Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 21, No. 1, 2026

A Multimodal Video Anomaly Detection Method Based on Vision-Language Alignment

Yueai Zhao[1, 2 *], Yan Zhang [1, 2], Shihao Wang [1, 2], Lipei Kong[1, 2]

[1] School of Computer Science and Technology, Taiyuan Normal University, Jinzhong 030602, China
[2] Shanxi Key Laboratory of Intelligent Optimization Computing and Blockchain Technology, Jinzhong 030619, China

Abstract: Existing weakly supervised anomaly detection methods suffer from insufficient semantic alignment of features and a lack of fine-grained localisation capabilities. Furthermore, purely visual models have limitations in understanding semantic information, making it difficult to effectively integrate visual features with semantic information, which results in poor discrimination of multi-class anomalies. To address these issues, this paper proposes a vision-language aligned video anomaly detection model. By introducing a vision-language alignment branch and utilising a pre-trained CLIP model to achieve fine-grained semantic alignment between video features and category text prompts, this model overcomes the limitations of purely visual models in understanding semantic information.It designs an adaptive gated fusion mechanism to dynamically fuse the global anomaly scores from the original visual branch with the semantically guided scores from the alignment branch, combining the complementary strengths of visual pattern recognition and semantic understanding to enhance the model's ability to distinguish between multiple anomaly categories.And it constructs a multi-task loss function to jointly optimise temporal localisation and fine-grained classification tasks, making full use of video-level weak supervision signals and cross-modal alignment information. Experimental results on the UCF-Crime and XD-Violence datasets demonstrate that this method effectively improves fine-grained anomaly localisation and classification performance, exhibiting significant advantages.

Keywords: Weakly supervised learning; Vision-language alignment; Text prompts; Fine-grained semantics; Multi-task loss.

1. Introduction

With the rapid development of smart cities and intelligent security systems, video surveillance has been widely deployed due to its advantages of being intuitive, information-rich and highly real-time, becoming a key pillar in maintaining public safety, traffic order and operational efficiency [1]. However, most existing surveillance systems are limited to data collection and storage, lacking the capability for intelligent understanding and analysis of video content; the identification of anomalous events remains heavily reliant on manual intervention. Manual detection suffers from low efficiency, high costs and a high error rate, and is heavily influenced by subjective judgement, making it difficult to meet the demands for real-time performance and accuracy posed by vast amounts of surveillance data[2].Consequently, video anomaly detection (VAD) technology based on intelligent computing has emerged and has become a major research direction in the field of computer vision.

Video Anomaly Detection (VAD) is one of the key technologies for addressing the challenges posed by complex spatiotemporal data. This technology aims to automatically identify anomalous events that deviate from normal behaviour within video sequences; such events typically manifest as atypical visual or motion features, or as typical behaviours occurring in inappropriate spatiotemporal contexts [3].

Existing video anomaly detection methods can be broadly categorised into two types: those based on traditional techniques and those based on deep learning. Video anomaly detection methods based on traditional techniques form the early foundation of this field. They primarily adopt unsupervised or weakly supervised approaches, relying on manually extracted low-level features such as optical flow, trajectories and texture, and identifying anomalous samples that deviate from the pattern by modelling normal behaviour patterns [4]. These methods primarily include clustering-based [5–8], trajectory analysis-based [9–10], optical flow-based [11–12], and hybrid frameworks combining multiple features [13], offering advantages such as high computational efficiency, minimal reliance on labelled data, and good practicality in simple static scenes. However, due to their over-reliance on manually crafted features, they are sensitive to noise, changes in lighting, dynamic backgrounds and occlusions, and exhibit weak generalisation capabilities, making it difficult to handle semantic-level anomalies and multi-object spatiotemporal interactions in complex scenes.

Deep learning-based video anomaly detection methods, leveraging end-to-end learning capabilities and the advantages of deep feature extraction, have achieved performance breakthroughs on mainstream datasets; currently, the integration of large-scale model technologies has further raised the upper limit of performance. These methods are primarily categorised into four paradigms: supervised learning [14–15], unsupervised learning [16–17], self-supervised learning [18–19], and weakly supervised learning [20–24]. Supervised learning requires a large amount of labelled data and is constrained by high annotation costs and poor generalisation. Unsupervised learning requires no labelled data but has limited capacity to model the distribution of normal samples and complex spatiotemporal relationships. Self-supervised learning learns representations through proxy tasks and can uncover complex spatiotemporal patterns, but model design and training are relatively challenging. Weakly supervised learning requires only video-level labels, combining the advantages of low annotation costs with clear semantic guidance. Relying on techniques such as multi-instance learning, attention mechanisms and prompt

learning,it demonstrates good generalisation capabilities in real-world scenarios.

Although existing deep learning methods have made significant progress, numerous shortcomings and challenges remain.Current weakly supervised video anomaly detection methods generally suffer from insufficient semantic alignment of features and inadequate fine-grained localisation capabilities;most methods still focus on coarse-grained video-level anomaly classification,with insufficient modelling of complex semantic contexts and long-term temporal dependencies, making it difficult to accurately identify small-scale, fine-grained local anomalies;Purely visual models have inherent limitations in understanding semantic information and struggle to effectively integrate visual features with high-level semantic information.Consequently, their ability to distinguish and classify multi-class anomalies is limited,and both anomaly localisation accuracy and multi-class anomaly discrimination capabilities require further improvement.

