Unsupervised Behavior Anomaly Detection Model Based on Contrastive Learning

Shujing Tong and Yongfei Wu *

Posted Date: 21 April 2026

doi: 10.20944/preprints202604.1458.v1

Keywords: contrastive learning; unsupervised anomaly detection; behavior recognition; spatio-temporal feature learning

 doi:10.20944/preprints202604.1458.v1

Article

Unsupervised Behavior Anomaly Detection Model Based on Contrastive Learning

Shujing Tong and Yongfei Wu *

Taiyuan University of Technology, No. 209 Daxue Street, Yuci District, Jinzhong 030600, Shanxi Province, China

* Correspondence: wuyongfei@tyut.edu.cn

Abstract

In the context of the rapid popularization of intelligent monitoring and edge perception, automatic identification of abnormal behaviors in complex scenarios has become a key issue in video understanding. This paper proposes an unsupervised behavior anomaly detection model based on contrastive learning. Through hierarchical organization of normal samples, joint spatio-temporal encoding, time attention aggregation, and "instance contrast - prototype traction - time smoothing" joint optimization, stable behavior embedding representations are learned. In the detection stage, a comprehensive anomaly score is constructed by integrating the recent prototype deviation, second-order temporal residual, and local neighborhood support information, and an adaptive threshold based on the median and absolute median difference is adopted for judgment. Experimental results show that the model achieves an AUC of 97.4% on UCSD Ped2, 91.8% on CUHK Avenue, and 83.7% on ShanghaiTech. The average AUC and average F1 are 91.0% and 88.1% respectively. The study demonstrates that this method can enhance the stability and generalization ability of anomaly detection in complex video scenarios, providing a reference technical path for video intelligent early warning in the absence of labels.

Keywords: contrastive learning; unsupervised anomaly detection; behavior recognition; spatio-temporal feature learning

1. Introduction

With the continuous deployment of video surveillance systems, intelligent sensing terminals and edge computing devices in scenarios such as public security, intelligent transportation, industrial inspection and park management, the demand for behavior analysis in complex environments is constantly increasing [1]. Although massive video data provide rich information sources for intelligent monitoring, it also brings practical problems such as low processing efficiency, high cost of manual screening and delayed response to abnormal events. Abnormal behaviors usually have suddenness, low frequency and uncertainty, and their manifestations are also affected by factors such as shooting angle, target density, background interference, light changes and scene semantic differences. Therefore, relying solely on manual review or simple rule matching is no longer able to meet the dual requirements of real-time performance and accuracy for practical applications [2,3]. In this context, behavior anomaly detection based on computer vision has gradually become an important research direction in the field of video understanding and intelligent analysis, and also a key issue that must be addressed in the evolution of intelligent monitoring systems towards automatic early warning and active response.

Unlike tasks such as object detection and action recognition that have clear category boundaries, anomaly behavior detection faces greater openness and higher modeling difficulty. Anomalies in real scenarios are difficult to exhaustively list, and the external features of the same category of anomalies in different scenarios often show significant differences. For example, anomalies may manifest as trajectory deviation, sudden changes in movement speed, local stagnation anomalies, or imbalance

 doi:10.20944/preprints202604.1458.v1

in group behavior. The determination criteria are not based on a single visual mode, but are related to the time evolution process, spatial positional relationship, and scene context [4–6]. This means that the model not only needs to extract local visual features, but also should effectively capture the dynamic changes of behavior in the temporal dimension and the structural correlation between the target and the environment. If the feature representation ability is insufficient, the detection results are prone to be interfered by background noise and the diversity of normal behaviors, thereby leading to an increase in false alarm rate or blurring of the anomaly boundary. For surveillance videos in open scenarios, such misjudgments will not only increase the burden on the system's alarm, but also may weaken the effectiveness of subsequent risk identification and linkage handling.

In current research, supervised learning methods have achieved good results on specific datasets, but these methods rely on large-scale labeled samples, especially requiring sufficient coverage of abnormal categories [7–9]. Due to the scarcity and imbalance of abnormal behaviors, it is not realistic to obtain complete, accurate and representative abnormal labels in actual monitoring systems, which largely limits the application of supervised methods [10]. In contrast, unsupervised anomaly detection better meets the needs of engineering practice. Its basic idea is to learn a stable distribution from normal behavior samples, and then use the deviation of test samples from the normal distribution to achieve anomaly identification. This type of method reduces the reliance on manual labeling, but also has problems such as insufficient representativeness, insufficient modeling of complex spatiotemporal relationships, and limited robustness in diverse scenarios. Especially in video data, the normal behavior itself has strong variability, and if the model can only learn shallow statistical patterns, it often fails to form a discriminative boundary with distinguishability. We believe that the key factor that truly affects the detection effect is not only whether the model has strong fitting ability, but also whether it can learn stable and separable normal behavior structures from complex video sequences.

In recent years, contrastive learning has demonstrated significant advantages in unsupervised representation learning, providing new research ideas for abnormal behavior detection [11]. This method constructs consistency constraints between samples, enabling the model to learn more discriminative feature representations without the need for manual labels. For behavior videos, if a reasonable positive-negative contrast relationship can be established around normal samples and combined with spatiotemporal information for feature encoding, it is possible to enhance the aggregation of normal patterns in the embedding space, thereby improving the stability and sensitivity of abnormal determination [12,13]. However, video abnormality detection cannot simply adopt the general image contrastive learning framework. Behavioral data has strong continuity, obvious context dependence, and complex semantic hierarchies. If the sample construction method is improper, the model is prone to bias towards background textures or appearance information, making it difficult to accurately depict the behavior changes themselves. Therefore, how to effectively combine contrastive learning with the spatiotemporal modeling process of video behaviors, and how to balance the internal differences of normal samples and the overall distribution stability under unsupervised conditions, remain issues worthy of in-depth research.

Based on this, this paper focuses on the unsupervised behavior anomaly detection task based on contrastive learning, and systematically designs the organization of normal behavior samples, spatio-temporal feature encoding, contrastive optimization mechanism, and abnormal score calculation method, aiming to improve the model's ability to identify abnormal behaviors in complex scenarios without relying on abnormal labels. During the research process, we attempted to start from modeling the normal behavior distribution, integrating sample organization, representation learning, and detection determination into a unified framework to enhance the model's perception of behavioral spatio-temporal changes and its ability to distinguish abnormal boundaries. Based on the review of related studies, this paper constructs an unsupervised detection model for video behavior analysis and validates the effectiveness of the method through experiments, with the aim of providing a valuable technical path for abnormal behavior recognition in intelligent video monitoring.

 doi:10.20944/preprints202604.1458.v1

2. Relevant Research Analysis

2.1. Main Limitations of Traditional Abnormal Behavior Detection Methods

Traditional abnormal behavior detection research is mostly based on artificially designed features and rule-based determination. Common practices include inter-frame difference analysis, optical flow analysis, target trajectory statistics, local motion descriptors extraction, and threshold-based anomaly discrimination. These methods have the advantages of simple implementation and low computational cost when the monitoring scenarios are relatively simple, with limited background changes and a small number of targets. Therefore, they were widely applied in early video analysis systems. However, from the development of computer vision tasks, traditional methods are more suitable for processing scenarios with clear structures and weak interference. Once they enter real open environments, their limitations will quickly become apparent. Considering the application requirements of abnormal behavior detection in complex video scenarios, we believe that the adaptability of these methods in open environments has difficulty meeting the current requirements for accuracy and stability in intelligent monitoring tasks.

Traditional methods rely on manual experience to select explicit features such as motion amplitude, direction changes, trajectory length, and regional stay time. These features can describe local abnormal phenomena, but they are difficult to cover the temporal and spatial correlations in complex behaviors. Abnormal behaviors are not single-point mutations in a frame but are determined by continuous actions, environmental relationships, and contextual semantics. Only shallow features are difficult to form stable representations. These methods are sensitive to scene conditions, such as camera jitter, illumination fluctuations, occlusion enhancement, and dynamic background changes, which can interfere with optical flow and trajectory extraction results, misclassifying normal behaviors or masking real abnormalities. Additionally, traditional methods usually rely on preset thresholds or rule templates for determination, and the crowd density, movement rhythm, and spatial layout in different scenes are not consistent. Fixed thresholds lack adaptability and tend to perform poorly when transferred to new scenes. When reviewing related research, we found that these methods still have certain application value in controlled scenarios, but once the complexity of the scene increases, their feature expression ability and determination flexibility will be significantly limited.

