Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

MSTAgent-VAD: Multi-scale video anomaly detection using time agent mechanism for segments' temporal context mining

Shili Zhao[a,b,c,d], Ran Zhao[a,b,c,d,*], Yan Meng[a,b,c,d,1], Xiaoyi Gu[a,b,c,d,2], Chen Shi[a,b,c,d], Daoliang Li[a,b,c,d,3]

[a] *National Innovation Center for Digital Fishery, China Agricultural University, Beijing 100083, China*
[b] *Key Laboratory of Smart Farming Technologies for Aquatic Animal and Livestock, Ministry of Agriculture and Rural Affairs, China Agriculture University, Beijing 100083, China*
[c] *Beijing Engineering and Technology Research Centre for Internet of Things in Agriculture, China Agriculture University, Beijing 100083, China*
[d] *College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China*

ARTICLE INFO

Keywords:
Surveillance video anomaly detection
Multi-scale temporal attention
Time agent mechanism
Anomaly scoring strategy
Temporal context mining

ABSTRACT

Due to the lack of frame-level annotations during training, video anomaly detection (VAD) requires developing learning methods without comprehensive supervision. Previous approaches have focused on modeling temporal relationships and learning discriminative features but often struggle with incomplete anomaly detection and weak segment separation. To address these issues, we propose a multi-scale VAD method using a time agent mechanism, called MSTAgent-VAD, which achieves significant innovation in method structure and feature learning. Firstly, in view of the diversity of anomaly events in videos on temporal scales, we design a multi-scale temporal attention module to capture temporal features of abnormal segments of varying lengths, enhancing temporal consistency and addressing limitations in detecting anomalies of diverse durations. Secondly, by generating temporal agent tokens with deformable convolution, the time agent mechanism can strengthen the distinction and improve separation between normal and abnormal segments in feature space, especially for incomplete anomalies and blurred boundaries, thus enhancing model discrimination. Finally, based on the multi-instance learning (MIL) strategy, an improved robust temporal feature magnitude (RTFM) learning method is used to detect multiple discrete abnormal segments, which solves the challenge that traditional methods are difficult to identify diverse anomalies in complex scenes and ensures the accuracy of detecting multiple types of anomaly events. Experimental results show that our method achieves state-of-the-art detection performance on the UCSD-Ped2, CUHK Avenue, ShanghaiTech and UCF-Crime datasets, accurately identifying diverse anomalies and showing strong generalization. This study provides an innovative VAD solution for surveillance applications, improving detection performance in real-world scenarios.

1. Introduction

In recent years, the rapid development of society has led to the emergence of various security threats and emergencies in daily life, highlighting the importance of video anomaly detection (VAD). VAD refers to the process of identifying abnormal data in expected normal data with large deviation from the normal data, so as to identify the events that do not conform to the expected behaviors, aiming to detect and locate the abnormal events automatically (Bai et al., 2021). As a crucial domain in computer vision applications, VAD has garnered significant attention from both academia and industry due to its potential application in autonomous surveillance systems (Sultani et al., 2018; Wu

* Corresponding author at: P. O. Box 121, China Agricultural University, 17 Tsinghua East Road, Beijing 100083, China.
E-mail addresses: slizhao@cau.edu.cn (S. Zhao), ran.zhao@cau.edu.cn (R. Zhao), s20233081756@cau.edu.cn (Y. Meng), S20233081784@cau.edu.cn (X. Gu), sc@cau.edu.cn (C. Shi), dliangl@cau.edu.cn (D. Li).
[1] 0009-0000-3909-9894.
[2] 0009-0000-9621-8394.
[3] 0000-0002-1842-419X.

https://doi.org/10.1016/j.eswa.2025.127154
Received 5 November 2024; Received in revised form 13 February 2025; Accepted 3 March 2025
Available online 8 March 2025
0957-4174/© 2025 Published by Elsevier Ltd.

et al., 2021; Yang et al., 2023). However, challenges such as complex backgrounds, high-dimensional data, and unclear definitions of anomalies make modeling unseen anomalies difficult, rendering VAD a challenging task (Wang et al., 2022).