To address these issues,this paper proposes a multimodal video anomaly detection model based on vision-language alignment. Firstly,a vision-language alignment branch is designed, utilising a pre-trained CLIP model to perform fine-grained semantic matching between video frame features and category text prompts, thereby achieving cross-modal semantic enhancement. A dual-branch adaptive fusion mechanism is proposed, which dynamically aggregates the global anomaly scores from the raw visual branch and the semantically guided scores from the vision-language alignment branch via learnable gates, thereby effectively combining the complementary strengths of visual pattern recognition and linguistic semantic understanding. A multi-task loss function based on alignment loss and fine-grained classification loss is introduced, enabling the model to distinguish between various anomaly categories.

2. Proposed method

In the task of weakly supervised video anomaly detection,detection and localisation are typically performed by analysing differences in the spatio-temporal patterns of normal and anomalous behaviours within a video. This is achieved primarily by extracting spatio-temporal, motion, appearance, and contextual semantic features from video sequences, learning the distribution patterns of normal behaviours and the discriminative patterns of anomalous events, and utilising the modelling capabilities of deep learning for spatio-temporal information and visual semantics to identify and localise anomalies in videos. To improve the accuracy of anomaly detection and fine-grained localisation,it is necessary to enhance the model's ability to capture spatio-temporal dynamic features, whilst simultaneously strengthening the alignment and fusion of visual features with semantic information.

2.1. Overall Architecture

This paper proposes a multimodal video anomaly detection method based on vision-language alignment. Taking the video anomaly detection model MSF-DMU [25], which is based on multi-scale fusion and dual memory units, as a baseline, we introduce a vision-language alignment branch to address issues such as weak semantic understanding and low feature utilisation efficiency. The overall framework of the model is shown in Figure 1. It primarily comprises:a visual branch, a vision-language alignment branch, a Prompt Constructor &Learner(PCL) module, and a Dual-Branch Adaptive Gating Fusion(DBAGF) module.

Given an input video, we first use a pre-trained I3D network to extract clip-level visual features $X_v \in R^{T\times D}$ (where T is the number of clips and D is the feature dimension). Through memory cells and a Transformer encoder, we output a visual feature representation $F_c \in R^{T\times D_c}$ that contains global spatiotemporal contextual information. In the vision-language alignment branch, we utilize a pre-trained CLIP model to align the visual features F_c with a set of text prompts describing anomaly categories, generating a fine-grained frame-category alignment map $A \in R^{T\times C}$ (where C is the number of categories). Through a learnable gated fusion module, the global anomaly score output by the dual memory units is adaptively fused with the semantically guided score from the vision-language alignment branch to obtain the final frame-level anomaly prediction score.

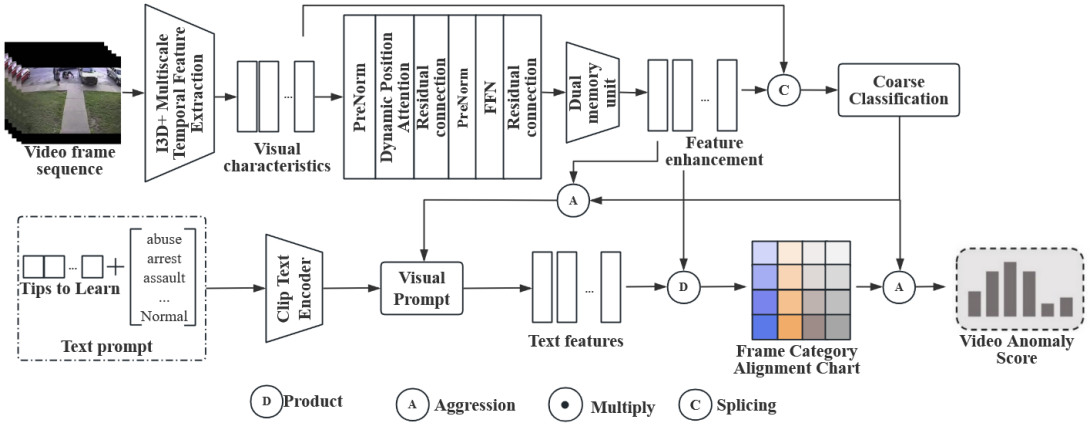


Figure 1. Model Organizational ChartText Prompt Construction and Learning Module (PCL)

To address the issue of missing semantic information, this paper designs a text prompt generation and learning module to support vision-language alignment, as illustrated in Figure 2.

For Ccategories in the dataset, we use multiple templates such as "a video of {} happening" and "a surveillance video of {}," filling in the category names into each template to generate textual descriptions for each category, resulting in a set of text prompts $P = [P_1, P_2, \ldots, P_C]$. We then encode these prompts using the CLIP text encoder to obtain text features $T_c \epsilon R^{D_t}$for each prompt, where D_t is the dimension of the CLIP text features. To further enhance the adaptability of the text representation, a learnable context vector Δc is

introduced for each category. The prompt text features T_c are added to the learnable offset Δc to obtain the modulated text features $\widetilde{T_c} = T_c + \Delta c.\widetilde{T_c}$ is then L2-normalized to obtain the final text features

