DOI: 10.19884/j.1672-5220.202512010

Weakly Supervised Video Anomaly Detection via Anomaly-Guided Learning

WANG Zhiqi[1], TANG Xuesong[1,2*], HAO Kuangrong[1,2]
1. School of Information and Intelligent Science, Donghua University, Shanghai 201620, China
2. Engineering Research Center of Digitized Textile and Fashion Technology, Ministry of Education, Donghua University, Shanghai 201620, China

Abstract: Video anomaly detection (VAD) aims to identify snippets or events that deviate from normal behavior in video sequences. With the growing demand for public safety and surveillance, this field has developed rapidly. In recent years, weakly supervised VAD (WSVAD) has attracted increasing attention. However, its performance remains limited due to the challenges of multiple instance learning (MIL), including high dataset complexity and substantial video noise, which severely hinder model learning. To address these issues, we propose a WSVAD method guided by anomaly representations. Specifically, our method leverages anomaly-related representations to correct biases in the learning process. It extracts key snippet and anomaly category representations through the key snippet learning module and anomaly classification module, respectively. These two types of representations are then embedded via an anomaly-guided attention mechanism. Extensive experiments on the ShanghaiTech and UCF-Crime datasets demonstrate that the proposed method significantly improves detection accuracy and achieves competitive performance.

Keywords: weakly supervised method; video anomaly detection; prompt-enhanced learning; attention mechanism; anomaly-guided learning; key snippet learning
CLC number: TP183 **Document code**: A
Article ID: 1672-5220(2026)04-0083-09

0 Introduction

Video anomaly detection (VAD) refers to detecting snippets or events that deviate from normal behavior in video sequences[1]. Increasing concerns for public safety and surveillance have greatly accelerated the development of this field. VAD is typically studied under three paradigms: supervised, semi-supervised, and weakly supervised. Due to the difficulty of annotating video data and the limited coverage of anomaly categories, supervised methods are severely restricted and usually impractical. Semi-supervised methods, also known as unsupervised methods[2-3], are trained only on normal videos and classify snippets based on their deviation from learned normal patterns. In recent years, reconstruction-based and prediction-based methods have been widely applied in semi-supervised VAD. However, the scarcity of normal video data, the unbounded diversity of normal behavior and the lack of anomaly modeling limit their effectiveness.

In contrast, weakly supervised methods leverage video-level labels and train models on both normal and anomalous videos to perform frame-level classification. Weakly supervised learning has demonstrated significant performance improvements with minimal annotation cost. Current binary-classification weakly supervised VAD (WSVAD) methods can be broadly divided into single-stage methods based on multiple instance learning (MIL) and two-stage methods based on self-training.

However, WSVAD has not yet achieved satisfactory accuracy under the mainstream MIL method. One key reason is that the uncertainty of anomaly occurrence and the diversity of visual manifestations in real-world scenarios introduce substantial complexity into video data. This introduces substantial temporal and visual noise, making the task significantly more challenging. To address this issue, we propose a novel anomaly-guided learning framework for WSVAD, which incorporates prior knowledge to guide the model in better understanding challenging anomalous behaviors. The guiding representations used in our method come from two sources: key snippets and anomaly categories.

The main contributions of this paper are summarized as follows.

1) We propose a key snippet learning module to identify and exploit potential anomalous temporal snippets, providing explicit temporal priors that guide the model to focus on critical moments of anomaly occurrence.

2) We introduce an anomaly classification module that leverages anomaly category descriptions as high-level semantic guidance, enabling effective cross-modal alignment between visual features and text anomaly

Received date: 2025-12-15
Foundation item: National Natural Science Foundation of China (No. 62176052)
* Correspondence should be addressed to TANG Xuesong, email: tangxs@dhu.edu.cn
Citation: Wang Z Q, Tang X S, Hao K R. Weakly supervised video anomaly detection via anomaly-guided learning[J]. *Journal of Donghua University (English Edition)*, 2026, 43(4): 83-91.

semantics.

3) We design an anomaly-guided attention module that integrates key snippet representations and anomaly category representations into a unified attention framework, enabling the model to jointly capture local temporal relevance and global anomaly semantics.

1 Related Work

1.1 Representative methods in WSVAD

Existing WSVAD methods can generally be categorized into MIL-based single-stage methods and two-stage self-training methods. The former directly optimizes anomaly detection models under the MIL framework, whereas the latter iteratively refines pseudo labels through self-training to improve detection performance. To learn anomaly representations under weak supervision, Sultani et al.[4] were the first to adopt an MIL-based method. In MIL, data are organized into bags, where each bag contains multiple instances. A bag is labeled as negative if all instances within it are negative, and as positive if it contains at least one positive instance. Additionally, Sultani et al. introduced the large-scale VAD dataset UCF-Crime and improved the model's ability to distinguish between positive and negative snippets by enhancing the ranking loss function. Tian et al.[5] proposed the robust temporal feature magnitude learning (RTFM) method, which strengthened MIL's ability to identify abnormal snippets through a feature magnitude learning function. It also incorporated extended convolutions and self-attention mechanisms to capture temporal dependencies. Fan et al.[6] introduced an anomaly attention mechanism for WSVAD, optimizing abnormal feature learning through snippet-level attention and multi-branch supervision modules without relying on pseudo-label supervision. Lv et al.[7] presented the unbiased multiple instance learning (UMIL) method, which removed context biases to learn unbiased anomaly features, significantly enhancing WSVAD performance.