The goal of VAD is to identify the time window in which the abnormal event occurs. The traditional detection method is to manually perform anomaly identification in the video. However, with more and more surveillance videos, manual detection will consume a lot of manpower and may also cause problems such as omission or error detection. With the development of computer technology, some methods that perform classical anomaly detection techniques based on manual features have been proposed (Cong et al., 2011; Wang & Miao, 2010). They model the action features and appearance features of the video data and thus identify abnormal behaviors. However, this method cannot effectively capture the discriminative information of anomalies, and the complexity of the feature engineering cannot be widely applied. Deep learning-based methods have dominated the VAD field in recent years (Chang et al., 2020; Gong et al., 2022; Reiss & Hoshen, 2022; Wan et al., 2020; Wang et al., 2022; Wu et al., 2023). Existing deep learning-based VAD methods are mainly divided into unsupervised anomaly detection methods and weakly supervised anomaly detection methods. Unsupervised anomaly detection methods discriminate data by learning the probability distribution of normal behavior data, and use unsupervised classifiers specifically trained with normal videos for anomaly detection (Doshi & Yilmaz, 2021; Huang et al., 2023; Lv et al., 2021; Wang et al., 2022; Wu et al., 2023; Yang et al., 2023). But this class of methods is only trained on normal data. In real-life scenarios, there are a wide variety of abnormal events, which are far less than the normal events. Therefore, lack of perception of abnormal data limits the detection performance of the model. Compared to unsupervised VAD methods, weakly supervised anomaly detection using training samples annotated with normal or abnormal video-level labels performs better (Gong et al., 2022; Sultani et al., 2018; Tian et al., 2021; Wan et al., 2020; Zhong et al., 2019). This weakly supervised setup achieves better anomaly classification accuracy at the cost of relatively less manual annotation work. Most of the existing weakly supervised methods are based on the multi-instance learning (MIL) paradigm, where only video-level annotations are given in the training set to indicate whether the video contains anomalies or not, without detailed abnormal information. In the testing phase, the model needs to predict an anomaly score between 0 and 1 for each frame.

MIL achieves VAD by balancing an equal number of abnormal and normal segments. However, most MIL-based approaches suffer from the following limitations: (1) Equal-length segments overlook the duration of anomalies, leading to incomplete representation of abnormal events. (2) Insufficient temporal correlation prevents effective separation of normal and abnormal segments. (3) A tendency to detect only the most unusual video segments while neglecting the remaining abnormal segments. Some methods (Deshpande et al., 2022; Le and Kim, 2023; Sun and Gong, 2024; Zhang et al., 2023), through attention mechanisms or multi-scale modeling strategies, better adapt to the temporal variations of abnormal events, thereby overcoming the limitations of static time segmentations and addressing the issue of neglecting anomaly duration caused by equal-length segment divisions. However, these methods face challenges when dealing with complex scenarios where the duration of abnormalities varies significantly, especially when the length of abnormal behaviors is inconsistent; in such cases, the model may lose crucial information. To address the issue of insufficient temporal correlation, some studies (Chen et al., 2023; Zhang et al., 2024; Zhong et al., 2022) use spatio-temporal encoders or graph convolutional networks to capture the spatio-temporal dependencies between video segments, which can effectively enhance the separation between normal and abnormal segments. Nevertheless, these methods may struggle with overly relying on local information while neglecting the global temporal structure when handling long time spans or complex temporal dependencies. Regarding the problem of tending to detect the most significant anomalies while neglecting other types, existing studies (Georgescu et al., 2021; Sun & Gong, 2023; Tian et al., 2022; Wang et al., 2020) have improved model sensitivity to more subtle or marginal anomalies through contrastive learning or multi-task learning frameworks. Despite enhancing the comprehensiveness of anomaly detection, these methods still have limitations, especially when dealing with multiple types of abnormal events, where some abnormal segments might be overlooked or misclassified.