From an engineering application perspective, traditional methods also have a prominent problem, namely, insufficient ability to recognize complex abnormalities. For behaviors with obvious movement differences such as running, gathering， and wandering， rule models can still provide basic judgments； however， when facing more concealed， longer-lasting， and less obvious local changes in abnormal events， it is difficult for artificially constructed features to accurately depict their evolution process. That is to say， traditional methods are better at identifying “clearly changing” abnormalities， but they are insufficient in responding to “ambiguous boundaries” or “strongly semantic-dependent” abnormalities. Therefore， although these methods laid the foundation for abnormal behavior detection research， they have been unable to meet the requirements of current video monitoring systems for robustness， generalization， and fine-grained discrimination capabilities. Table 1 summarizes the applicable characteristics and main limitations of traditional abnormal behavior detection methods.

Table 1. Main Characteristics and Limitations of Traditional Abnormal Behavior Detection Methods.

Method Type	Typical Thought	Extractable Information	Primary Advantage	Primary Limitation
Inter-frame Difference Method	Compare the pixel change areas of adjacent frames	Motion position, rough change range	Computationally simple, suitable forchanges, rapid detection	Prone to be affected by noise and illumination difficult to describe complex behaviors

 doi:10.20944/preprints202604.1458.v1

Optical Flow Field Analysis Method	Calculate the direction and speed of pixel or area movement	Local motion intensity, direction distribution	Reflect short-term motion characteristics	Sensitive to occlusion, camera jitter, and background disturbances
Trajectory Statistics Method	Trajectory tracking and analysis	Motion path, stay time, speed change	Has certain explanatory power for continuous behaviors	Tracking is unstable in multi-target scenarios, and the trajectory loss leads to false determination
Methods Based on Artificial Features	Extract direction gradients, local motion descriptors, etc.	Local texture and action cues	Easy to implement, convenient for combination with classifiers	Feature expression is shallow, difficult to model long-term dependence and scene semantics
Rule or Threshold Discrimination Method	Establish abnormal conditions based on experience	Region boundary crossing, speed anomaly, stay anomaly	Deployment is convenient, suitable for simple monitoring tasks	Rules are rigid, have weak cross-scenario adaptability, and have more false alarms and missed detections

2.2. Research Progress and Challenges of Unsupervised Deep Abnormality Detection Methods

With the widespread application of deep learning in computer vision, unsupervised anomaly detection has gradually become an important research direction in the field of behavior analysis. Such methods usually only use normal samples for training and learn the spatiotemporal distribution patterns of normal behaviors to complete anomaly determination in the test stage based on the degree of deviation of the samples from the normal mode. Compared with supervised methods that rely on manual annotation, unsupervised methods are more suitable for real monitoring scenarios and better fit the characteristics of scarce abnormal samples, high annotation costs, and difficult enumeration of abnormal categories. Based on this background, we will also take unsupervised modeling as the core research path in the subsequent method design.

From existing research, unsupervised deep anomaly detection mainly forms three technical paths. One is the reconstruction method, which uses autoencoders or generative models to restore the input content and then identifies anomalies based on the reconstruction error; the second is the prediction method, which models the temporal evolution process of video frames or behavioral features and judges based on the deviation between the predicted results and the real observations; the third method places more emphasis on deep representation learning, hoping to use feature embedding to depict the distribution of normal behaviors and discover abnormal samples through distance, density, or clustering relationships. Compared with traditional artificial feature methods, these methods have significantly improved feature extraction and nonlinear expression capabilities in complex scenarios, and have to some extent expanded the modeling space of video anomaly detection.

However, these methods still have some shortcomings. The reconstruction model sometimes restores abnormal areas along with the normal ones, weakening the difference between anomalies and normal cases; the prediction model is highly dependent on temporal continuity and has poor error stability when affected by occlusion, illumination fluctuations, and background changes; the representation learning methods, although more focused on the distribution of feature space, are prone to bias towards appearance textures and ignore the actual behavior changes if they lack effective constraints. In addition, video anomaly detection also faces problems such as large internal differences in normal behaviors, difficulty in fully characterizing long-term dependencies, and insufficient cross-scenario generalization ability. Thus, although unsupervised deep anomaly detection has made significant progress, in complex environments, a more stable feature learning mechanism is still needed to improve anomaly discrimination ability, which is the direct motivation for us to introduce contrastive learning for modeling.

 doi:10.20944/preprints202604.1458.v1

2.3. Adaptability Analysis of the Introduction of the Abnormal Detection Task Through Contrastive Learning

Contrastive learning was originally used for unsupervised representation learning. Its core idea is to bring samples closer in the feature space and pull apart dissimilar samples, thereby enhancing the ability to distinguish representations. For anomaly detection tasks, this mechanism has strong adaptability. Anomaly behavior samples are limited in number and have unstable forms. If only relying on explicit labels for training, the model often fails to cover the abnormal patterns in complex scenarios; while contrastive learning can build a stable distribution structure around normal samples without relying on a large amount of annotations, making the model focus more on the intrinsic differences between behavioral fragments.

Let sample xi be mapped by the encoder to feature vector zi, and its corresponding positive sample is zi+, while the other samples are negative samples. Then, the commonly used contrastive loss can be expressed as:

$$L_{cl}=-\log\frac{\exp\left(\mathrm{sim}(z_i,z_i^+)/\tau\right)}{\exp\left(\mathrm{sim}(z_i,z_i^+)/\tau\right)+\sum_{j=1}^{N}\exp\left(\mathrm{sim}(z_i,z_j^-)/\tau\right)} \quad (1)$$

Among them, sim(·) represents the feature similarity function, τ is the temperature coefficient, and N is the number of negative samples. This loss enhances the consistency of positive sample pairs and suppresses the similarity of negative sample pairs, thereby forming a more compact aggregation structure for normal behaviors in the embedding space. After introducing contrastive learning into anomaly detection, the model is no longer limited to reconstruction error or prediction bias, but can depict normal patterns from the feature distribution perspective. Thus, abnormal samples, due to their deviation from the normal aggregation area, are more easily identified in the feature space. For video behavior analysis, this approach is particularly suitable for handling scenarios with complex normal patterns and blurred abnormal boundaries, and provides a more stable representation basis for subsequent anomaly score calculation.

3. Model Design

3.1. Unsupervised Behavior Abnormality Detection Framework Based on Contrastive Learning

To address the issues of scarce abnormal behavior samples, high cost of manual annotation, and unstable abnormal pattern boundaries in complex scenarios, this paper constructs an unsupervised behavior anomaly detection framework based on contrastive learning. This framework only utilizes normal behavior videos for training, integrating video modeling, temporal and spatial feature extraction, representation mapping, and anomaly scoring into a single process, enabling the model to learn the stable distribution of normal behaviors without abnormal labels. The architecture of the proposed unsupervised behavior anomaly detection model based on contrastive learning is illustrated in Figure 1.

In the computer implementation process, the original surveillance video is first segmented by a sliding time window, resulting in a set of video segments of fixed length:

$$X=\{x_1,x_2,\ldots,x_n\} \quad (2)$$

Among them, x_n represents the n-th behavior segment, which includes consecutive frame images and their corresponding local motion information. To reduce the interference of single-frame noise on representation learning, this paper adopts a unified sampling strategy to standardize the segments, and then feeds them into the encoding network to complete feature extraction. Let the feature encoder be f(·), then the basic representation of the input segment can be written as:

$$h_i=f(x_i),\ h_i\in R^d \quad (3)$$

Considering that abnormal behaviors are often influenced by both spatial layout changes and temporal dynamic evolution, this paper introduces a spatio-temporal joint representation mechanism into the framework, integrating the appearance structure information extracted from the spatial

 doi:10.20944/preprints202604.1458.v1

branches with the action evolution information extracted from the temporal branches to obtain the segment-level behavior representation:

$$z_i=\phi(h_i^s,h_i^t) \tag{4}$$

Here, h_i^s represents the spatial feature component, h_i^t represents the temporal feature component, and $\phi(\cdot)$ represents the feature fusion mapping function. This design can avoid the model relying solely on local textures or a single motion signal for judgment, thereby enhancing the ability of the representation to depict complex behavioral changes.