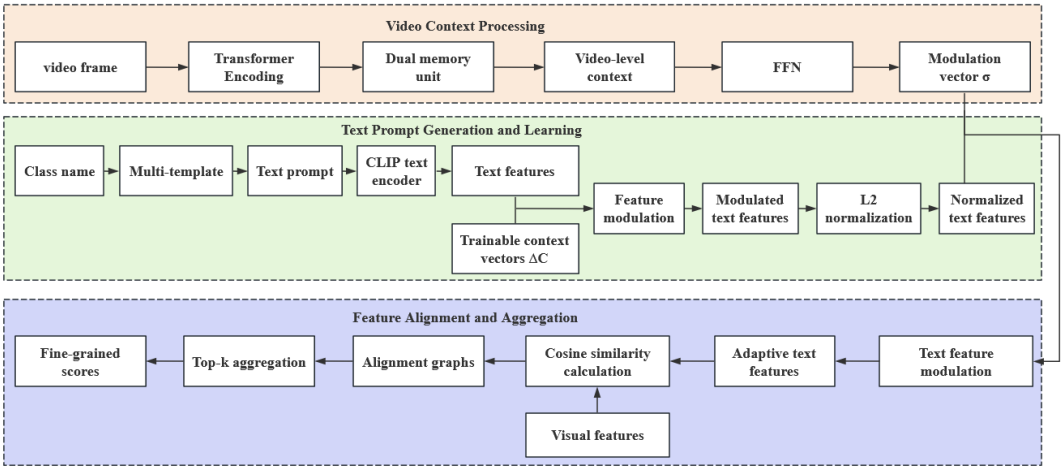


Figure 2. Diagram of the vision-language alignment architecture

The video frame feature $F_c \epsilon R^{T \times D_v}$(where T is the time step and D_v is the dimension of the visual feature) is mapped to the CLIP text feature space via a linear projection layer:

$$V' = W_v F_c + b_v \quad (1)$$

Here, $W_v \epsilon R^{D_t \backslash times D_v}$ and $b_v \epsilon R^{D_t}$ are learnable parameters. The projected visual features $V' \epsilon R^{T \times D_t}$are then L2-normalized.Using the video-level context vector $C_v \epsilon R^{D_c}$—derived from the memory module and Transformer encoding as input, a feedforward network consisting of fully connected layers and layer normalization is used to generate a modulation vector $\delta \epsilon R^{D_t}$:

$$\delta = \mathrm{LayerNorm}(W_c C_v + b_c) \quad (2)$$

Next, the text features are modulated using this modulation vector $\widetilde{T_c} = \widetilde{T_c} + \delta$. The modulated text features are then normalized again to obtain video-adaptive text features. The alignment map A is obtained by calculating the cosine similarity between the projected visual features V' and the text features $\widehat{T_c}$ for all categories:

$$A_{t,c} = \frac{V'.\widehat{T_c}}{||V'||||\widehat{T_c}||} \quad (3)$$

To obtain video-level fine-grained category scores from the alignment map, a $top-k$ aggregation strategy is applied for each category: the k frames with the highest similarity in the temporal dimension are selected, and the average similarity of these frames is calculated as the score for that category:

$$S_c = \frac{1}{k} \sum_{t \in topk(A:,c)} A_{t,c} \quad (4)$$

2.2. Dual-Branch Adaptive Gated Fusion Module (DBAGF)

The visual branch, based on sequence modeling and memory enhancement mechanisms, excels at capturing spatiotemporal anomaly patterns in videos but has limited capabilities in understanding fine-grained semantic information. The semantic branch introduces rich prior knowledge through vision-language alignment and can effectively distinguish anomaly categories, but may over-rely on the accuracy of text prompts. To effectively integrate the global pattern information from the visual branch with the fine-grained semantic information from the semantic branch, a dual-branch adaptive gated fusion mechanism is proposed. The specific process is illustrated in Figure 2.

Let $S_c \epsilon R^{B \times T}$ denote the clip-level anomaly score output by the visual branch, and let $A \epsilon R^{B \times T \times C}$ denote the frame-category alignment map output by the semantic branch. Derive the frame-level anomaly score $S_a \epsilon R^{B \times T}$ for the semantic branch from the alignment map A:

$$S_a = \sum_{c=1}^{C-1} A_{:,:,c} \quad (5)$$

Here,$C-1$represents the number of anomaly categories after excluding the normal category and the frame-level anomaly probability is obtained by summing the scores of the anomaly categories.The fusion weight is calculated using the learnable control parameter $g \in R$:

$$w = \sigma(g) = \frac{1}{1+e^{-g}} \quad (6)$$

Here, σ denotes the sigmoid function, which constrains the gating parameter to the interval (0,1). The final frame-level fusion score is:

$$S_{fused} = w.S_c + (1-w).S_a \quad (7)$$

Here,S_c represents the anomaly score output by the visual branch, and S_{fused} is the final fused score.The video-level fusion score is calculated by taking the $top-k$ average of the frame-level scores:

$$S_{video} = \frac{1}{k} \sum_{i=1}^{k} S_{fused}^{(i)} \quad (8)$$

As training progresses, g is automatically adjusted via gradient descent, enabling the model to find the optimal balance based on the characteristics of the dataset. This design allows the model to adaptively determine, in a data-driven manner, whether to rely more on visual patterns or semantic cues in its final decision, thereby enhancing the model's flexibility and robustness.

2.3. Multi-task Loss Function

To effectively utilize video-level weakly supervised labels and cross-modal alignment information, while simultaneously improving the model's ability to localize abnormal segments and classify them, this paper proposes a multi-task loss function.

The total loss function L_{total} consists of several components:

$$L_{total} = L_m + \lambda_1 L_{align} + \lambda_2 L_{cls} + \lambda_3 L_{con} \quad (9)$$

Specifically,L_m is a ranking-based main anomaly detection loss, L_{align} is an alignment loss based on multi-instance learning,L_{cls} is a fine-grained classification loss, L_{con} is a consistency loss, and λ1,λ2 and λ3 are hyperparameters used to balance the weights of these losses.