To address the above problems, we propose a multi-scale VAD method based on a time agent mechanism, called MSTAgent-VAD. Specifically, to overcome the issues caused by equal-length segment divisions, we have designed a multi-scale temporal attention module that can flexibly adapt to context representations at different time scales. This module extracts rich contextual and detailed information from segments of varying lengths, enabling the model to more accurately capture the persistence and feature details of abnormal events. It enhances the feature learning ability of anomalous segments and avoids the problem of incomplete anomaly representation due to the equal-length segment limitation in traditional methods. To address the challenge of insufficient temporal correlation, we propose a time agent mechanism. Temporal agent tokens, generated through deformable convolution, can flexibly capture dynamic correlations between segments. These tokens are then processed through the self-attention mechanism of an agent Transformer, allowing the model to capture temporal dependencies and improving the separation between normal and anomalous segments in the feature space, thereby enhancing the model's sensitivity to temporal information. To further improve model performance, particularly when handling videos containing multiple discrete abnormal segments, we adopt the existing MIL-based RTFM method, as traditional MIL methods often overlook smaller anomalies. RTFM effectively handles multiple anomalous segments in videos. To capture the subtle differences between normal and abnormal video segments and enhance sensitivity to abnormal information, we introduced a category contrast loss to optimize the RTFM anomaly scoring strategy, thereby maximizing the distinction between abnormal and normal videos. Moreover, we further improved the model by incorporating temporal smoothing and sparse constraints, which enable it to better capture multiple abnormal segments within a video. These improvements lead to more accurate detection of complex videos containing discrete anomalies, enhancing the model's ability to identify all abnormal segments and significantly improving its robustness and reliability. The main contributions of this paper are as follows:

(1) Multi-scale Temporal Attention Module: We propose a module that flexibly captures segment features at different temporal scales, enhancing the model's ability to represent and detect various abnormal events by leveraging richer temporal context.
(2) Time Agent Mechanism: We use deformable convolution to generate agent tokens that adaptively refine the feature distribution of normal and abnormal segments, forming clearer category boundaries. This mechanism enhances the model's ability to capture temporal correlations and improves detection accuracy for diverse abnormal segments.
(3) RTFM Improvement: We enhance RTFM by incorporating contrastive learning, temporal smoothing, and sparsity constraints to promote robust learning of feature amplitudes, thus enabling accurate detection of complex and heterogeneous anomalous events across diverse datasets. Experiments on four widely used datasets (UCSD-Peds, CUHK Avenue, ShanghaiTech and UCF-Crime) show that our method outperforms several state-of-the-art methods.

2. Related work

2.1. Unsupervised video anomaly detection

Due to the unpredictable nature of abnormal events, it is challenging to collect comprehensive data, and the cost of frame-level annotation for videos is high. Consequently, unsupervised VAD has been proposed. The abnormal video detection methods based on unsupervised learning do not rely on video annotation information. They only utilize the similarity among sample data to train normal samples. During testing, videos that deviate from the normal distribution are classified as anomalies, facilitating effective anomaly detection (Ramachandra et al., 2020). Unsupervised anomaly detection is mainly categorized into two types: reconstruction-based methods and prediction-based methods.

The reconstruction-based methods follow the assumption that abnormal events cannot be accurately reconstructed and expressed on a model trained with only normal data. In particular, the reconstruction or prediction error will be larger for irregularly distributed data during testing. Since then, methods based on autoencoders to model regular patterns and reconstruct video frames have been proposed. Full convolutional autoencoders, spatio-temporal convolutional autoencoders, and memetic autoencoders, etc. have been widely used for abnormal event detection (Li et al., 2020; Liu et al., 2021). These methods work on well-designed reconstruction models that implicitly model the context and learn robust normal spatio-temporal patterns with the help of encoding–decoding capabilities of deep networks. In the recent approach, Tur et al. (2023) first used compact motion representation of motion for VAD, using a conditional diffusion model to extract motion and appearance features of a given video segment. The method utilized data-driven thresholding and considers high reconstruction error as an indicator of abnormal events, showing better generalization performance across different datasets. Al-Lahham et al. (2024) proposed a two-stage coarse–fine pseudo label (C2FPL) generator framework, which generated segment-level pseudo labels, which can be further used to train a segment-level anomaly detector in a supervised manner. The trained anomaly detector is able to identify abnormal video segments from a set of completely unlabeled videos and obtain segment-level and subsequent frame-level anomaly predictions. Wang et al. (2022) proposed a new self-encoder model for augmented sequential frame reconstruction, the spatio-temporal rotary encoder, to bring the reconstruction-based approach back into play. The method employed both input perturbation technology and object-level reconstruction to make the test reconstruction error of normal frames smaller than that of abnormal frames.