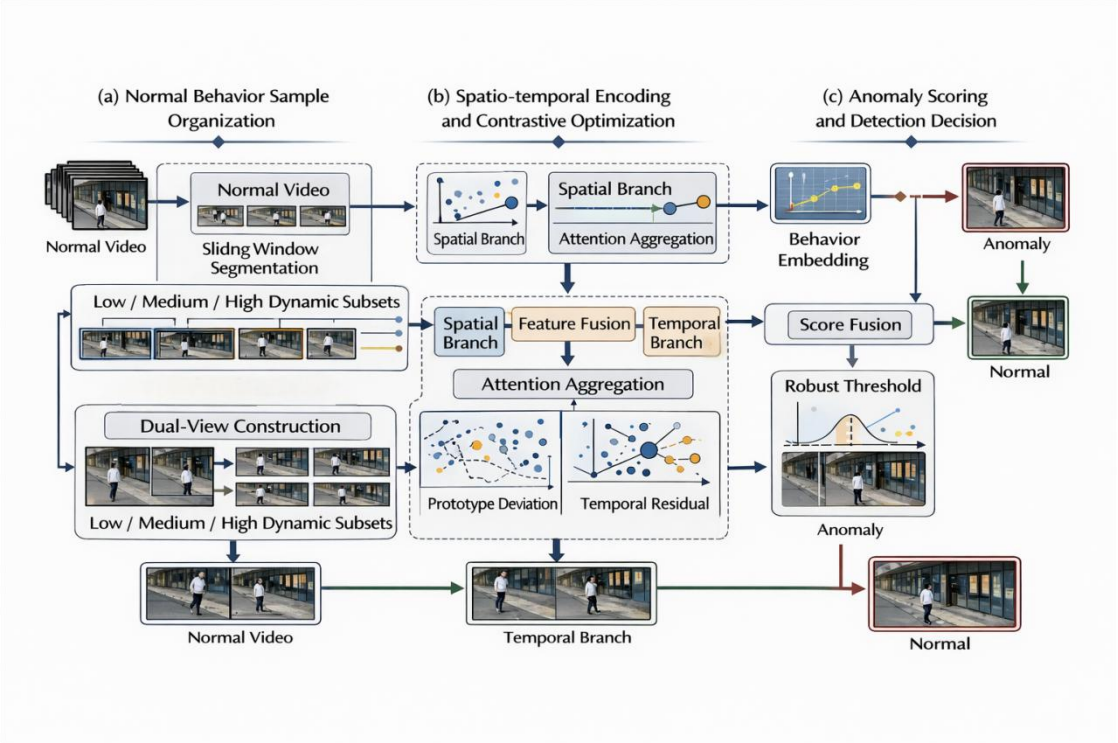

Figure 1. Architecture of the proposed contrastive-learning-based unsupervised behavior anomaly detection model.

Based on this, the model further maps the fused representation to the contrastive learning embedding space, gradually forming a stable depiction of normal behavior patterns. For any test segment xi, this paper no longer simply attributes the abnormality degree to a certain central distance or a single threshold comparison result, but instead uses the deviation of its representation from the normal distribution as the basis for detection. Specifically, in the test stage, based on the aforementioned spatio-temporal encoding and representation learning results, the distribution deviation degree between the segment and the normal behavior prototype, the consistency of changes in adjacent segments in the temporal dimension, and the support strength of normal samples in the local neighborhood will be comprehensively examined, thereby constructing a more stable abnormal score. Abnormal determination no longer relies on a single piece of evidence, but can jointly identify abnormal segments from three levels: distribution position, dynamic change, and local structure, providing a unified scoring basis for the subsequent detection and determination method design. Let the center of the normal behavior prototype be c, then the abnormal score of the segment is defined as:

$$S_i=\|z_i-c\|_2^2 \tag{5}$$

 doi:10.20944/preprints202604.1458.v1

Among them, S_i represents the anomaly score of the i-th video segment, and $||\cdot||2$ represents the Euclidean distance. When Si exceeds the set threshold δ, the segment will be determined as abnormal behavior:

$$y_i=\begin{cases}1, & S_i\geq\delta\\ 0, & S_i<\delta\end{cases} \tag{6}$$

The above framework organizes the model process from four levels: "video segment input - spatiotemporal encoding - representation mapping - anomaly scoring". It effectively connects the unsupervised representation learning with the anomaly determination process. On one hand, normal samples can gradually form a more concentrated feature distribution during the training stage; on the other hand, the distance difference between the test sample and the normal center also provides a more stable quantitative basis for subsequent anomaly identification. Overall, this framework takes into account the spatiotemporal characteristics of video behavior analysis and the feasibility of unsupervised learning conditions, laying the foundation for the organization of normal samples, comparison optimization, and detection determination method design.

3.2. Organization of Normal Behavior Samples and Feature Representation Learning

Although unsupervised anomaly detection only uses normal samples for training, normal behavior itself is not a single pattern. Taking surveillance videos as an example, in the same scene, there are not only low-speed passage and short stops, but also turning, avoidance, merging and other daily actions of varying intensities. If these segments are directly input into the model without distinction, it is likely to cause the internal structure of the normal samples to be too loose, which is not conducive to the formation of subsequent stable representation. Based on this, in the sample organization stage, this paper mainly organizes normal segments from two aspects: temporal continuity relationship and motion intensity difference, and maintains the consistency of representation of similar segments in the subsequent dual-view construction and prototype aggregation process, so that the feature learning results are closer to the distribution structure of the real behavior.

Let the normal training video set be $\mathrm{V^N}=\{\mathrm{v_1},\mathrm{v_2},\ldots,\mathrm{v_M}\}$, and after sliding window segmentation, the normal behavior segment set can be obtained:

$$\mathrm{X^N}=\{\mathrm{x_1},\mathrm{x_2},\ldots,\mathrm{x_n}\} \tag{7}$$

To avoid excessive redundancy in the samples, this paper calculates the motion response intensity for each segment and conducts a preliminary screening based on this. The average motion intensity of the i-th segment over T frames is defined as:

$$\varrho_i=\frac{1}{\mathrm{T}}\sum_{\mathrm{t=1}}^{\mathrm{T}}\|\mathrm{F_t}\|_1 \tag{8}$$

Here, F_t represents the optical flow response corresponding to the t-th frame. Based on the distribution of ρ_i, normal segments can be divided into three subsets: low dynamic, medium dynamic, and high dynamic. The purpose of this processing is not to artificially create category labels, but to reduce the disorderly overlap between different intensity normal behaviors, so that the model can see the internal differences of normal samples during training and not weaken the distribution structure due to overly loose sample organization.

In the feature representation learning stage, this paper constructs two views for each normal segment: one retains the original temporal structure, and the other is formed through lightweight perturbation to form an enhanced view, which is used to enhance the model's perception of the stable attributes of normal behaviors. Let the two views be xia and xib, then the segment representation can be written as:

$$z_i^a=g(f(x_i^a)),z_i^b=g(f(x_i^b)) \tag{9}$$

 doi:10.20944/preprints202604.1458.v1

Among them, f(·) is the spatio-temporal encoder, and g(·) is the projection mapping function. To avoid the representation results being completely dependent on a single sample, this paper further introduces intra-group prototype representation to aggregate the features within the same dynamic level:

$$p_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} z_i \tag{10}$$

Among them, C_k represents the kth normal subset, and p_k is its prototype center. During training, the sample representation not only maintains consistency between the two views but also requires a high degree of proximity to the prototype of the corresponding subset. Therefore, it is possible to enhance the local aggregation ability of normal behaviors without introducing abnormal labels, and reduce the interference of background noise and accidental disturbances on representation learning.

The organization strategy and role of normal behavior samples are shown in Table 2.

Table 2. Organization Strategy and Role of Normal Behavior Samples.

Organization Step	Processing Method	Output Result	Main Role
Normal Segment Extraction	Using sliding window to divide continuous video	Fixed-length behavior segments	Reserve basic temporal information
Motion Intensity Filtering	Calculating dynamic degree based on optical flow response	Low, medium, and high dynamic subsets	Reduce sample aliasing
Dual View Construction	Parallel input of original view and lightweight enhanced view	Paired training samples	Improve representation stability
Prototype Aggregation	Computing the average of group-specific features to obtain the center representation	Normal behavior prototype	Enhance distribution compactness
Representation Learning	Joint training of spatio-temporal encoding and mapping	Fragment embedding features	Provide a discriminative basis for anomaly detection

Overall, this design transforms the originally loose normal training data into an input set with a hierarchical structure, enabling the feature representation to gradually form a stable embedding oriented towards behavioral patterns, providing a more reliable representation basis for subsequent spatio-temporal contrast optimization and anomaly score calculation.