Apart from MIL-based single-stage methods, some studies adopt two-stage self-training methods. Feng et al.[8] proposed the multi-instance self-training (MIST) method, which generated reliable snippet-level pseudo-labels through a sparse continuous sampling method. It further combined self-guided attention to enhance the feature encoder, enabling automatic focus on abnormal regions to extract relevant features. Li et al.[9] introduced the multi-sequence learning (MSL) method and hinge-based MSL ranking loss, leveraging Transformer networks to learn video-level anomaly probabilities and snippet-level anomaly scores, which were iteratively refined through a self-training method. Zhang et al.[10] proposed a two-stage self-training method by utilizing pseudo-label completeness and uncertainty. This method incorporated a multi-head classification module and an iterative pseudo-label optimization method, enabling the coverage of more abnormal events and the progressive refinement of pseudo-labels.

1.2 Prompt-enhanced methods in WSVAD

Recent advances in WSVAD have increasingly explored prompt-enhanced paradigms that inject auxiliary semantic guidance into model learning. Text prompts are most widely adopted due to their flexibility and low deployment cost.

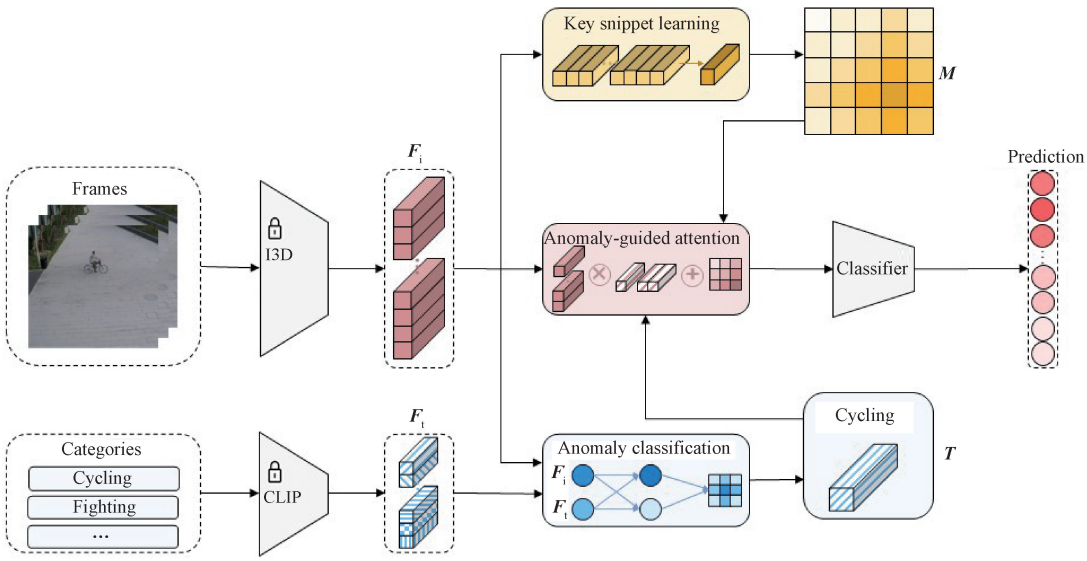
To enrich the semantic expressiveness of text prompts, Pu et al.[11] have leveraged external knowledge bases to construct prompts by aggregating concepts across multiple relationships. While such designs introduce structured semantic priors, their reliance on predefined knowledge sources inevitably limits semantic coverage, potentially leading to biased prompt representations and unclear performance gains. Alternatively, an increasing number of studies[12-14] replaced handcrafted or knowledge-based prompts with learnable prompt embeddings, enabling adaptive alignment between text representations and specific datasets. For example, VadCLIP[12] explores a vision-language formulation by computing anomaly scores based on the similarity between video snippets and anomaly categories, which differs from MIL-based methods. Although this method improved data-dependent compatibility, the resulting prompts were often difficult to interpret and lacked explicit semantic grounding. More recently, Zanella et al.[15] employed vision-language models to generate frame-level text descriptions as prompts, thereby capturing fine-grained visual semantics. However, generating text annotations for individual frames incurs substantial computational overhead and limits scalability in long video sequences.

In contrast to existing prompt-enhanced WSVAD methods, anomaly-guided learning does not treat anomaly-related text representations as predefined or externally constructed prompts. Instead, the proposed method predicts the anomaly category of each video during training, and the corresponding category-level text features are directly incorporated into the inference process to guide anomaly detection. Furthermore, to better exploit the temporal nature of video data, we introduce an additional positional prompt stream by learning auxiliary anomaly-related binary predictions during training, which complements anomaly detection with temporal localization cues.