The prediction-based methods follow the assumption that normal behaviors in videos are regular events, while abnormal behaviors are unpredictable irregular events. Future frame prediction is performed by learning the intrinsic representation and continuous context dependencies in normal videos. However, abnormal videos often violate these dependencies, resulting in unpredictable future frames. The anomaly detection accuracy of this type of method is usually higher than that of reconstruction-based detection methods. Liu et al. (2018) was the first to utilize a prediction-based future framework for VAD by generating an adversarial network for future frame prediction while introducing motion (temporal) constraints in addition to appearance (spatial) constraints for future prediction of normal events. Ye et al. (2019) proposed a deep predictive coding network for generating future frames of videos, and achieved VAD through a predictive coding module and an error optimization module. Zaheer et al. (2022) proposed a new unsupervised generative cooperative learning method for VAD. This method used the low frequency of anomalies to establish cross-supervision between the generator and the discriminator, and achieved VAD through cooperative training. Zhao et al. (2022) used a spatio-temporal long and short-term memory network to extract and memorize the spatial appearance and temporal changes in a unified memory unit, while introducing a discriminator with spatiotemporal long short-term memory for adversarial learning to enhance learning ability. Li et al. (2023) proposed a new adversarial composite prediction framework for future video frames, which uses an adversarial learning and bidirectional composite prediction strategy to normal video dynamics for rational modelling and learning a more compact representation of normal video dynamics to improve the detection sensitivity of motion anomalies. There are also some methods that combine frame reconstruction methods with frame prediction methods to achieve better results (Chang et al., 2022; Liu et al., 2021; Morais et al., 2019; Tang et al., 2020). However, these methods are all based on pixel-level generation, and the model has high computational cost and difficulty in training, so much so that they cannot meet the needs of real application scenarios.

2.2. Weakly supervised video anomaly detection

Unsupervised learning methods effectively reduce the computational complexity of VAD, but due to the complexity of the scenes, the detection accuracy is generally lower. To address this issue, weakly supervised video anomaly detection (WVAD) has emerged, which requires only video-level labels during training and relies on frame-level labels during testing. However, one of the main challenges faced by WVAD is the lack of temporal judgment for anomaly events. To tackle this challenge, existing methods adopt different strategies to identify potential anomalous segments in labeled abnormal videos, improving the discriminative power between abnormal and normal features by designing more effective loss functions (Pu & Wu, 2022; Sapkota & Yu, 2022) or modeling temporal context relationships (Chen et al., 2023; Huang et al., 2022). Additionally, Zhou et al. (2024) proposed using the divergence of the mean vector from BatchNorm as an anomaly criterion and designed a batch-level selection strategy to filter anomalous segments. Zhang et al. (2023) introduced a dynamic erasure network (DENet) based on multi-scale temporal modeling to evaluate the completeness of anomalous segments, but these methods still struggle to effectively handle the subtle visual differences between abnormal and normal events. In Chen et al. (2023), the authors pointed out that the feature magnitude representing the degree of anomaly often overlooks the impact of scene changes, proposing a feature amplification mechanism and magnitude contrast loss to enhance the discriminability of features for anomaly detection. In terms of temporal modeling, the transformer-style temporal feature aggregator and self-guided feature encoder improve feature discriminability (Huang et al., 2022) and enhance the accuracy of detecting anomalous frames (Zhou et al., 2023). Furthermore, some studies (Fan et al., 2024; Feng et al., 2021) generate pseudo-time annotations or pseudo-labels to provide supervision signals, helping to distinguish abnormal and normal events. By using generated video anomaly scores as pseudo-labels for video segments and employing anomaly attention mechanisms and multi-branch supervision modules to address the lack of frame-level labels during training, these methods improve performance. However, the quality and accuracy of pseudo-label generation remain key factors, and ensuring robustness in complex scenarios remains challenging.