L_m serves as the base objective function for weakly supervised anomaly detection, driving the model to learn the classification boundary between normal and anomalous states at the video level. For each video, the mean of the scores from the top k segments with the highest anomaly scores($k = T//16+1$,where T is the total number of segments)is selected as the video-level predicted score.This score is then used to compute the binary cross-entropy loss against the video-level labels (0 for normal, 1 for abnormal), ensuring that the model learns to distinguish between normal and abnormal videos from the video-level labels.

To leverage the category-level semantic information provided by the vision-language alignment branch, a fine-grained classification loss L_{cls} is introduced.This loss acts directly on the category scores $S_{fine} \epsilon R^{B\times C}$ output by the vision branch:

$$L_{cls} = \frac{1}{B}\sum_{i=1}^{B} CrossEntropy(S_{fine}^{(i)}, y_{cls}^{(i)}) \quad (10)$$

Here, B represents the batch size, C represents the number of classes, and $y_{cls^{(i)}}$ represents the true class index of sample i .For normal samples(y_{cls}=0),an auxiliary loss is added to constrain the model such that the predicted scores for all abnormal classes, excluding the normal class, should be close to 0:

$$L_{cls} \leftarrow L_{cls} + \alpha MSE(S_{fine}[y_{cls} = 0,1:],0) \quad (11)$$

Here, α represents the weighting coefficient, and MSE stands for mean squared error. This approach further strengthens the decision boundary between normal and abnormal data.

To address the issue of missing frame-level labels in weakly supervised settings and fully leverage the fine-grained information provided by alignment map $A \in R^{B\times T\times C}$,we propose the MIL alignment loss L_{align} for an anomalous video, the predictions on the alignment map for the top k frames most relevant to its true class should also tend toward that true class. Let the alignment map be $A \in R^{B\times T\times C}$,where each element $A_{t,c}$ represents the similarity between frame t and class c .For each sample, select the top k frames most relevant to the ground-truth class $c^*(k = \lfloor T \cdot r \rfloor$, where r is the $top-k$ ratio hyperparameter), and compute the cross-entropy loss between the predictions for these frames and the ground-truth class:

$$L_{align} = \frac{1}{B}\sum_{i=1}^{B} CrossEntropy(A_{topk,c^*}^{(i)}, c^*) \quad (12)$$

Here,$A_{topk,c^{*(i)}}$ denotes the score vector for the $top-k$ frames in the i-th sample with respect to the ground-truth class c^*.

The consistency loss L_{con}ensures consistency between the predictions of the visual branch and the semantic branch, preventing conflicting predictions between the two branches. This loss calculates the mean squared error between the video-level anomaly scores of the two branches. Let $S_c \in R^B$ denote the video-level anomaly score of the semantic branch and $S_a \in R^B$ denote the video-level anomaly score of the visual branch:

$$L_{con} = \frac{1}{B}\sum_{i=1}^{B} (S_c^{(i)}, S_a^{(i)})^2 \quad (13)$$

This loss serves as a regularization term, encouraging the two branches to work in concert and converge together to a consistent decision boundary.

3. Experimental Design and Results

3.1. Experimental setup

(1) Datasets: To validate the effectiveness of the method, this paper employs two public datasets, UCF-Crime [26] and XD-Violence [27], for experimental validation. The UCF-Crime dataset comprises 1,900 uncropped videos sourced from real-world surveillance scenarios, with a total duration of approximately 128 hours. It covers 13 categories of abnormal events, including abuse, robbery, explosions and traffic accidents, and spans diverse environments such as streets, indoor spaces and shopping centres, thereby exhibiting significant real-world complexity. The XD-Violence dataset comprises 4,754 uncropped video clips with a total duration of approximately 217 hours. Video sources include surveillance footage, films, dashcam recordings and game footage, with content covering six categories of violent behaviour:abuse,traffic accidents, explosions,brawls,riots and shootings. Approximately 82% of the videos contain complex visual changes such as camera movement, sudden changes in viewpoint and scene transitions. The specific division of the dataset into train and test sets is shown in Table 1.

Table 1. Sample information for the dataset

Dataset	Train set	Test set	Total
UCF-Crime	1610	290	1900
XD-Violence	3954	800	4754

During the data pre-processing stage, to ensure the comparability and consistency of the research results, this paper follows the technical approach adopted in mainstream studies [27-28], employing an I3D model pre-trained on the Kinetics-400 dataset as the base network for feature extraction. The raw video is read frame-by-frame using OpenCV. First, the colour space is converted from BGR to RGB, then each frame is adjusted to pixel values consistent with the I3D model's input specifications, mapping the pixel values to the [-1,1] range. In terms of spatial dimensions, the video frames are centred and cropped to extract a 224×224 region, which serves as the input for I3D. I3D performs feature extraction in segments of 16 frames; therefore, the video is first divided into several 16-frame segments along the timeline, with a step size of 16. If the total number of frames is not a multiple of 16, the last frame is repeated to pad the sequence to the next multiple of 16, ensuring that each segment consists of 16 frames and can thus be fed into the I3D network in its entirety. During training, the model utilises pre-extracted and saved I3D features. All video features of varying lengths are uniformly divided into N=200 fixed-length segments to ensure standardisation of the feature

extraction process.