2 Anomaly-Guided Learning

Figure 1 illustrates an overview of the proposed method. We employ standard Inflated 3D ConvNet (I3D)[16] and contrastive language-image pre-training (CLIP)[17] models to extract video and text features, respectively, which are then fed into our framework. The proposed model consists of three main components: the key snippet learning module that provides critical anomalous snippets, the anomaly classification module that supplies anomaly category representations, and the

main backbone equipped with anomaly-guided attention for anomaly detection. During training, the anomaly classification module is supervised using ground-truth video-level anomaly categories, which enables the model to learn category-aware text representations and their alignment with snippet features. During inference, the anomaly category is first predicted by the anomaly classification module, and the corresponding category-level text feature is subsequently incorporated into the anomaly-guided attention module to guide anomaly detection. This design ensures that the proposed framework does not rely on ground-truth category representations at inference time and remains consistent with the weakly supervised setting.



$\boldsymbol{F}_i$ —snippet feature; $\boldsymbol{F}_t$ —text feature; $\boldsymbol{M}$ —temporal matrix based on indices of key snippets; $\boldsymbol{T}$ —category text feature obtained by classification.

Fig. 1 Overview of proposed method

2.1 Key snippet learning

To provide the model with key snippet representations, we design a module that remains as lightweight as possible while selecting the most likely anomalous snippets from each anomalous video. During the key snippet learning process, the snippet feature $\boldsymbol{F}_i$ is fed into a two-layer feed-forward neural network and a classifier to generate a coarse prediction $\hat{\boldsymbol{p}}$. The detailed formulation of the process is presented as

$$\hat{\boldsymbol{p}} = f_{\text{cls}}(f_{\text{mlp}}(\boldsymbol{F}_i)), \tag{1}$$

where $f_{\text{mlp}}(\cdot)$ denotes the function of the two-layer feed-forward neural network; $f_{\text{cls}}(\cdot)$ represents the function of the classifier.

Instead of aggregating multiple top-k snippets to compute a video-level prediction, we intentionally select the top-1 scoring snippet from each video during training, which enforces more targeted supervision for learning discriminative key snippets, encouraging the module to focus on the most salient temporal snippet associated with anomalous patterns. This process is denoted as

$$\hat{p} = \begin{cases} \dfrac{1}{k}\displaystyle\sum_{i\in\mathcal{T}_k}(p_i,\ y) = 1 \ (\text{if anomalous video}), \\ \max\limits_{1\leqslant i\leqslant N}(p_i,\ y) = 0 \ (\text{if normal video}), \end{cases} \tag{2}$$

$$\mathcal{L}_{\text{BCE2}}(\hat{\boldsymbol{p}},\ \boldsymbol{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \ln(\hat{p}_i) + (1-y_i)\cdot\ln(1-\hat{p}_i)], \tag{3}$$

where $p_i \in [0,1]$ is the prediction of the ith snippet; N is the total number of snippets in a video; $y \in \{0,1\}$ is the ground truth label; $\hat{p}$ is the aggregated video-level prediction obtained from the snippet-level scores; k is the number of selected top-scoring snippets; $\mathcal{T}_k$ is the index set of the top-k snippets with the highest anomaly scores. The maximum value in $\hat{\boldsymbol{p}}$ corresponds to the most likely anomalous snippet, whose temporal position is then used to provide effective key anomalous snippet representations to the model.

2.2 Anomaly classification

The anomaly classification module is trained using video-level anomaly category labels, which are available in standard WSVAD benchmarks. The objective of this module is to learn high-level semantic representations associated with anomaly categories. These representations are subsequently encoded as category-aware text features. These features are used as auxiliary semantic guidance within the attention mechanism to enhance feature discrimination.

Based on the learned category-aware representations, we quantify the correlation between videos and categories through cross-modal interactive feature fusion. Specifically, pairs of $\boldsymbol{F}_i$ and $\boldsymbol{F}_t$ are alternately used as the query and the key-value pairs in cross-modal attention $\text{CA}(\cdot)$ to achieve alignment. Afterward, the two aligned feature sets are multiplied to generate a similarity matrix

$\boldsymbol{S}$, thereby constructing the cross-modal relationship between videos and categories. This process is denoted as

$$\boldsymbol{C} = \mathrm{Norm}(\mathrm{CA}(\boldsymbol{F}_{\mathrm{t}}, \boldsymbol{F}_{\mathrm{i}}) + \boldsymbol{F}_{\mathrm{t}}), \tag{4}$$

$$\boldsymbol{V} = \mathrm{FC}(\mathrm{Norm}(\mathrm{CA}(\boldsymbol{F}_{\mathrm{i}}, \boldsymbol{F}_{\mathrm{t}}) + \boldsymbol{F}_{\mathrm{i}})), \tag{5}$$

$$\boldsymbol{S} = \boldsymbol{C} \times \boldsymbol{V}, \tag{6}$$

where $\boldsymbol{C}$ is the category-aware cross-modal representation obtained by attending $\boldsymbol{F}_{\mathrm{t}}$ to $\boldsymbol{F}_{\mathrm{i}}$; $\boldsymbol{V}$ is the snippet-aware cross-modal representation obtained by attending $\boldsymbol{F}_{\mathrm{i}}$ to $\boldsymbol{F}_{\mathrm{t}}$; $\boldsymbol{S}$ is the resulting similarity matrix that captures the cross-modal relationship between video snippets and anomaly categories. Both $\boldsymbol{F}_{\mathrm{i}}$ and $\boldsymbol{F}_{\mathrm{t}}$ are L2-normalized before similarity computation. The similarity is implemented as cosine similarity, ensuring that all similarity scores are bounded within a consistent range.