3.3. Spatio-Temporal Feature Encoding and Contrastive Optimization Mechanism

Video anomalies are not simply the cumulative of static appearance differences, but are formed by local action changes, short-term speed fluctuations, and long-term behavioral evolution. For this characteristic, this paper introduces a joint mechanism of "local motion enhancement - global temporal aggregation - hierarchical contrast optimization" in the feature learning stage, enabling the behavior representation to not only reflect the dynamic changes within the segment but also maintain the structural consistency across segments. The spatio-temporal feature encoding and contrastive optimization process is illustrated in Figure 2.

Let the input behavior segment be $x_i=\{I_1,I_2,\ldots,I_T\}$, where It represents the image of the t-th frame. The spatial branch extracts the appearance feature st for each frame, while the temporal branch constructs the motion response mt based on the differences between adjacent frames. The single-step representation after fusion is defined as:

$$u_t=\sigma(W_s s_t+W_m m_t+b) \tag{11}$$

 doi:10.20944/preprints202604.1458.v1

Here, W_s and W_m are learnable parameters, and σ(·) is a nonlinear mapping function. Different from direct concatenation, this weighted fusion method can automatically adjust the proportion of static and dynamic information according to the training process, reducing the interference of background texture on the behavior representation.

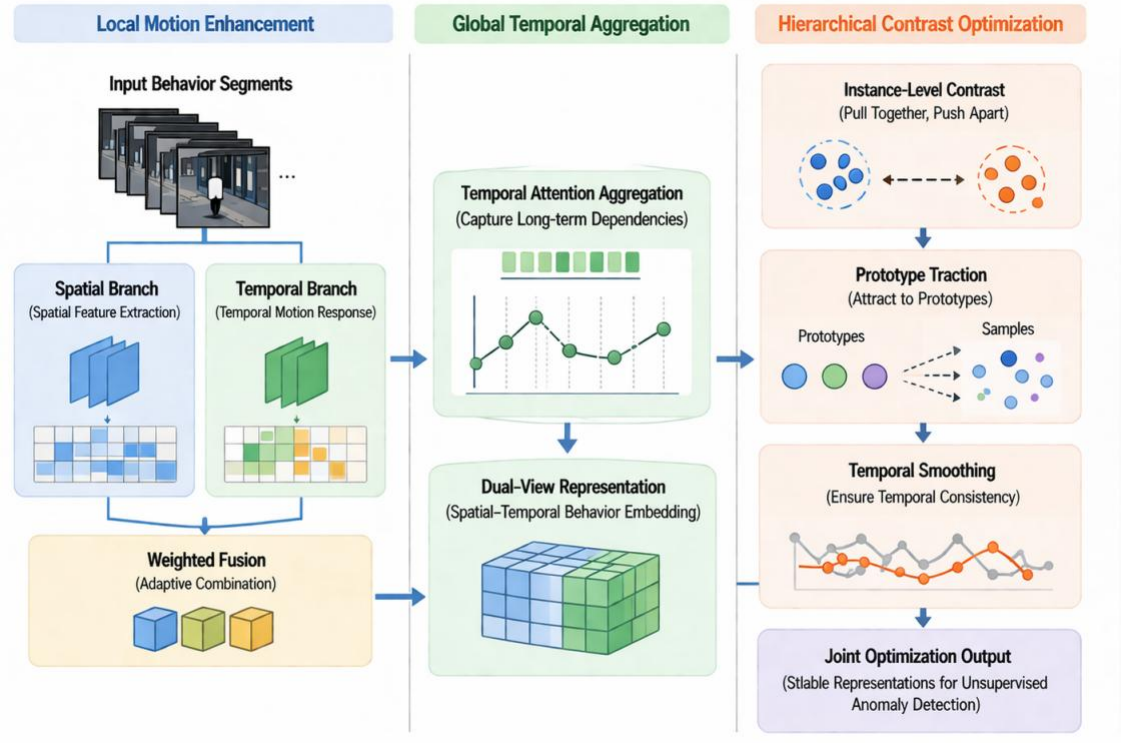

Figure 2. Network Diagram of Spatio-Temporal Feature Encoding and Contrastive Optimization.

Considering that abnormal behaviors often have phased evolution characteristics, this paper further introduces an attention aggregation mechanism in the time dimension to give higher weights to key frames. Let the segment-level representation be zi, then:

$$\alpha_t=\frac{\exp(q^\top u_t)}{\sum_{k=1}^{T}\exp(q^\top u_k)},z_i=\sum_{t=1}^{T}\alpha_t u_t \tag{12}$$

Here, α_t represents the time weight of the t-th frame, and q is the attention query vector. This processing no longer assumes that each frame contributes equally, but enables the model to focus more attention on action segments with more discriminative value such as turning, accelerating, pausing, and aggregating edges, thereby enhancing the response ability of the representation result to key behavioral changes.

In the comparative optimization stage, this paper does not follow the single instance-level comparison method, but constructs a dual constraint of "sample comparison + prototype traction". For the two view representations z_i^a and z_i^b of the same normal segment, the model requires them to maintain high consistency in the embedding space; for the segments in different dynamic subsets, it inhibits their indiscriminate convergence. The instance-level loss is written as:

$$L_{ins}=-\log\frac{\exp(\sin(z_i^a,z_i^b)/\tau)}{\sum_{j=1}^{N}\exp(\sin(z_i^a,z_j)/\tau)} \tag{13}$$

Here, sim(·) represents the cosine similarity, and τ is the temperature coefficient. To prevent the internal structure of normal samples from being too loose, this paper introduces prototype constraints, bringing the current sample closer to the center of the dynamic group p_k:

$$L_{pro}=\| z_i-p_k \|_2^2 \tag{14}$$

 doi:10.20944/preprints202604.1458.v1

Considering that video behavior is continuous, if the fluctuations in adjacent segments are too large, it will often reduce the stability of the detection. Therefore, a time smoothing term is added:

$$L_{tem}=\sum_{i=2}^{n} \| z_i-z_{i-1} \|_2^2 \tag{15}$$

The final optimization objective is:

$$L=\lambda_1 L_{ins}+\lambda_2 L_{pro}+\lambda_3 L_{tem} \tag{16}$$

Here, λ_1, λ_2, and λ_3 are weight coefficients. After the above optimization, the distribution of normal behavior samples in the feature space will be more concentrated, and the representation changes between adjacent segments will be more stable. After this processing, the model can better distinguish normal fluctuations from abnormal deviations, reducing misjudgments caused by local noise, background disturbances, or short-term action changes. At the same time, the encoded results retain more complete spatiotemporal correlation information, which can provide a more stable input for subsequent abnormal score calculation.

3.4. Abnormal Score Calculation and Detection Decision Method

After completing the modeling of normal behavior samples and spatiotemporal representation learning, the core task in the testing phase becomes: how to stably map the abnormal deviations in the segment representations to a determinable numerical result. If only relying on a single distance or a single similarity for recognition, the model is easily affected by viewpoint switching, local occlusion, short-term pauses, and internal fluctuations of normal behavior, resulting in excessive fluctuations in abnormal scores. Considering that video anomalies often manifest as distribution deviations, imbalance in behavior rhythm, and weakened neighborhood support, this paper divides the detection process into three parallel scoring stages, and then obtains the final abnormal score through robust fusion.

Let the test video be segmented by the sliding window to obtain the segment xi. Through the aforementioned encoding network, it is mapped to an embedding representation zi. In the training stage, K behavior prototypes are formed in the normal sample space, denoted as $P=\{p_1,p_2,\ldots,p_K\}$. Due to the differences in the distribution shapes of different normal patterns, this paper does not adopt a uniform spherical distance, but matches the current segment to the nearest prototype based on the statistical deviation degree. The selection method is defined as:

$$k_i^*=arg\min_{1\le k\le K}(z_i-p_k)^T\Sigma_k^{-1}(z_i-p_k) \tag{17}$$

where, Σk represents the covariance matrix of the kth prototype. Thus, the deviation score of the segment relative to the nearest normal pattern can be obtained:

$$D_i=(z_i-p_{k_i}^*)^T\sum_{k_i}^{-1}(z_i-p_{k_i}^*) \tag{18}$$

This reflects "how far the current segment is from the normal behavior center", and can retain the relevant information between the feature dimensions. It is more suitable for describing the degree of discretization of complex behavior representations than the ordinary Euclidean distance.