Some existing methods (Park et al., 2023; Wu et al., 2021), inspired by MIL, address the issue of lacking annotated anomalous segments, especially when noisy labels are present. Specifically, these methods group video segments into bags and assign bag-level labels, thereby effectively reducing the need for large-scale frame-level annotations. Sultani et al. (2018) were the first to introduce an MIL framework with a ranking loss function, learning abnormal patterns by calculating score differences between normal and abnormal bags. Building on this, Zhu and Newsam (2019) added an attention mechanism, incorporating learned temporal context information into the MIL model to help the model better differentiate between normal and abnormal data. Yu et al. (2021) optimized the training approach of previous MIL methods to improve loss calculation accuracy. However, these methods still face limitations when handling noise in positive bags, which may lead to misclassification of normal segments as anomalous. To address this, Zhong et al. (2019) proposed a noise removal method based on Graph Convolutional Networks (GCN), though it incurs high training overhead due to the presence of action classifiers. The RTFM model resolves this issue with an L2-norm-based ranking loss function, but still uses conventional CNNs to extract video features, resulting in insufficient latent space constraints, which affects accuracy (Tian et al., 2021). To address this, temporal and dilated convolutions are employed to enhance the feature extraction capability of video segments (Wan et al., 2020; Zhang et al., 2019). Furthermore, MIL-based methods require complete video input during each training iteration, and the data correlation significantly impacts the training of anomaly detection networks. To minimize this correlation, CLAWS Net adopts a random batch selection approach, reducing data correlation by randomly selecting temporally consistent batches (Zaheer et al., 2020). Li et al. (2022) introduced multi-sequence learning for the first time, incorporating a self-training method for the iterative update of anomaly score prediction networks based on Transformer layers.

However, existing MIL methods perform poorly when handling anomalous segments with subtle visual differences, especially when these anomalous segments are rare and similar to normal segments. Additionally, most methods overlook the temporal dependencies within videos, leading to poor performance when processing sequential information. To overcome these issues, this paper employs the VideoSwin Transformer backbone network to extract video segment features and designs a multi-scale video anomaly detection method based on the time agent mechanism. This approach effectively addresses the issues of incomplete anomaly event detection and weak segment separation, significantly improving the accuracy of anomalous segment detection.

3. Methodology

3.1. Video anomaly detection problem description

Given a set of training videos, defined as $\boldsymbol{B}$ bags (i.e., videos) $\{V_1, V_2, \ldots, V_B\}$, and video-level labels $Y_i \in \{0,1\}$ for each video $V_{i,}$ for which $Y_i = 1$ if there is at least one abnormal instance in the video V_i; otherwise, $Y_i = 0$. Each bag contains t video segments, represented as $V_i = \{v_{i,1}, v_{i,2}, \ldots, v_{i,j}, \ldots, v_{i,t}\}$, $i \in \{0,B\}$ and $j \in \{0,t\}$. The goal of VAD is to build a model that predicts a frame-level anomaly score for each instance $v_{i,j}$ in each test video, i.e., $S_i = \{s_{i,1}, s_{i,2}, \ldots, s_{i,j}, \ldots, s_{i,t}\}$, where $s_{i,j} \in [0,1]$, so as to judge whether the current video segment is abnormal based on the anomaly score.