(2) Experimental environment: In this paper, Visual Studio Code is used as the development environment, Python as the programming language, and PyTorch 1.8.0 as the deep learning framework. Experiments were conducted on a server equipped with an NVIDIA GeForce RTX 3090 graphics card. The experimental study employs the Adam optimiser, with an initial learning rate set to 2e-4 and a batch size of 32, and a total of 3,000 training steps. For the CLIP text encoder, ViT-B/32 is selected as the backbone network [29-30], and learning rate scheduling utilises the ReduceLROnPlateau strategy.

(3) Evaluation Criteria: In video anomaly detection research, model performance is typically evaluated using the following core metrics: the area under the receiver operating characteristic (ROC) curve AUC, the area under the precision-recall curve (PR), and the false alarm rate (FAR), which reflects the model's ability to distinguish between normal and anomalous samples.

The ROC curve is formed by plotting the relationship between the true positive rate (TPR) and the false positive rate (FPR), with the x-axis representing the FPR and the y-axis representing the TPR. The TPR and FPR are defined as follows:

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{14}$$

$$FPR = \frac{FP}{N} = \frac{FP}{TN + FP} = FAR \tag{15}$$

Here, P represents the total number of samples that are actually positive, N represents the total number of samples that are actually negative, TP refers to the number of samples that are actually positive and correctly identified as such, FN represents the number of samples that are actually positive but incorrectly classified as negative, FP is the number of samples that are actually negative but incorrectly classified as positive, and TN represents the number of samples that are actually negative and correctly classified as such.

The AUC metric represents the area under the ROC curve, with values ranging from 0 to 1. A higher value indicates better overall classification performance of the model. Its formula is:

$$AUC = \int_0^1 ROC(FPR)d(FPR) \tag{16}$$

The higher the AP score, the better the model performs in identifying positive samples, i.e., anomalous events,and is particularly suitable for scenarios where the number of positive and negative samples is imbalanced. The formula for calculating it is as follows:

$$AP = \sum_n (R_n - R_{n-1}) \cdot P_n \tag{17}$$

The definitions of precision (P) and recall (R) are as follows:

$$R_n = \frac{TP}{TP + FN}, P_n = \frac{TP}{TP + FP} \tag{18}$$

These metrics collectively serve as a key tool for evaluating the overall performance of video anomaly detection models, providing insights into the model's stability,accuracy and generalisation ability when identifying anomalous events from various perspectives.

3.2. Comparative Experiments

In the evaluation on the UCF-Crime dataset, this paper employs a 10-crop augmentation strategy comprising the center, four corners, and their horizontal mirror images, and compares the resulting AUC scores with those of various existing deep learning-based video anomaly detection methods, as shown in Table 2. The data show that the method proposed in this paper achieves an AUC of 87.83%, outperforming most existing models. Although the false alarm rate (FAR) is 2.54 percentage points higher than that of the PEL model, the AUC—a core evaluation metric—improves by 1.77 percentage points compared to the PEL model's 86.06%, validating the effectiveness of this method.

Table 2. Comparison of AUC Values for Different Methods on the UCF-Crime Dataset

Method	AUC (%)	FAR (%)
VAD-CLIP [31]	87.35	N/A
CLIP-TSA [32]	86.93	N/A
MSF-DMU [25]	86.71	0.80
RTFM [21]	84.03	N/A
PEL [24]	86.06	0.7
DMU [33]	86.02	N/A
FusionNet+LSTM+GAN [34]	83.97	N/A
BN-SVP [35]	83.39	N/A
Ours	87.83	3.24

Figure 3 shows the frame-level score changes corresponding to the occurrence of four types of abnormal behaviors: arson, shoplifting, theft and vandalism. The curves represent the real-time score changes for each type of behavior, the shaded areas indicate the corresponding abnormal regions, and the video frames associated with these regions are listed below the axes. As shown in the figure, the score rises significantly when an anomaly occurs, indicating that the model is highly sensitive to anomalies and can accurately detect them.

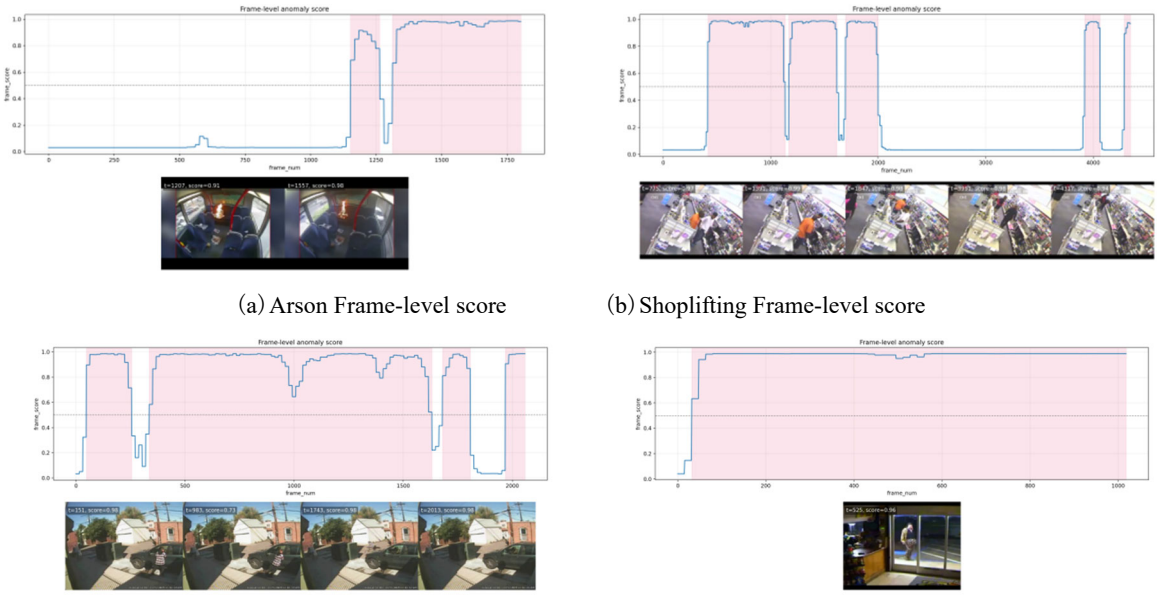
In tests on the XD-Violence dataset, this paper employs a 5-crop enhancement strategy [25] covering the central and four corner regions. Using Average Precision (AP) as the primary evaluation metric, we compare our method with various existing deep learning-based video anomaly detection methods. The results are listed in Table 3. Experiments demonstrate that the method proposed in this paper achieves an AP of 83.30%, outperforming most existing methods, while maintaining a false alarm rate(FAR) of 5.63%. Although the FAR value is slightly higher than that of the MSF-DMU model, the key evaluation metric, AP, is similarly superior.