To train this module, we adopt a method based on the maximum similarity among anomaly categories, allowing the model to focus on inter-category distinctions. At each training stage, we extract the maximum similarity value of each anomaly category from the similarity matrix and convert the fine-grained video annotations into binary labels via one-hot encoding: the correct category j is assigned a label of 1, while all others are assigned 0. This design ensures that the module emphasizes the strongest association between each video snippet and anomaly category, thereby guiding the learning of discriminative features. We employ the L1 loss as the optimization objective, as it effectively measures the similarity error between video snippets and anomaly categories while maintaining robustness in multi-category settings. Since both $\boldsymbol{F}_{\mathrm{i}}$ and $\boldsymbol{F}_{\mathrm{t}}$ are L2-normalized and the cosine similarity is adopted, the similarity values are naturally bounded within a fixed range, which avoids scale ambiguity and ensures stable optimization when using L1 loss. Compared with classification-based losses, L1 loss allows the model to focus on relative similarity alignment without enforcing overly confident predictions at early training stages. The overall loss calculation method and the L1 loss function are defined as

$$\boldsymbol{p} = [\max(S_{1,1}, S_{1,2}, \cdots, S_{1,N}), \max(S_{2,1}, S_{2,2}, \cdots, S_{2,N}), \cdots, \max(S_{J,1}, S_{J,2}, \cdots, S_{J,N})], \tag{7}$$

$$\boldsymbol{y} = \mathrm{OneHot}(j, J), \tag{8}$$

$$\mathcal{L}_{l1}(\boldsymbol{y}, \boldsymbol{p}) = \frac{1}{K}\sum_{i=1}^{K} |y_i - p_i|, \tag{9}$$

where $\boldsymbol{p}$ is the predicted anomaly-category probability vector; $S_{J,N}$ is the element at the Jth row and Nth column of the similarity matrix; $\boldsymbol{y}$ is the one-hot encoded ground-truth category label generated by the OneHot $(\cdot)$ operation; K is the number of categories. The anomaly classification module outputs the anomaly category corresponding to each video, and the associated features $\boldsymbol{T}$ are used as text representations to guide the model.

2.3 Anomaly-guided attention

The attention mechanism[18] can adaptively assign weights to different video snippets and effectively model long-range temporal dependencies, making it well-suited for WSVAD. Moreover, it enables efficient alignment and fusion of cross-modal features, which motivates us to embed both key snippet and anomaly category representations into the attention framework.

For the key snippet, we generate a temporal index matrix $\boldsymbol{M}$. The indexing formulation is

$$m_{i,j} = \frac{1}{1 + |w|i-j|^2 - b|}, \tag{10}$$

where i and j denote the temporal indices within the video; $|i-j|$ represents the absolute distance between each pair of temporal positions; $m_{i,j}$ is the temporal prior weight that measures the relative importance between the ith and jth temporal positions; w is a position weight that controls the sensitivity of the temporal distance; b is a bias used to adjust the peak position of the temporal weighting function.

The temporal index matrix $\boldsymbol{M}$ is designed to encode relative temporal relevance with respect to the key anomalous snippet. Since the absolute temporal position of anomalies varies significantly across videos, we adopt a distance-based formulation to model transferable temporal relationships within each video. The linear form provides a general temporal prior, softly emphasizing snippets closer to the anomaly center without imposing strong assumptions on anomaly duration or continuity. The position weight w and bias b are used to adjust the strength of this prior when embedded into the attention mechanism.

$\boldsymbol{M}$ is further integrated into the attention mechanism to explicitly modulate the interaction between temporal snippets. Specifically, $\boldsymbol{M}$ is embedded as a temporal prior into the attention score matrix computed from the query and key. This embedding does not alter the original feature distribution but reflects local anomalous representations. In contrast, category guidance provides global semantic representations, aiming to measure anomaly correlations at a global scale. To enable effective interaction among $\boldsymbol{M}$, $\boldsymbol{T}$, and $\boldsymbol{F}_{\mathrm{i}}$, we align $\boldsymbol{T}$ with $\boldsymbol{F}_{\mathrm{i}}$ along the temporal dimension to obtain $\boldsymbol{T}_{\mathrm{exp}}$. $\boldsymbol{T}_{\mathrm{exp}}$ serves as the query, which interacts with the original snippet features (key) to adjust the attention regions over the video content. The original snippet features are then updated through the attention layer embedded with anomaly-guided representations, yielding new features $\boldsymbol{F}_{\mathrm{a}}$, which are subsequently fed into the classifier for anomaly prediction. The process of this attention mechanism is formulated as