3.2. Method description

This paper proposes a weakly supervised learning anomaly detection method based on video-level labels, which is divided into three main stages (as shown in Fig. 1), and multiple modules are designed to improve the detection accuracy and robustness of abnormal events. Each video is divided into a fixed number of segments to ensure that each segment contains the same number of frames even if the video duration is different. This method assumes that the segments obtained from the abnormal video contain at least one abnormal segment, while the segments obtained from the normal video contain all normal segments. The first stage is video feature extraction and multi-scale feature modeling. In this stage, each video segment is processed using a pre-trained video feature extraction model to extract rich spatial and semantic features from it. These features contain the basic information within the segment and are input into a multi-scale temporal attention module. This module captures short-term and long-term temporal features from different time scales. By combining dilated convolutions and paired self-attention, it handles anomalous events across various time spans, enhancing the fusion of local and global features, thereby improving the model's ability to perceive anomalies of different durations. The second stage involves temporal feature modeling using the time agent mechanism. The time agent mechanism operates at the feature space level, generating temporal agent tokens from multi-scale contextual features using deformable convolutions, which enhance the temporal dependencies between video segments. These agent tokens can adaptively focus on key temporal information from anomalous segments and enhance anomaly detection capabilities through the self-attention mechanism, especially when handling boundary-fuzzy or incomplete anomalous segments, thus improving the discriminative representation of anomalous segments. The core self-attention mechanism enables both modules to enhance the effectiveness of video anomaly detection in multiple dimensions. The third stage is anomaly detection and optimization. In the final stage, the enhanced RTFM model is used to detect anomalies on the extracted segment features. To further optimize detection performance, we introduce contrastive loss, temporal smoothing, and sparsity constraint loss. These improvements ensure that anomaly detection remains stable over time, reduce sensitivity to irrelevant features, and enhance the model's overall robustness.

3.2.1. Multi-scale temporal attention module

In this paper, the pre-trained VideoSwin Transformer model is used for spatial semantic feature extraction. The model is pre-trained on large-scale datasets such as ImageNet and can effectively capture rich spatial information in video segments. By dividing the image into multiple windows and calculating self-attention in each window, the model greatly reduces the computational cost of the traditional Transformer model, which is particularly suitable for processing long videos. VideoSwin Transformer uses a sliding window mechanism to capture local features of each window, so that the model can not only focus on local details, but also form global associations in the entire image sequence. This feature solves the problem of insufficient spatial features in anomaly detection, allowing the model to more comprehensively perceive abnormal objects and background features in the video screen, and improves the ability to identify abnormal behaviors in complex

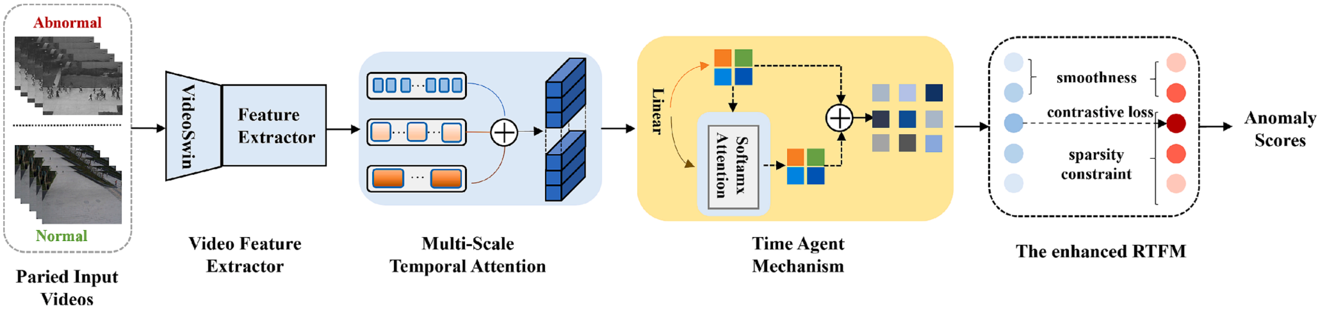


Fig. 1. The overall architecture of our proposed network. First, video features are extracted from input video frames using the pre-trained VideoSwin Transformer backbone. Next, the multi-scale temporal attention module is employed to model temporal dependencies at various time scales. Following this, the generated temporal agent token mechanism is used to enhance the modeling of cross-segment temporal relationships. Finally, the enhanced RTFM with contrastive learning, smoothness and sparsity regularization is applied for anomaly scoring.

scenesa.

According to the problem description, we regard every 16 frames in each segment as a snippet from t equal-length segments in each video. Then, the pre-trained VideoSwin Transformer network is used to extract features from each segment. Each video has RGB channels, so as to obtain the input dimension $\boldsymbol{F} = C \times T \times H \times W$, where T is the number of snippets and C is the number of channels. We input the extracted spatial features into the multi-scale temporal attention module to obtain local and global context information at different time scales. The multi-scale temporal attention module is shown in Fig. 2.