Table 3. Comparison of AP Scores for Different Methods on the XD-Violence Dataset

Method	AP (%)	FAR (%)
VAD-CLIP [31]	83.02	N/A
CLIP-TSA [32]	82.17	N/A
MSF-DMU [25]	82.42	2.82
RTFM [21]	77.81	N/A
DMU [33]	79.62	N/A
MGFN [36]	79.19	N/A
Ours	83.30	5.63

Figure 4 shows the frame-level score change curves corresponding to the occurrence of four types of abnormal

behaviors in the XD-Violence dataset. The curves illustrate the real-time score changes for each behavior, while the shaded rectangular areas indicate the corresponding abnormal regions; the video frames corresponding to the current anomaly are displayed below the shaded areas. As shown in the figure, the score also rises significantly when an anomaly occurs, enabling accurate detection of the anomaly.



(a) Arson Frame-level score (b) Shoplifting Frame-level score

(c) Stealing Frame-level score (d) Vandalism Frame-level score

Figure 3. Frame-level scores for video anomaly detection on the UCF-Crime dataset

3.3. Ablation Studies

To verify the impact of each improved module on model performance, this paper conducts ablation studies on each module using the UCF-Crime and XD-Violence datasets. The results of the ablation studies are shown in Table 4.

Table 4. Model Ablation Experiment

				UCF-Crime (%)	XD-Violence (%)
baseline	PCL	DBAGF	Ltotal	AUC	AP
√				86.71	82.42
√	√			86.82	82.58
√	√	√		87.35	82.89
√	√	√	√	87.83	83.30

The results of the ablation experiments show that, after introducing text prompts into the purely visual model, the comprehensive evaluation metrics AUC and AP for both datasets improved compared to the baseline model, reaching 86.71% and 82.42%, respectively, thereby validating the positive impact of text prompts on anomaly detection. Building on this, we further introduced the dual-branch adaptive gated fusion module (DBAGF), which increased the AUC value for the UCF-Crime dataset from 86.71% to 86.82% and raised the AP value for the XD-Violence dataset to 82.58%, representing an increase of 0.16 percentage points. This indicates that the fine-grained semantic information from the language branch has been effectively fused with the global representations from the visual branch; Finally, by introducing the multi-task loss function Ltotal to adaptively adjust the weights of each loss, the AUC value on the UCF-Crime dataset was further improved to 87.83%, and the AP value on the XD-Violence dataset reached 83.30%, representing increases of 1.12 percentage points and 0.88 percentage points, respectively, compared to the baseline model. The above experimental results systematically validate the effectiveness of the video anomaly detection method based on vision-language alignment.

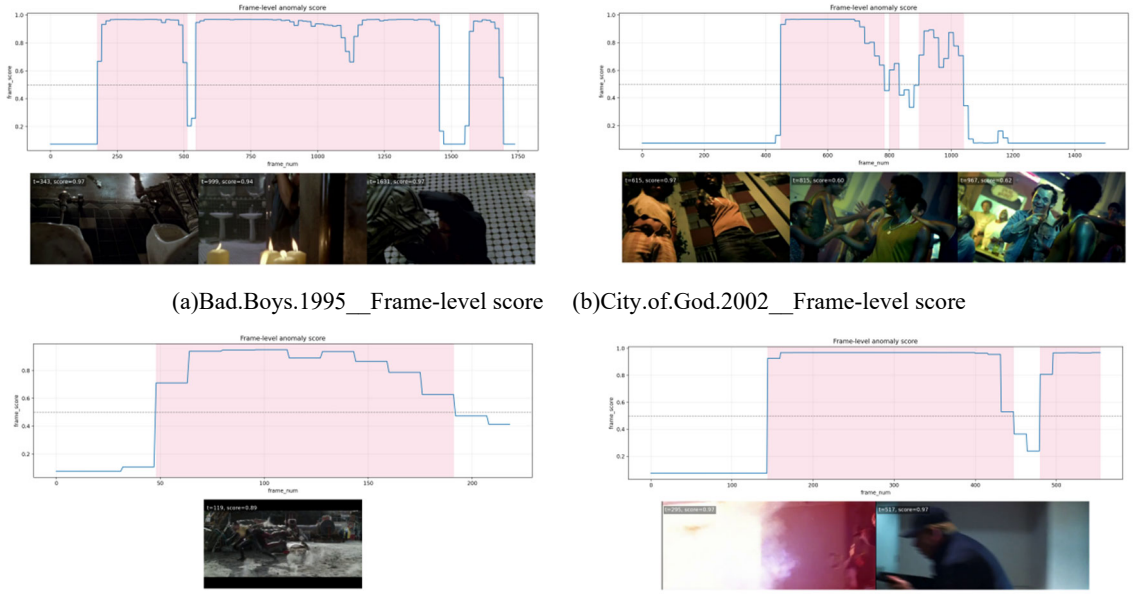
To validate the effectiveness of each sub-loss component in the model's multi-task loss function, we conducted ablation experiments on the UCF-Crime and XD-Violence datasets, designing four progressive experiments to quantify the independent contribution of each loss term. "baseline" refers to the results obtained using the base loss without the multi-task loss function Ltotal,"fine" refers to the fine-grained loss and "align" refers to the alignment loss and "cons" refers to the consistency loss. The experimental results are shown in Table 5.