Just relying on the deviation from the distribution is not sufficient to characterize the abnormal changes in the video. Some segments, although not significantly deviating from the normal prototype, may have discontinuous jumps in the time dimension, such as sudden turns, abnormal pauses, or local accelerations. To describe this dynamic imbalance, this paper introduces a second-order temporal residual term. Based on the representation changes of adjacent windows, it is defined as:

$$T_i=\| z_{i+1}-2z_i+z_{i-1} \|_2 \tag{19}$$

This formula essentially describes the curvature change of the fragment representation along the time axis. When the behavior evolution is relatively stable, T_i is smaller; if there is a significant rhythm

 doi:10.20944/preprints202604.1458.v1

break between the current fragment and the previous and subsequent fragments, this value will increase rapidly. Compared with the first-order difference, the second-order residual is more sensitive to short-term sudden changes, and is less likely to mistakenly identify normal behaviors with slow changes as abnormal.

In addition to prototype deviation and temporal residual, this paper also considers the support strength in the local neighborhood. Normal fragments usually have more similar samples in the training memory bank, while abnormal fragments are often in sparse areas and lack sufficient normal samples for support. Therefore, let be the m nearest neighbors of z_i in the normal memory bank, then the local support deficiency score is defined as:

$$R_i=\frac{1}{m}\sum_{z_j\in N_m(z_i)}\| z_i-z_j \|_2 \tag{20}$$

If R_i is small, it indicates that there is a stable normal neighborhood around the current segment; if Ri is large, it means that the segment is far from the normal sample aggregation area and has a higher possibility of being abnormal. This item can complement the deficiency of prototype centrality measurement in distinguishing edge samples.

Since the scales of the three types of scores are not consistent, this paper uses a robust standardization method to achieve unified mapping. Let Q(·) represent the quantile normalization operator based on the validation set statistics, then the comprehensive abnormal score is written as:

$$A_i=\alpha Q(D_i)+\beta Q(T_i)+\mu Q(R_i) \tag{21}$$

where α, β, and μ are weight coefficients, and they satisfy $\alpha + \beta + \mu = 1$. This formulation avoids one item dominating the overall result due to its excessively large numerical range, allowing the final score to simultaneously absorb information from distribution shift, temporal mutation, and neighborhood sparsity.

In the detection and determination stage, this paper does not use a fixed empirical threshold, but instead constructs a robust boundary based on the median and absolute median difference of the normal validation sample scores. Let the median of the comprehensive score of the normal validation set be Med(A), and the absolute median difference be MAD(A), then the threshold is defined as:

$$\delta=Med(A)+\lambda\, MAD(A) \tag{22}$$

where λ is the adjustment coefficient. Compared to the threshold method based on mean and standard deviation, this formulation is less sensitive to extreme values and is more suitable for the common long-tail distribution in video anomaly detection. The final determination result is:

$$y_i=\begin{cases}1, & A_i\geq\delta\\ 0, & A_i<\delta\end{cases} \tag{23}$$

When $y_i = 1$, the current segment is marked as abnormal behavior. Through the above design, the detection stage no longer relies on a single piece of evidence, but completes the determination from three dimensions: deviation from the normal mode, temporal evolution anomaly, and insufficient local support. The resulting abnormal score has better stability and is more convenient for analyzing the contribution of different modules in subsequent experiments.

4. Experiment and Result Analysis

4.1. Experimental Scheme and Parameter Settings

To verify the effectiveness of the proposed model in the unsupervised behavior anomaly detection task, this paper conducts experiments on three public video anomaly detection datasets: UCSD Ped2, CUHK Avenue, and ShanghaiTech. These datasets differ in terms of scene complexity, target density, background changes, and abnormal behavior types, which can effectively test the model's adaptability and detection stability in different monitoring environments. During the experiment, only normal behavior segments are used for training, and in the testing stage, the input videos are segmented using a sliding window, and the fragment-level anomaly scores are output,

 doi:10.20944/preprints202604.1458.v1

followed by video-level detection determination. To reduce the impact of resolution differences and scene scale differences, all input frames are uniformly adjusted to 224×224. The sample fragment length is set to 16 frames, and the sliding step size is set to 4 frames. The model training uses the Adam optimizer, with an initial learning rate of 1×10^{-4}, a batch size of 32, and a maximum training cycle of 120. The embedding dimension is set to 256, the contrastive learning temperature coefficient is set to 0.07, the number of normal behavior prototypes is set to 8, and the number of neighboring samples is set to 10. The experiment is implemented in Python and PyTorch, with a hardware platform of an Intel i7 processor, 32 GB of memory, and an NVIDIA RTX 3080 graphics card. To reduce the influence of random initialization fluctuations, each group of experiments is repeated 3 times and the average result is taken. The comparison methods are as much as possible implemented publicly and reproduced under the same data partition; the evaluation indicators are mainly frame-level AUC, combined with fragment-level Precision, Recall, and F1 for auxiliary analysis. To ensure the stability of score normalization and threshold setting in the detection stage, a normal validation set is proportionally divided from the training stage's normal samples to estimate the required statistics for quantile normalization and robust threshold parameters. The fragment-level anomaly scores are mapped to the corresponding frame positions according to the window coverage relationship on the time axis, and the frame-level AUC is calculated accordingly; the fragment-level Precision, Recall, and F1 are used to assist in analyzing the model's ability to identify local abnormal fragments. The basic information of the used datasets is shown in Table 3.

Table 3. Basic Information of Experimental Datasets.

Dataset	Training Video Number	Testing Video Number	Resolution	Dataset Characteristics
UCSD Ped2	16	12	360×240	Scene relatively simple, target size small, suitable for testing basic anomaly detection capabilities
CUHK Avenue	16	21	640×360	Scene more diverse, abnormal behavior forms are more numerous, has certain complexity
ShanghaiTech	274	330	Resolution not uniform	Scene quantity is large, background differences are significant, abnormal types are complex, suitable for testing model generalization ability

4.2. Detection Results and Performance Comparison

To comprehensively test the effectiveness of the proposed model, this paper conducts analysis from three aspects: overall detection performance, training convergence stability, and cross-dataset adaptability, and compares it with traditional feature methods, reconstruction-based methods, prediction-based methods, and basic contrastive learning methods.

(1) Overall Detection Performance Comparison

The detection results of different methods on the three datasets are shown in Table 4. Overall, the method proposed in this paper has achieved higher detection accuracy on UCSD Ped2, CUHK Avenue, and ShanghaiTech, indicating that the constructed spatio-temporal encoding and multi-source anomaly scoring mechanism can stably distinguish normal behavior from abnormal behavior. Compared with the baseline contrastive learning method, the proposed method has increased the AUC by 1.6, 2.2, and 3.2 percentage points on the three datasets respectively. The improvement is more significant in complex scenarios, indicating that the introduced prototype constraints and temporal residual terms have a positive effect on the characterization of abnormal boundaries. Especially on the ShanghaiTech dataset, the AUC of the proposed method reaches 83.7%, which is 4.4 percentage points higher than that of MemAE, demonstrating that this method still has good robustness in multi-scenario and strong interference conditions.

 doi:10.20944/preprints202604.1458.v1

Table 4. Comparison of detection performance of different methods on each dataset.

Method	UCSD Ped2/AUC(%)	CUHK Avenue/AUC(%)	ShanghaiTech/AUC(%)	Average AUC(%)	Average F1(%)
Traditional feature method	89.6	81.7	71.8	81.0	75.6
Conv-AE	91.8	84.6	74.9	83.8	80.7
Future Prediction	93.2	86.1	76.8	85.4	82.1
MemAE	95.1	88.4	79.3	87.6	84.5
Baseline contrastive learning method	95.8	89.6	80.5	88.6	85.3
Proposed method	97.4	91.8	83.7	91.0	88.1

(2) Analysis of training convergence and result stability

The convergence changes of the model during the training process are shown in Figure 3.