In video anomaly detection, abnormal events may manifest as different features at multiple time scales (such as short-term abnormal actions or continuous abnormal activities), so the model needs to be able to flexibly adapt to various time scales. The multi-scale temporal attention module flexibly captures local and global information in short and long time periods through a set of dilated convolution operations, that is, by inserting holes (i.e., zeros) between the elements of the standard convolution kernel, allowing the model to refine features at different time scales. This module effectively increases the receptive field through one-dimensional convolution with different dilation factors, ensuring the acquisition of high-detail features in a short time period while capturing richer temporal context in the long time dimension. This design solves the problem that traditional anomaly detection models are not able to adapt to the temporal scale of abnormal events, and ensures efficient perception of abnormal events with different time spans. Specifically, one-dimensional convolution $\boldsymbol{F}^{\mathbf{l}} \in$ conv1D($\boldsymbol{F}^{\mathbf{d,i}}$), i∈{1,2,4} with dilation factors **d** of 1, 2, 4, a convolution kernel size of 3, and a step size of 1 is adopted to build local temporal context layer by layer. This method solves the problem of capturing the temporal continuity between video segments, ensuring that the temporal information of adjacent segments can be smoothly integrated by the model, thereby improving the local perception accuracy of the model for abnormal behavior.

In order to further improve the comprehensiveness of temporal features, this module also calculates the global temporal context through the paired temporal self-attention mechanism, further improving the modeling ability of global temporal features. In anomaly detection, relying only on local temporal information may lead to misjudgment, because abnormal events may show specific temporal patterns at the global level of the video. Specifically, through paired temporal self-attention calculation on the obtained local temporal features $\boldsymbol{F}^{\mathbf{l}}$, local temporal features $\boldsymbol{F}^{\mathbf{l1}}$, $\boldsymbol{F}^{\mathbf{l2}}$, and $\boldsymbol{F}^{\mathbf{l3}}$ of different scales are combined in the form of matrix multiplication ($\boldsymbol{F}^{\mathbf{g}}$ = conv1D(($\boldsymbol{F}^{\mathbf{l1}}$*($\boldsymbol{F}^{\mathbf{l2}}$)$^{\mathrm{T}}$) * $\boldsymbol{F}^{\mathbf{l3}}$)) to generate a fused global temporal feature representation, and finally the model stability and gradient flow are maintained through residual connection to obtain the final multi-scale temporal feature output $\boldsymbol{M}$. This mechanism processes features in pairs through temporal self-attention, solves the problem of representing abnormal events in global temporal information, enables the model to more accurately distinguish normal and abnormal segments in the temporal dimension, reduces the risk of missing abnormal segments, and thus greatly improves the model's anomaly detection accuracy and temporal consistency.

3.2.2. Time agent mechanism

We use the multi-scale information fusion in the video segments as the input of the time agent mechanism module, and generate enhanced temporal agent tokens (denoted as $\boldsymbol{A}$) through deformable convolution. These tokens are designed to capture the temporal dependencies between segments, thereby helping the model to better understand the temporal information across segments. Deformable convolution enables the model to adaptively focus on the key time points or regions of abnormal segments by flexibly adjusting the convolution receptive field, avoiding redundant processing of irrelevant information. This feature solves the problem that traditional fixed convolution cannot adapt to abnormal changes in the temporal dimension, especially for those abnormal events with irregular time. The generated $\boldsymbol{A}$ is input into the agent Transformer model for autocorrelation calculation, so that the model can effectively model long-term temporal dependencies under the agent mechanism. In the anomaly detection task, short-term features alone are often not enough to identify anomalies that appear to accumulate slowly or over a long period of time. Through the autocorrelation calculation in the agent mechanism, the model can enhance the understanding of the cross-temporal relationship between segments, thereby solving the problem that the features of abnormal events are not easy to capture over a long span of time. This autocorrelation process effectively improves the quality of the feature map obtained in the first stage, especially enriches the feature modeling of abnormal segments, enhances the discriminative representation of learning normal and abnormal segments, and adapts to the diverse temporal patterns of abnormal events. The proposed time agent mechanism is shown in Fig. 3.