Table 5. Loss Term Ablation Experiment

				UCF-Crime(%)	XD-Violence(%)
baseline	fine	align	cons	AUC	AP
√				87.35	82.89
√	√			87.51	83.01
√	√	√		87.79	83.27
√	√	√	√	87.83	83.30

Experimental results show that, for baseline anomaly detection, the AUC on the UCF-Crime dataset is 87.35%, and the AP on the XD-Violence dataset is 82.89%. After incorporating the fine-grained loss, the AUC on the UCF-Crime dataset and the AP on the XD-Violence dataset improved by 0.16 percentage points and 0.12 percentage points, respectively, demonstrating the necessity of fine-grained classification; After introducing alignment loss, the AUC on the UCF-Crime dataset increased from 87.51% to 87.79%, and the AP on the XD-Violence dataset increased from 83.01% to 83.27%, demonstrating the effectiveness of vision-language alignment in anomaly detection; After incorporating the consistency loss, the AUC on the UCF-Crime dataset reached a maximum of 87.83%, and the AP on the XD-Violence dataset reached 83.30%. This demonstrates that the introduction of a multi-task loss function enables the model to more fully leverage cross-modal alignment

information to enhance its ability to localize and classify anomalies.



(a)Bad.Boys.1995__Frame-level score (b)City.of.God.2002__Frame-level score

(c)Deadpool.2016__Frame-level score (d)Mission.Impossible.II.2000__Frame-level score

Figure 4. Frame-level scores for video anomaly detection on the XD-Violence dataset

4. Conclusion

In response to the limitations of purely visual models in terms of semantic understanding and fine-grained localisation in weakly supervised video anomaly detection, this paper proposes a video anomaly detection model based on vision-language alignment. By introducing a vision-language alignment branch, designing an adaptive gating fusion mechanism, and proposing a multi-task loss function based on an alignment graph, the model constructs an end-to-end framework capable of effectively integrating visual patterns with semantic information. Experiments on public datasets demonstrate that the model achieves an AUC of 87.83% on the UCF-Crime dataset and an AP of 83.30% on the XD-Violence dataset, outperforming many current mainstream methods. This indicates that the video anomaly detection model proposed in this paper is effective and capable of accurately identifying video anomalies.In the future,we plan to design efficient and lightweight neural network architectures and apply techniques such as knowledge distillation, quantization, and pruning to adapt the models to the computational constraints of edge devices.

Funding

This research was supported by the National Natural Science Foundation of China (61273294); National Social Science Fund of China (20BJL080); Shanxi Provincial Key Research and Development Programme (201803D121088); Shanxi Provincial Natural Science Research General Programme (202303021221173)

References

[1] Zhang H M, Yan D D, Tian Q Q. Improved spatio-temporal graph convolutional networks for video anomaly detection[J]. Opto-Electron Eng, 2024, 51(05): 48-60.DOI: 10.12086/oee.2024.240034

[2] Li N J, Nie X S, Li T, et al. A review of state-of-the-art video anomaly detection methods based on deep learning[J].Computer Applications Research, 2025, 42(03): 663-676.DOI: 10.19734/j.issn.1001-3695.2024.06.0241

[3] Zhang Y, Song J, Jiang Y, et al. Online video anomaly detection[J]. Sensors, 2023, 23(17): 7442.DOI: 10.3390/s23177442

[4] Ramachandra, Bharathkumar, Michael J. Jones, et al. A survey of single-scene video anomaly detection[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(5): 2293-2312.DOI: 10.1109/TPAMI.2020.3040591

[5] Li N, Wu X, Xu D, et al. Spatio-temporal context analysis within video volumes for anomalous-event detection and localization[J]. Neurocomputing, 2015, 155: 309-319.

[6] Breitenstein M D, Reichlin F, Leibe B, et al. Robust tracking-by-detection using a detector confidence particle filter[C]//Proceedings of the 2009 IEEE 12th International Conference on Computer Vision (ICCV). Los Alamitos: IEEE Computer Society, 2009: 1515-1522.DOI: 10.1109/ICCV.2009.5459278

[7] Piciarelli C, Micheloni C, Foresti G L. Trajectory-based anomalous event detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2008, 18(11): 1544-1554.