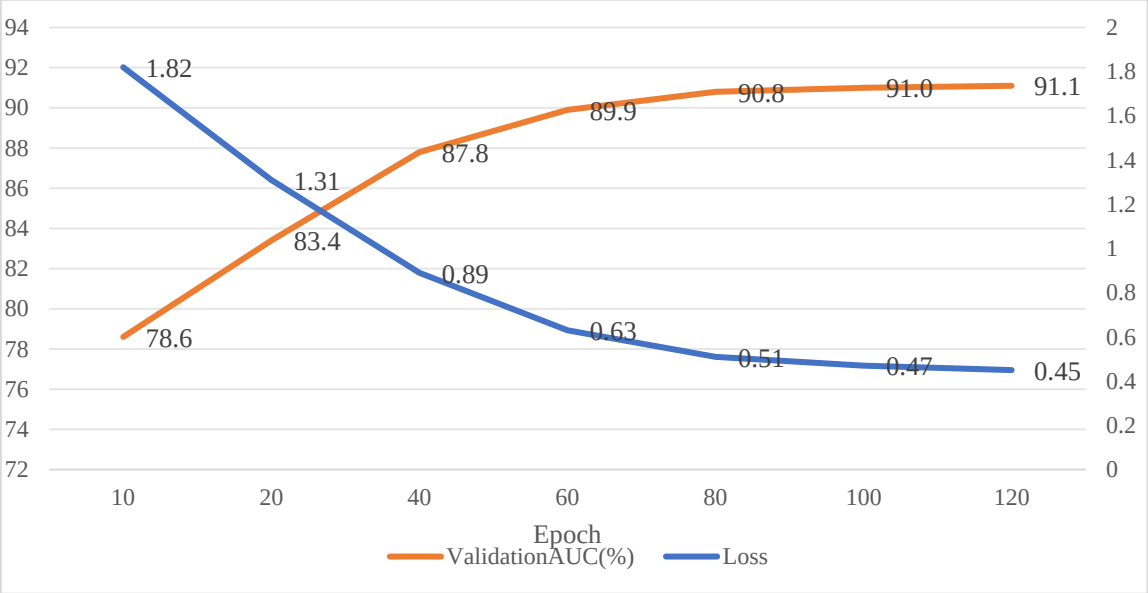

Figure 3. Convergence Curve of Model Training Process.It can be seen that in the first 40 training rounds, the loss value decreased rapidly, and the validation AUC increased simultaneously, indicating that the model was able to establish the basic distribution structure of normal behavior relatively early. After entering the 60th round, the curve became relatively flat, and after 80 rounds, the AUC remained above 90.8%. This indicates that the model did not show significant fluctuations and the training process was relatively stable. At the 120th round, the loss value decreased from 1.82 to 0.45, with a reduction of approximately 75.3%; the validation AUC increased from 78.6% to 91.1%, indicating that the proposed model has good convergence efficiency.

(3) Cross-dataset adaptability analysis

To further test the model's adaptability in different complexity scenarios, the Precision, Recall, and F1 values compared on different datasets are shown in Figure 4.

The Precision, Recall, and F1 scores of the method in this paper on UCSD Ped2 are 95.8%, 94.6%, and 95.2% respectively, on CUHK Avenue they are 91.7%, 89.8%, and 90.7% respectively, and on ShanghaiTech they are 86.4%, 84.1%, and 85.2% respectively. As the complexity of the scene increases, the indicators show a certain decline, but the overall change range is within an acceptable range, indicating that the model still retains a strong recognition ability in more complex video backgrounds and more diverse abnormal types.

 doi:10.20944/preprints202604.1458.v1

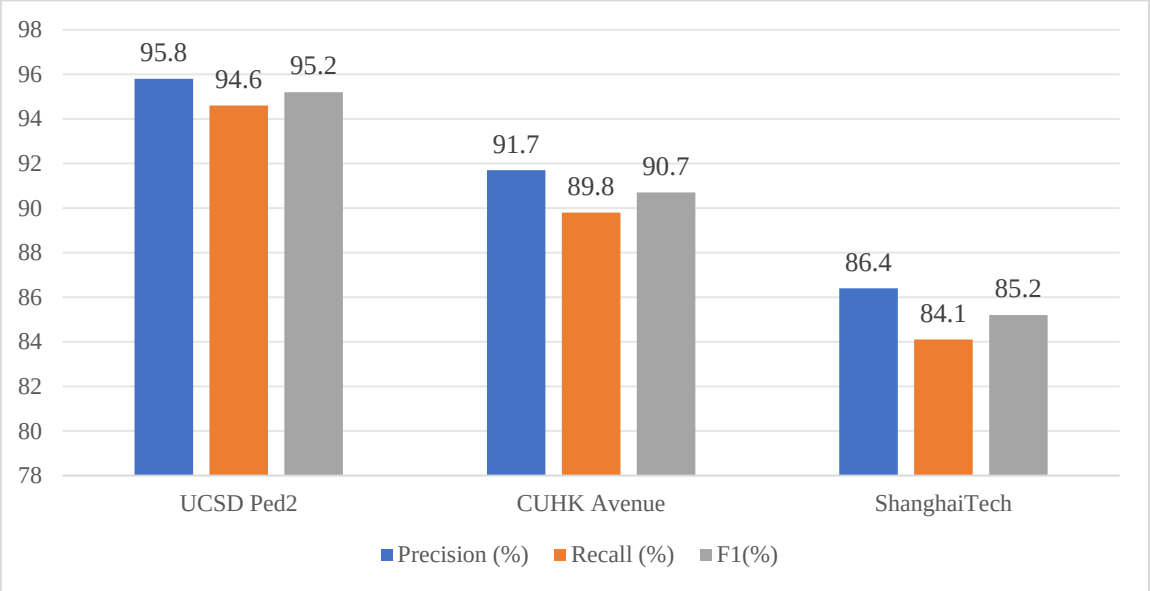

Figure 4. Bar chart comparing fine-grained performance of different datasets.

4.3. Ablation Analysis and Result Discussion

To further determine the specific impact of each module on the detection results, this paper conducts an abandonment experiment under the same training configuration, successively examining the effects of spatio-temporal joint encoding, prototype constraints, second-order temporal residual terms, and local neighborhood support terms. The experiment starts with the basic contrastive learning model and gradually adds different modules on the basis of this, and calculates the average AUC and average F1 on three datasets. The abandonment results are shown in Table 5.

Table 5. Abandonment Experiment Results of Different Module Combinations.

Model Number	Spatio-temporal Joint Encoding	Prototype Constraints	Second-order Temporal Residual	Local Neighborhood Support	Average AUC (%)	Average F1 (%)
M1 Basic Contrastive Learning Model	×	×	×	×	88.6	85.3
M2	√	×	×	×	89.7	86.2
M3	√	√	×	×	90.2	86.8
M4	√	√	√	×	90.7	87.4
M5	√	√	×	√	90.5	87.1
M6 This method	√	√	√	√	91.0	88.1

From Table 5, it can be seen that the basic contrastive learning model already has a certain detection ability, but its results still mainly rely on the global embedding distribution, and do not respond adequately to the subtle changes in complex scenes. After adding spatio-temporal joint encoding, the average AUC increases from 88.6% to 89.7%, and the average F1 increases from 85.3% to 86.2%, indicating that by incorporating spatial structure information and time-evolution clues into the feature representation, the model can better depict the behavior fragments and can distinguish short-term anomalies that were previously easily confused earlier.

On this basis, adding prototype constraints further improves the average AUC to 90.2%. This change indicates that normal behaviors are not simply clustered into a single center during training, but have hierarchical differences and local distribution structures. The role of prototype constraints is to reorganize scattered normal samples into more stable representation regions, making the distance relationship between the test fragment and the normal distribution clearer, and thus

 doi:10.20944/preprints202604.1458.v1

reducing the false alarm rate. For unsupervised anomaly detection, this step is very crucial because it directly affects the separability of subsequent anomaly scores.

The results of M4 and M5 show that the second-order temporal residual term and local neighborhood support term can both bring gains, but their emphasis is not the same. After adding the second-order temporal residual term, the average AUC reaches 90.7%, with a slightly higher improvement than M5 which only added neighborhood support terms. This indicates that in this task, some anomalies are not "unlike normal in appearance", but "do not conform to the normal rhythm of changes", such as sudden pauses, abnormal turns, or local accelerations, which are more easily identified through the discontinuity on the time axis. The contribution of the neighborhood support term mainly lies in stabilizing the judgment results of boundary samples, reducing the false detection of a few normal fragments due to position offset. The complete model M6 achieved the best results under the combined effect of three modules, with an average AUC of 91.0%, an increase of 2.4 percentage points compared to the basic model, and an increase of 2.8 percentage points in average F1. Based on the previous results, it can be concluded that the improvement of this method is not from a single module, but from the collaborative effect of representation modeling, time variation characterization, and local density constraints. Overall, this set of ablation experiments verified the rationality of the model design and also demonstrated that the multi-source anomaly scoring mechanism has good practical value for anomaly recognition in complex video scenarios.