$$\boldsymbol{A} = \mathrm{softmax}\left(\frac{f_{\mathrm{q}}(\boldsymbol{T}_{\mathrm{exp}}) \cdot f_{\mathrm{k}}(\boldsymbol{F}_{\mathrm{i}})^{\mathrm{T}}}{D_{\mathrm{h}}} + \boldsymbol{M}\right), \tag{11}$$

$$\boldsymbol{F}_{\mathrm{a}} = \mathrm{Norm}(\boldsymbol{A} \cdot f_{\mathrm{v}}(\boldsymbol{F}_{\mathrm{i}}) + \boldsymbol{F}_{\mathrm{i}}), \tag{12}$$

where f_{q}, f_{k} and f_{v} represent the three distinct linear layers; D_{h} is the dimension of the hidden layer.

Subsequently, $\boldsymbol{F}_a$ is passed through a two-layer feed-forward neural network and a classifier to produce the predictions. Then, the predictions are trained with the same MIL-based loss. The training of the proposed framework follows the standard MIL paradigm commonly adopted in WSVAD, which is consistent with the learning strategy used in the key snippet learning module.

In summary, during the training phase, the overall objective function of our model is denoted as

$$\mathcal{L} = \mathcal{L}_{\mathrm{BCE1}} + \alpha \mathcal{L}_{\mathrm{BCE2}} + \beta \mathcal{L}_{l1}, \quad (13)$$

where α and β are coefficients to adjust the loss function; $\mathcal{L}_{\mathrm{BCE1}}$ represents the loss of the final predicted values; $\mathcal{L}_{\mathrm{BCE2}}$ represents the loss of the coarse prediction.

3 Experiments

3.1 Experimental setting

Datasets. We conduct extensive evaluations of the proposed method on two widely used public benchmark datasets: ShanghaiTech[19] and UCF-Crime[20]. UCF-Crime consists of 1 900 long untrimmed surveillance videos with a total duration of 128 h. It covers 13 real-world anomaly categories, comprising abuse, arrest, arson, assault, road accident, burglary, explosion, fighting, robbery, shooting, stealing, shoplifting, and vandalism. The training set contains 800 normal videos and 810 anomalous videos, while the test set contains 150 normal videos and 140 anomalous videos. ShanghaiTech includes 437 videos from 13 campus scenes. The original version is designed for semi-supervised anomaly detection, containing 330 normal videos and 107 anomalous videos. The dataset is later reorganized into a weakly supervised setting, consisting of 238 training videos and 199 test videos[21]. For fair comparison, we follow prior WSVAD work and adopt the frame-level area under the curve (AUC) as the evaluation metric. A higher AUC indicates better VAD performance.

Implementation details. Following Ref. [21], we employ the I3D network pretrained on Kinetics-400 for snippet feature extraction. To ensure fair comparison, we adopt the 10-crop augmentation method, which includes the center crop, four corners, and their horizontal flips. For category text feature extraction, we use the pretrained CLIP model to obtain the corresponding representations. For hyperparameters, the hidden dimension of all modules is set to 128. The two-layer feed-forward neural network consists of two fully connected layers with 512 and 300 output units, respectively. All dropout rates are set to 0.1. For both ShanghaiTech and UCF-Crime, w is set to 0.6, b is set to 0.2, α is set to 5, and β is set to 15. During training, the model is optimized using the Adam optimizer[22] with a batch size of 128 for 50 epochs. The initial learning rate is set to 5×10^{-4}, and a cosine decay schedule is applied.

3.2 Comparison with state-of-the-art methods

Table 1 presents the comparison between our proposed method and other state-of-the-art methods on both datasets. On ShanghaiTech, Su et al.[23] achieved an AUC of 96.88%, while on UCF-Crime, Thakare et al.[24] achieved an AUC of 83.56%. Moreover, AUCs of 95.00% and 84.05% were achieved by the recent BE-WSVAD method on ShanghaiTech and UCF-Crime, respectively. In contrast, AUCs of 97.85% and 84.97% were achieved by our proposed method on ShanghaiTech and UCF-Crime, respectively, demonstrating its superior performance in WSVAD tasks.

Table 1 Performance comparison with state-of-the-art methods on ShanghaiTech and UCF-Crime

Method	AUC/%	
	ShanghaiTech	UCF-Crime
GCN[20]	84.44	82.12
CLAWS[25]	89.67	83.03
Contrastive attention[26]	92.25	84.62
MIST[8]	94.83	82.30
RTFM[5]	97.21	84.30
CRFD[21]	97.48	84.89
BN-SVP[27]	96.00	83.39
BE-WSVAD[28]	95.00	84.05
VPE-WSVAD[23]	96.88	—
PEL[11]	97.82	—
RareAnom[24]	—	83.56
LAVAD[15]	—	80.28
Ours	**97.85**	**84.97**

3.3 Ablation study

Multiple ablation experiments are considered to validate the effectiveness of each component in our model. Table 2 reports the impact of different modules. The baseline consists of a single self-attention layer, a two-layer feed-forward neural network, and a classifier. Incorporating the key snippet representations yields an absolute improvement of 1.35% and 1.51% on ShanghaiTech and UCF-Crime, respectively. Further introducing the anomaly category representations brings an additional gain of 1.27% and 2.12%, respectively. These results confirm that both types of guidance contribute effectively to enhancing the model's performance.