[8] Wang X, Tieu K, Grimson E. Learning semantic scene models by trajectory analysis[C]//European Conference on Computer Vision. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 110-123.DOI: 10.1007/11744078_9

[9] Andrade E L, Blunsden S, Fisher R B. Hidden markov models for optical flow analysis in crowds[C]//Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006). Los Alamitos: IEEE Computer Society, 2006: 460-463.DOI: 10.1109/ICPR.2006.621

[10] Hu W, Tan T, Wang L, et al. A survey on visual surveillance of object motion and behaviors[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2004, 34(3): 334-352.DOI: 10.1109/TSMCC.2004.829274

[11] Mehran R, Oyama A, Shah M. Abnormal crowd behavior detection using social force model[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE Computer Society, 2009: 935-942.

[12] Sharif M H, Djeraba C. An entropy approach for abnormal activities detection in video streams[J]. Pattern Recognition, 2012, 45(7): 2543-2561.DOI: 10.1016/j.patcog.2012.01.009

[13] Feizi A, Aghagolzadeh A, Seyedarabi H. Using optical flow and spectral clustering for behavior recognition and detection of anomalous behaviors[C]//Proceedings of the 2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP). Los Alamitos: IEEE Computer Society, 2013: 210-213.

[14] Zhou S, Shen W, Zeng D, et al. Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes[J]. Signal Processing: Image Communication, 2016, 47: 358-368.DOI: 10.1016/j.image.2016.07.007

[15] Sabokrou M, Fathy M, Hoseini M. Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder[J]. Electronics Letters, 2016, 52(13): 1122-1124.DOI: 10.1049/el.2016.1026

[16] Nguyen T N, Meunier J. Anomaly detection in video sequence with appearance-motion correspondence[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Los Alamitos: IEEE Computer Society, 2019: 1273-1283.DOI: 10.1109/ICCV.2019.00136

[17] Zaheer M Z, Mahmood A, Khan M H, et al. Generative cooperative learning for unsupervised video anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE Computer Society, 2022: 14744-14754.DOI: 10.1109/CVPR52688.2022.01435

[18] Huang C, Wen J, Xu Y, et al. Self-supervised attentive generative adversarial networks for video anomaly detection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 34(11): 9389-9403.DOI: 10.1109/TNNLS.2022.3155154

[19] Huang C, Wu Z, Wen J, et al. Abnormal event detection using deep contrastive learning for intelligent video surveillance system[J]. IEEE Transactions on Industrial Informatics, 2021, 18(8): 5171-5179.DOI: 10.1109/TII.2021.3121891

[20] Lv H, Yue Z, Sun Q, et al. Unbiased multiple instance learning for weakly supervised video anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE Computer Society, 2023: 8022-8031.DOI: 10.1109/CVPR52729.2023.00774

[21] Tian Y, Pang G, Chen Y, et al. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Los Alamitos: IEEE Computer Society, 2021: 4975-4986.DOI: 10.1109/ICCV48922.2021.00493

[22] Cho M A, Kim M, Hwang S, et al. Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE Computer Society, 2023: 12137-12146.DOI: 10.1109/CVPR52729.2023.01167

[23] Chen J, Li L, Su L, et al. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE Computer Society, 2024: 18319-18329.

[24] Pu Y, Wu X, Yang L, et al. Learning prompt-enhanced context features for weakly-supervised video anomaly detection[J]. IEEE Transactions on Image Processing, 2024, 33: 4923-4936.

[25] Zhang Y, Zhao Y A, Kong L P, et al. Video anomaly detection based on multi-scale fusion and dual memory units[J]. Computer Technology and Development,2026,36(04):69-77.DOI:10.20165/j.cnki.ISSN1673-629X.2025.0286.

[26] Zhang C, Li G, Xu Q, et al. Weakly supervised anomaly detection in videos considering the openness of events[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(11): 21687-21699.

[27] Wu P, Liu J, Shi Y, et al. Not only look, but also listen: Learning multi modal violence detection under weak supervision[C]//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer International Publishing, 2020: 322-339.DOI: 10.1007/978-3-030-58536-5_20

[28] Wu P, Liu J. Learning causal temporal relation and feature discrimination for anomaly detection[J]. IEEE Transactions on Image Processing, 2021, 30: 3513-3527.DOI: 10.1109/TIP.2021.3062204

[29] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE Computer Society, 2022: 11976-11986.DOI: 10.1109/CVPR52688.2022.01167

[30] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[C]//Proceedings of the International Conference on Machine Learning (ICML). New York: PMLR, 2021: 10347-10357.

[31] Wu P, Zhou X, Pang G, et al. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Palo Alto: AAAI Press, 2024: 6074-6082.

[32] Joo H K, Vo K, Yamazaki K, et al. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection[C]//Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP). Los Alamitos: IEEE Computer Society, 2023: 3230-3234.

[33] Zhou H, Yu J, Yang W. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Palo Alto: AAAI Press, 2023: 3769-3777.DOI: 10.1609/aaai.v37i3.25492

[34] Zhao Y G, Yang Y J, Xiang T, et al. Video anomaly detection framework based on bidirectional spatio-temporal feature fusion GAN[J]. Journal of Jilin University (Information Science Edition), 2025, 43(05): 1128-1137.DOI:10.19292/j.cnki.jdxxp.20250623.001

[35] Sapkota H, Yu Q. Bayesian nonparametric submodular video partition for robust anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE Computer Society, 2022: 3212-3221.DOI: 10.1109/CVPR52688.2022.00320

[36] Chen Y, Liu Z, Zhang B, et al. Mgfn:Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Palo Alto: AAAI Press, 2023: 387-395.DOI: 10.1609/aaai.v37i1.25100