4.4. Interpretability Analysis of Abnormal Concern Regions

In order to further illustrate the attention basis of the model in the judgment of abnormal behavior, this paper selects typical samples of normal and abnormal segments in the testing phase, performs visual analysis on the high-level features of the coding network, and maps their responses back to the original video frames to obtain the visualization results of abnormal attention regions, as shown in Figure 5.

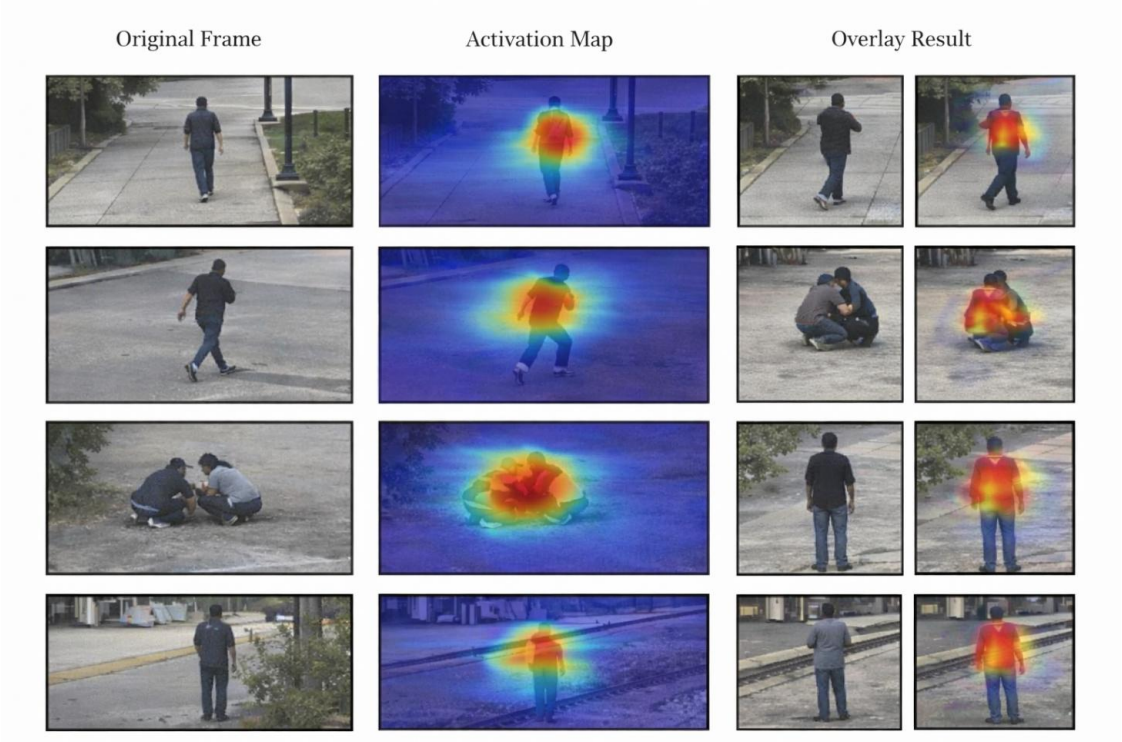

Figure 5. Visualization result of activation map of abnormal attention region.As can be seen from Figure 5, in the normal behavior segment, the high response area of the model is mainly concentrated near the contour of the target subject and its continuous motion trajectory, the overall distribution is relatively compact, and the heat diffusion range is small, indicating that the model has formed a relatively stable feature aggregation for the

 doi:10.20944/preprints202604.1458.v1

normal behavior pattern. When there are abnormal actions, motion direction deviation or local behavior aggregation in the segment, the high response region will obviously concentrate to the location of the abnormal behavior, and form a clearer heat enhancement band along the direction of the action change, while the invalid response of the background region is better suppressed. Especially in complex scenes, if the model can still focus the main response on the abnormal subject and the edge of action change, it shows that the spatio-temporal joint coding and attention aggregation mechanism constructed in the previous section improve the ability of the model to capture key behavior segments, and enhance the adaptability of the model to complex video environments.

5. Conclusions

This paper addresses the issues of scarce abnormal behavior samples, difficult labeling, and ambiguous boundaries of normal patterns in complex monitoring scenarios, and constructs an unsupervised behavior anomaly detection model based on contrastive learning. Compared with the approach that only relies on reconstruction error or short-term prediction deviation, this paper focuses on the stable modeling of normal behavior distribution, through normal sample hierarchical organization, dual-view representation learning, spatio-temporal joint encoding, and time attention aggregation, to retain the key dynamic information of behavior evolution; in the optimization stage, instance-level contrast constraints, prototype traction constraints, and time smoothing constraints are introduced to make the aggregation structure of normal fragments in the embedding space clearer; in the detection stage, a comprehensive anomaly score is constructed by combining the recent prototype deviation, second-order temporal residual, and local neighborhood support strength, and a robust threshold based on the median and absolute median difference is used for determination. After such processing, the model no longer makes judgments based on a single piece of evidence, but identifies abnormal fragments from three aspects: distribution position, time variation, and local density. From the experimental results, this method shows good detection capabilities on three public datasets. Among them, the AUC on UCSD Ped2, Avenue, and ShanghaiTech datasets reached 97.4%, 91.8%, and 83.7%, respectively, with an average AUC of 91.0% and an average F1 of 88.1%. Compared with the basic contrastive learning method, the average AUC increased by 2.4 percentage points and the average F1 by 2.8 percentage points; on the more complex ShanghaiTech dataset, the AUC was 4.4 percentage points higher than MemAE, indicating that this method still maintains good robustness in multi-scenario and strong interference conditions. The ablation experiments further verified the effectiveness of the model design: with the addition of spatio-temporal joint encoding, the average AUC increased from 88.6% to 89.7%; further adding prototype constraints increased it to 90.2%; and with the addition of second-order temporal residual and local neighborhood support terms, the overall performance of the model reached the optimal level. Thus, it can be seen that the results of this paper do not come from a single module, but are formed by the joint effect of representation learning, dynamic modeling, and anomaly scoring strategies.

Overall, the proposed method in this paper effectively addresses the key issues in unsupervised video anomaly detection, such as “large internal differences in normal samples, unstable anomaly boundaries, and insufficient cross-scenario generalization”, and provides a relatively complete implementation path for the application of contrastive learning in behavior anomaly detection tasks. Of course, this paper still has areas that can be further deepened, such as the current model’s characterization of long-time series dependence can be strengthened, the cross-camera scene migration ability still has room for further improvement, and the localization and interpretation ability of more granular anomaly types need to be improved. In the future, it can be combined with lightweight spatio-temporal Transformer, memory enhancement mechanism, and open scene incremental update strategy to further enhance the practical value of the model in edge deployment, online learning, and complex anomaly interpretation.