Table 2 Ablation results on ShanghaiTech and UCF-Crime

Anomaly-guided attention	Key snippet	Anomaly category	AUC/%	
			ShanghaiTech	UCF-Crime
√	×	×	95.23	81.34
√	√	×	96.58	82.85
√	√	√	97.85	84.97

3.4 Sensitivity analysis of temporal prior parameters

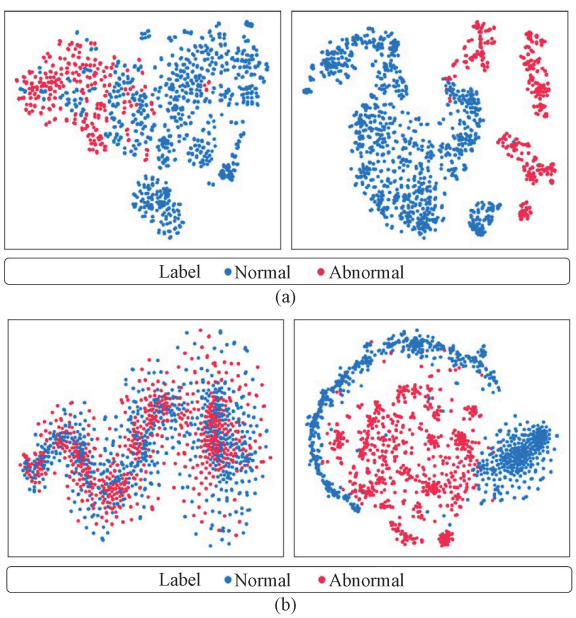
The sensitivity of the proposed model is analyzed to w and b by varying them within reasonable ranges on the ShanghaiTech and UCF-Crime datasets. The results are illustrated in Table 3. The performance across different parameter settings remains stable, suggesting that the proposed temporal index matrix is robust and does not rely on precise hyperparameter tuning.

Table 3 Sensitivity analysis of temporal prior parameters on ShanghaiTech and UCF-Crime

w	b	AUC/%	
		ShanghaiTech	UCF-Crime
0.2	0.2	97.41	84.52
0.4	0.2	97.68	84.81
0.6	**0.2**	**97.85**	**84.63**
0.8	0.2	97.85	84.74
0.6	0.0	97.62	84.63
0.6	0.4	97.57	84.69

3.5 Qualitative results

To evaluate the detection capability of the anomaly-guided attention module and the classification capability of the anomaly classification module, we visualize the features before and after each module using t-distributed stochastic neighbor embedding (t-SNE) at both coarse-grained and fine-grained levels, as shown in Fig. 2. Figures 2(a) and 2(b) illustrate the inter-category feature distribution before and after anomaly-guided attention, while Figs. 2(c) and 2(d) present the intra-category feature distribution of anomalies before and after anomaly classification. These results demonstrate that the two modules learn complementary discrimination capabilities: the anomaly-guided attention module focuses on separating anomalous events from normal ones, while the anomaly classification module captures discriminative representations among different anomaly categories.



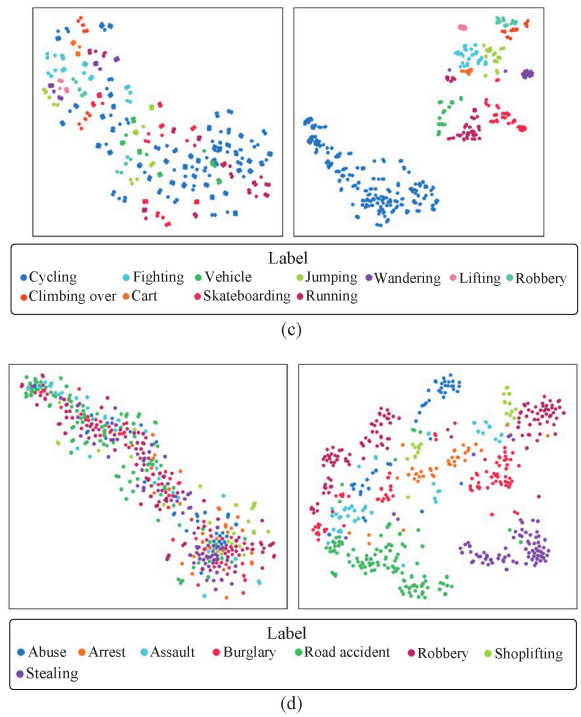


Fig. 2 Distribution of discriminative features using t-SNE: (a) inter-category feature distribution before anomaly-guided attention; (b) inter-category feature distribution after anomaly-guided attention; (c) intra-category feature distribution of anomalies before anomaly classification; (d) intra-category feature distribution of anomalies after anomaly classification

