the Appendix.

5. Conclusion

In this paper, we propose a novel framework, TLMA, for weakly supervised video anomaly detection, aiming to mitigate the adverse impact of Weakly Labeled Information (WLI) without incurring additional annotation cost. To alleviate frequent misclassifications caused by WLI, we design a triplet learning strategy that increases the feature distance between WLI and truly anomalous patterns, thereby enhancing the model's discriminative capability. In addition, we introduce a motion-aware feature enhancement module to emphasize foreground dynamics and suppress background interference, leading to more robust segment representations. Extensive experiments on three public benchmarks demonstrate that our method achieves state-of-the-art performance while significantly reducing false alarms. These results confirm that TLMA effectively mitigates the influence of WLI and enables more precise and robust anomaly localization under weak supervision.

6. Acknowledgments

This work was partly supported by the National Natural Science Foundation of China under Grant (62436001, 62536001); the National Key Research and Development Program of China under Grant 2024YFE0202900; the Joint Foundation of the Ministry of Education for Innovation team (8091B042235); the Fundamental Research Funds for the Central Universities (No. 2025JBZX064); the National Natural Science Foundation of China (62506028); Postdoctoral Innovation Talent Support Program (K25M200080); and the Talent Fund of Beijing Jiaotong University (2025JBZX029 and No. 2024XKRC090).

References

[1] Giacomo D' Amicantonio, Snehashis Majhi, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, François Brémond, and Egor Bondarev. Mixture of experts guided by gaussian splatters matters: A new approach to weakly-supervised video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10275–10285, 2025. 3, 6

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

[3] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009. 1

[4] Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, and David Aik-Aun Khoo. Tevad: Improved video anomaly detection with captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5559, 2023. 6

[5] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 387–395, 2023. 3, 6

[6] MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun Lee. Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12137–12146, 2023. 3

[7] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14009–14018, 2021. 3, 6

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[9] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1395–1403, 2022. 3, 6

[10] Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, Jun Liu, Azzedine Boukerche, Peng Sun, and Liang Song. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *ACM Computing Surveys*, 56(7):1–38, 2024. 1

[11] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8022–8031, 2023. 3, 6

[12] Snehashis Majhi, Giacomo D'Amicantonio, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Egor Bondarev, and François Brémond. Just dance with pi! a poly-modal inductor for weakly-supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24265–24274, 2025. 3, 6

[13] David Marr and Ellen Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217, 1980. 8

[14] Seongheon Park, Hanjae Kim, Minsu Kim, Dahye Kim, and Kwanghoon Sohn. Normality guided multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2665–2674, 2023. 1, 3, 6

[15] Yujiang Pu, Xiaoyu Wu, Lulu Yang, and Shengjin Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Transactions on Image Processing*, 2024. 3, 6

[16] 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 *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PmLR, 2021. 3

[17] Bharathkumar Ramachandra, Michael J Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2293–2312, 2020. 1

[18] Hanno Scharr. *Optimal operators in digital image processing*. PhD thesis, Ruprecht-Karls-Universität Heidelberg, Germany, 2000. 8

[19] Irwin Sobel, Gary Feldman, et al. A 3x3 isotropic gradient operator for image processing. *a talk at the Stanford Artificial Project in*, 1968:271–272, 1968. 5

[20] Yong Su, Jiahang Li, Simin An, Hengpeng Xu, and Weilong Peng. Dual-detector reoptimization for federated weakly supervised video anomaly detection via adaptive dynamic recursive mapping. *IEEE Transactions on Industrial Informatics*, 2025. 3, 5

[21] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018. 1, 3, 6

[22] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4975–4986, 2021. 3, 6

[23] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 35:10078–10093, 2022. 3

[24] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015. 3

[25] Jinwen Wang, Youfang Lin, Xiaobo Hu, Siyu Yang, Sheng Han, Shuo Wang, and Kai Lv. From pixels to temporal correlations: Learning informative representations for reinforcement learning pre-training. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 679–688, 2025. 3

[26] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 3, 5

[27] 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 *Proceedings of the European Conference on Computer Vision*, pages 322–339. Springer, 2020. 3, 6

[28] Peng Wu, Xuerong Zhou, Guansong Pang, Zhiwei Yang, Qingsen Yan, Peng Wang, and Yanning Zhang. Weakly supervised video anomaly detection and localization with spatio-temporal prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9301–9310, 2024. 3

[29] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6074–6082, 2024. 1, 3, 6

[30] Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18899–18908, 2024. 6

[31] Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16271–16280, 2023. 3, 6

[32] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1237–1246, 2019. 6

[33] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3769–3777, 2023. 1, 2, 3, 5, 6

[34] Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. *Advances in Neural Information Processing Systems*, 37:89943–89977, 2024. 6

[35] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *arXiv preprint*, 2019. 1