 doi:10.20944/preprints202604.1458.v1

References

1. Zhang W, Shi H, Qiu J, et al. EdgeAD: Unsupervised Learning Model Based on Prior Knowledge Enhanced Image Anomaly Detection of Heavy Railway Freight Cars[J].IEEE Transactions on Instrumentation and Measurement, 2025(Pt.1):74.DOI:10.1109/TIM.2025.3547481.
2. Li M, Ying Z, Li G, et al. Unsupervised anomaly detection with memory bank and contrastive learning[J].Array,2025,28(c):100548.DOI:10.1016/j.array.2025.100548.
3. Wang H W, Wu R T. Unsupervised anomaly detection for tile spalling segmentation using synthetic outlier exposure and contrastive learning[J].Automation in construction, 2025(Feb.):170.DOI:10.1016/j.autcon.2024.105941.
4. Wu W, Gu Y. Advancing unsupervised graph anomaly detection: A multi-level contrastive learning framework to mitigate local consistency deception[J].Neurocomputing, 2025, 646.DOI:10.1016/j.neucom.2025.130507.
5. Zhu W, Li W, Dorsey E R, et al. Unsupervised anomaly detection by densely contrastive learning for time series data[J].Neural Networks, 2023, 168:450-458.DOI:10.1016/j.neunet.2023.09.038.
6. Ghanim J, Awad M. An Unsupervised Anomaly Detection in Electricity Consumption Using Reinforcement Learning and Time Series Forest Based Framework[J].JOURNAL OF ARTIFICIAL INTELLIGENCE AND SOFT COMPUTING RESEARCH, 2025, 15(1):20.DOI:10.2478/jaiscr-2025-0001.
7. Seo J, Kim Y, Ha J, et al. Unsupervised anomaly detection for earthquake detection on Korea high-speed trains using autoencoder-based deep learning models[J].Scientific Reports, 2024, 14(1).DOI:10.1038/s41598-024-51354-7.
8. Li R, Ma H, Wang R, et al. Application of unsupervised learning methods based on video data for real-time anomaly detection in wire arc additive manufacturing[J].Journal of manufacturing processes,2025(Jun.):143.DOI:10.1016/j.jmapro.2025.03.113.
9. Ding F, Li B, Ben X, et al. ALAD: A New Unsupervised Time Series Anomaly Detection Paradigm Based on Activation Learning[J].IEEE Transactions on Big Data, 2025, 11(3):1285-1297.DOI:10.1109/TBDATA.2024.3453762.
10. Wang X, Wang Y, Dai Y, et al. UDTL: Anomaly Detection Based on Unsupervised Deep Transfer Learning[J].2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD),2024:2650-2655.DOI:10.1109/cscwd61410.2024.10580439.
11. Pu Y, Sun J, Tang N, et al. Self-supervised distributional and contrastive learning model for image anomaly detection[J].Knowledge-Based Systems,2025,316.DOI:10.1016/j.knosys.2025.113348.
12. Wang X, He J, Huang F, et al. Abnormal cell cause localization based on contrastive pre-training and unsupervised data-driven model for lithium-ion battery manufacturing[J].Journal of Energy Storage, 2024(Nov.Pt.A):101.DOI:10.1016/j.est.2024.113743.
13. Wang L, Cheng Y, Gong H, et al. Research on Dynamic Data Flow Anomaly Detection based on Machine Learning[J].2024 3rd International Conference on Electronics and Information Technology (EIT), 2024:953-956.DOI:10.1109/eit63098.2024.10762641.
14. Ho W J, Hsieh H Y, Tsai C W. Anomaly Detection Model of Time Segment Power Usage Behavior Using Unsupervised Learning[J].Journal of Internet Technology, 2024(3):25.
15. Vyshkvarkova E V, Grekov A N, Kabanov A A, et al. Anomaly Detection in Biological Early Warning Systems Using Unsupervised Machine Learning[J].Sensors (Basel, Switzerland), 2023, 23(5):2687-2687.DOI:10.3390/s23052687.
16. Hiruta T, Maki K, Tetsuji K, et al. Unsupervised Learning Based Diagnosis Model for Anomaly Detection of Motor Bearing with Current Data[J].Procedia CIRP, 2021, 98:336-341.DOI:10.1016/J.PROCIR.2021.01.113.
17. Fan C. Unsupervised anomaly detection based on improved skip-gannomaly[J].Proceedings of SPIE, 2022, 12348(000):7.DOI:10.1117/12.2641921.
18. Ea P, Vo Q, Salem O, et al. Unsupervised Anomaly Detection in IoMT Based on Clustering and Online Learning[J].2024 IEEE International Conference on E-health Networking, Application & Services (HealthCom), 2024:1-6.DOI:10.1109/healthcom60970.2024.10880810.

 doi:10.20944/preprints202604.1458.v1

19. Wang X, Bian W, Zhao X. Robust Unsupervised Anomaly Detection for Surface Defects Based on Stacked Broad Learning System[J].IEEE/ASME Transactions on Mechatronics, 2024:1-11.DOI:10.1109/tmech.2024.3465563.
20. Xu Q, Xie T, Jiang C, et al. Adaptive Working Condition Recognition With Clustering-Based Contrastive Learning for Unsupervised Anomaly Detection[J].IEEE transactions on industrial informatics,2024(10):20.DOI:10.1109/TII.2024.3413952.
21. Wu L, Ali M K M, Tian Y. Supervision and early warning of abnormal data in Internet of Things based on unsupervised attention learning[J].Computer communications, 2024, 216(Feb.):229-237.DOI:10.1016/j.comcom.2023.12.043.
22. Lu J, Cao Y, Shi R, et al. An Efficient Driver Anomaly State Detection Approach Based on End-Cloud Integration and Unsupervised Learning[J].2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC),2023:5824-5830.DOI:10.1109/itsc57777.2023.10422424.
23. Xi Y, Lei Z, Wen G, et al. Unsupervised Fault Detection Method via Time-Series Segmentation and Contrastive Masking Learning[J].Instrumentation and Measurement, IEEE Transactions on, 2025, 74(000):1-10.DOI:10.1109/TIM.2025.3568963.
24. Shang X, Zhang J, Jiang X, et al. Anomaly Detection for Multivariate Time Series Based on Contrastive Learning and Autoformer[J].2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2024:2614-2619.DOI:10.1109/cscwd61410.2024.10580672.
25. Xiao P, Jia T, Duan C, et al. LogCAE: An Approach for Log-based Anomaly Detection with Active Learning and Contrastive Learning[J].2024 IEEE 35th International Symposium on Software Reliability Engineering (ISSRE), 2024:144-155.DOI:10.1109/issre62328.2024.00024.
26. Hamza A, Ali Z, Dudley S, et al. Optimizing PV Array Performance: A2 LSTM for Anomaly Detection and Predictive Maintenance based on Machine Learning[J].2024 IEEE Energy Conversion Congress and Exposition (ECCE), 2024:1681-1688.DOI:10.1109/ecce55643.2024.10861733.
27. Song D, Lee N, Kim J, et al. Anomaly Detection of Deepfake Audio Based on Real Audio Using Generative Adversarial Network Model[J].IEEE Access,2024,12:184311-184326.DOI:10.1109/access.2024.3506973.
28. Qian J, Wu Z, Cao Y, et al. Unsupervised anomaly detection for radar active deception jamming based on denoising diffusion implicit model[J].IET Conference Proceedings, 2024, 2023(47):2454-2457.DOI:10.1049/icp.2024.1471.
29. Li S, Song W, Zhao C, et al. An Anomaly Detection Method for Multiple Time Series Based on Similarity Measurement and Louvain Algorithm[J].Procedia Computer Science,2022,200(c):1857-1866.DOI:10.1016/j.procs.2022.01.386.
30. Lei Y, Nieuwoudt M, Matsumoto H, et al. Unsupervised Anomaly Detection for Mild Cognitive Impairment Using Diffusion Model[J].2023 Asia Conference on Cognitive Engineering and Intelligent Interaction (CEII), 2023:41-46.DOI:10.1109/ceii60565.2023.00016.
31. Roseline S A, Karthik S, Sruti I N V D. Intelligent Human Anomaly Detection using LSTM Autoencoders[J].2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), 2024:1-7.DOI:10.1109/accai61061.2024.10602454.
32. Natsumeda M, Mizoguchi T, Cheng W, et al. Unsupervised anomaly detection under a multiple modeling strategy via model set optimization through transfer learning[J].2023 26th International Conference on Information Fusion (FUSION),2023:1-8.DOI:10.23919/fusion52260.2023.10224107.
33. Song S, Yang K, Wang A, et al. A Mura Detection Model Based on Unsupervised Adversarial Learning[J].IEEE Access, 2021, PP(99):1-1.DOI:10.1109/ACCESS.2021.3069466.
34. Ali M, Scandurra P, Moretti F, et al. Anomaly Detection in Public Street Lighting Data Using Unsupervised Clustering[J].IEEE Transactions on ConsumerElectronics,2024(1):70.DOI:10.1109/TCE.2024.3354189.
35. He Y, Ding X, Tang Y, et al. Unsupervised Multivariate Time Series Anomaly Detection by Feature Decoupling in Federated Learning Scenarios[J].IEEE Transactions on Artificial Intelligence,2025:1-15.DOI:10.1109/tai.2025.3533437.
36. Qiu H, Jiang H. RLIF-Net:Unsupervised Trace-SPC Fault Detection Solution Based on Representation Learning and Isolation Forest[J].2024 2nd International Symposium of Electronics Design Automation (ISEDA), 2024:552-557.DOI:10.1109/iseda62518.2024.10618034.

 doi:10.20944/preprints202604.1458.v1

37. Duan M, Mao L, Liu R, et al. Unified Model Based on Reinforced Feature Reconstruction for Metro Track Anomaly Detection[J].IEEE sensors journal, 2024(4):24.DOI:10.1109/JSEN.2023.3348118.