Figure 3 further presents the predicted anomaly scores of the anomaly-guided learning framework on representative samples from the ShanghaiTech and UCF-Crime datasets. The predicted anomaly score curves exhibit clear and consistent responses that are well aligned with the ground-truth anomalous snippets, while remaining relatively low and stable in normal intervals. This indicates that the anomaly-guided learning method can effectively distinguish anomalous events from background activities at the snippet level. Notably, even in complex and long-duration videos from UCF-Crime, the anomaly-guided learning method maintains reliable temporal localization of anomalies, demonstrating its robustness and strong temporal discrimination capability.

However, the predicted anomaly scores in certain surrounding normal regions do not strictly collapse to zero, as observed in Fig. 3(a). This indicates that the model may produce elevated responses in temporally adjacent normal snippets, leading to false positives and imprecise anomaly boundaries. Such behavior is a common limitation in WSVAD, where the absence of frame-level supervision makes it challenging to enforce sharp temporal localization. Therefore, addressing more precise temporal boundary modeling and reducing false alarms in normal regions remains an important direction in future research.

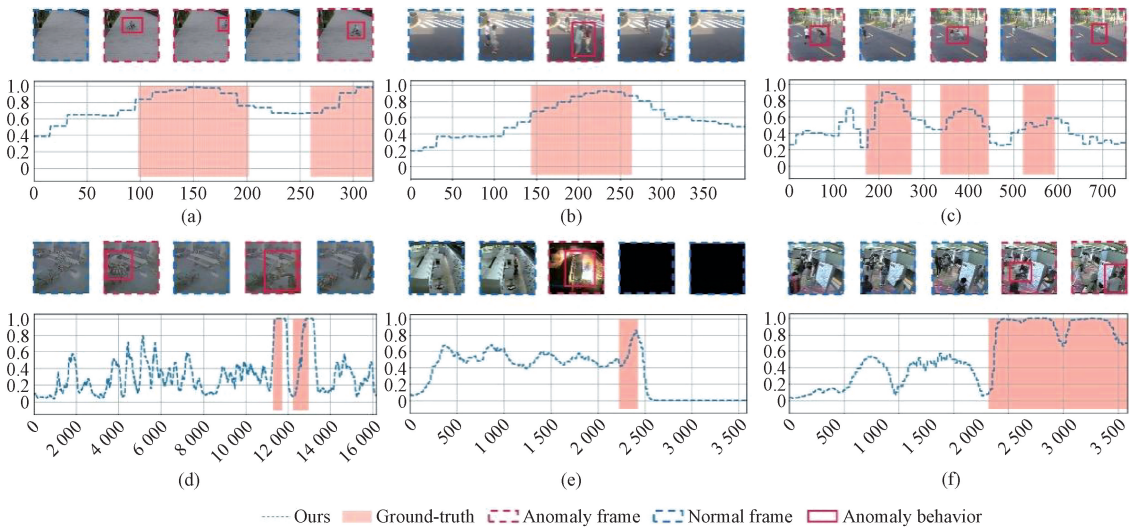


Fig. 3 Qualitative results of anomaly detection performance on ShanghaiTech and UCF-Crime: (a) example of video ID 010051 in ShanghaiTech; (b) example of video ID 030032 in ShanghaiTech; (c) example of video ID 050023 in ShanghaiTech; (d) example of video ID Assault010 in UCF-Crime; (e) example of video ID Explosion022 in UCF-Crime; (f) example of video ID Fighting003 in UCF-Crime (*x*-axis denotes frame index and *y*-axis represents predicted anomaly score)

4 Conclusions

In this paper, we propose an anomaly-guided learning method for WSVAD. Specifically, we leverage a key snippet learning module to capture critical anomalous snippets and anomaly classification to obtain text representations of anomaly categories. These two sources of guidance are then integrated through an anomaly-guided attention module, enabling the model to achieve a more precise and comprehensive understanding of anomalies from both global and local perspectives. Extensive experiments on the ShanghaiTech and UCF-Crime datasets validate the effectiveness of the proposed guidance and its fusion mechanism. In future work, we plan to explore more effective anomaly-guided methods and deeper cross-modal interaction mechanisms to further enhance WSVAD.

References

[1] Ramachandra B, Jones M J, Vatsavai R R. A survey of single-scene video anomaly detection [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(5): 2293-2312.

[2] Pang G S, Shen C H, Cao L B, et al. Deep learning for anomaly detection: a review [J]. *ACM Computing Surveys*, 2022, 54(2): 1-38.

[3] Yang X Y, Yang Y Z, Deng H P. An unsupervised online detection method for foreign objects in complex environments[J]. *Journal of Donghua University (English Edition)*, 2026, 43(1): 140-151.

[4] Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6479-6488.

[5] Tian Y, Pang G S, Chen Y H, et al. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2022: 4955-4966.

[6] Fan Y D, Yu Y X, Lu W H, et al. Weakly-supervised video anomaly detection with snippet anomalous attention [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, 34(7): 5480-5492.

[7] Lv H, Yue Z Q, Sun Q R, et al. Unbiased multiple instance learning for weakly supervised video anomaly detection [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2023: 8022-8031.

[8] Feng J C, Hong F T, Zheng W S. MIST: multiple instance self-training framework for video anomaly detection [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2021: 14004-14013.

[9] Li S, Liu F, Jiao L C. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 36(2): 1395-1403.

[10] Zhang C, Li G R, Qi Y K, et al. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection [C]//2023 IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2023: 16271-16280.
[11] Pu Y J, Wu X Y, Yang L L, et al. Learning prompt-enhanced context features for weakly-supervised video anomaly detection [J]. *IEEE Transactions on Image Processing*, 2024, 33: 4923-4936.
[12] Wu P, Zhou X R, Pang G S, et al. VadCLIP: adapting vision-language models for weakly supervised video anomaly detection [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 38(6): 6074-6082.
[13] Yang Z W, Liu J, Wu P. Text prompt with normality guidance for weakly supervised video anomaly detection [C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2024: 18899-18908.
[14] Chen J X, Li L, Su L, et al. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection [C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2024: 18319-18329.
[15] Zanella L, Menapace W, Mancini M, et al. Harnessing large language models for training-free video anomaly detection [C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2024: 18527-18536.
[16] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 4724-4733.
[17] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [PP/OL]. arXiv (2021-02-26) [2025-11-10]. https://arxiv.org/pdf/2103.00020.
[18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. Long Beach, CA, USA: Curran Associates, Inc., 2017: 5998-6008.
[19] Luo W X, Liu W, Gao S H. A revisit of sparse coding based anomaly detection in stacked RNN framework [C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017: 341-349.
[20] Zhong J X, Li N N, Kong W J, et al. Graph convolutional label noise cleaner: train a plug-and-play action classifier for anomaly detection [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2020: 1237-1246.
[21] Wu P, Liu J. Learning causal temporal relation and feature discrimination for anomaly detection [J]. *IEEE Transactions on Image Processing*, 2021, 30: 3513-3527.
[22] Kingma D P, Ba J. Adam: a method for stochastic optimization [PP/OL]. arXiv (2017-01-30) [2025-11-10]. https://arxiv.org/pdf/1412.6980.
[23] Su Y, Tan Y Y, Xing M, et al. VPE-WSVAD: Visual prompt exemplars for weakly-supervised video anomaly detection [J]. *Knowledge-Based Systems*, 2024, 299: 111978.
[24] Thakare K V, Dogra D P, Choi H, et al. RareAnom: a benchmark video dataset for rare type anomalies [J]. *Pattern Recognition*, 2023, 140: 109567.
[25] Zaheer M Z, Mahmood A, Astrid M, et al. CLAWS: clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection [C]//Computer Vision-ECCV 2020. Cham: Springer, 2020: 358-376.
[26] Chang S N, Li Y C, Shen S M, et al. Contrastive attention for video anomaly detection [J]. *IEEE Transactions on Multimedia*, 2022, 24: 4067-4076.
[27] Sapkota H, Yu Q. Bayesian nonparametric submodular video partition for robust anomaly detection [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2022: 3202-3211.
[28] Yang Z, Guo Y F, Wang J F, et al. Towards video anomaly detection in the real world: a binarization embedded weakly-supervised network [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, 34(5): 4135-4140.

基于异常引导学习的弱监督视频异常检测

王执祺[1]，唐雪嵩[1,2*]，郝矿荣[1,2]

1. 东华大学 信息与智能科学学院，上海 201620

2. 东华大学 数字化纺织服装技术教育部工程研究中心，上海 201620

摘　要：视频异常检测（video anomaly detection，VAD）旨在识别视频序列中偏离正常行为模式的异常片段或事件。随着公共安全与智能监控需求的不断增长，该领域得到了快速发展。近年来，弱监督视频异常检测（weakly supervised video anomaly detection，WSVAD）因其较低的标注成本而受到广泛关注。然而，由于多实例学习（multiple instance learning，MIL）框架面临数据集复杂度高和视频噪声干扰严重等挑战，其检测性能仍受到较大限制，这些问题显著阻碍了模型对异常模式的有效学习。为解决上述问题，本文提出了一种基于异常引导学习的弱监督视频异常检测方法。具体而言，该方法利用异常相关表征对模型学习过程中的偏差进行校正。首先，通过关键片段学习模块和异常分类模块分别提取关键片段表征与异常类别表征；其次，设计异常引导注意力机制对两类表征进行嵌入与融合，从而增强模型对异常语义信息的感知能力。在 ShanghaiTech 和 UCF-Crime 数据集上的大量实验结果表明，所提出的方法能够显著提升异常检测精度，并取得具有竞争力的性能表现。

关键词：弱监督方法；视频异常检测；提示增强学习；注意力机制；异常引导学习；关键片段学习