Specifically, we use the multi-scale temporal features $\boldsymbol{M}$ obtained in the previous stage as the $\mathbf{M_Q}$, $\mathbf{M_K}$, and $\mathbf{M_V}$ input features of the time agent mechanism module, where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V} \in \boldsymbol{R}^{\mathrm{T \times C}}$ represent query, key, and value matrices, respectively. By modeling the temporal features of video segments through agent, the problem of missing context in the temporal information of abnormal events across segments is solved, which helps to build a complete temporal feature chain. Our time agent mechanism can be written as:

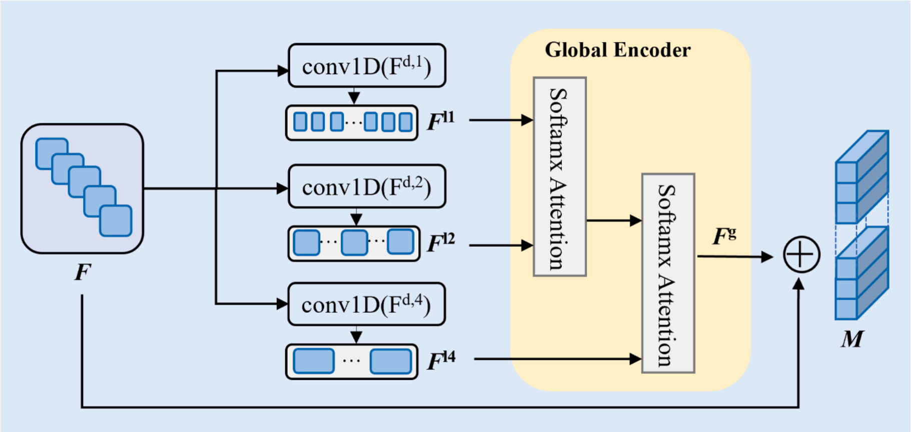


Fig. 2. The structure of the multi-scale temporal attention module, which captures temporal dependencies at different scales using a set of dilated convolution operations and the paired temporal self-attention mechanism.

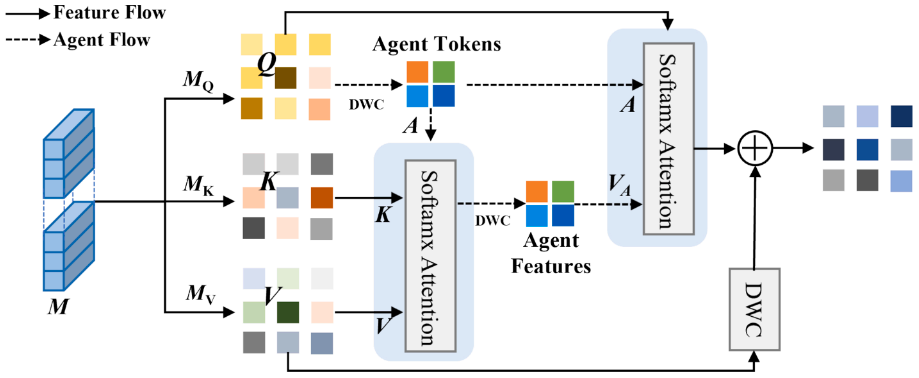


Fig. 3. The structure of the time agent mechanism module, which utilizes deformable convolution to generate temporal agent tokens and continuous temporal self-attention to model cross-segment temporal dependencies, thereby enhancing anomaly detection.

$$\boldsymbol{O}^{\mathrm{A}} = \underbrace{\mathrm{Attn}^{\mathbf{S}}\big(\boldsymbol{Q},\boldsymbol{A},\underbrace{\mathrm{Attn}^{\mathbf{S}}(\boldsymbol{A},\boldsymbol{K},\boldsymbol{V})\big)}_{\text{Agent Aggregation}}}_{\text{Agent Broadcast}}. \tag{1}$$

where $\mathrm{Attn}^{\mathbf{S}}$ represents self-attention and $\boldsymbol{A}$ is our custom temporal agent token.

As shown in Eq. (1), we initially treat the temporal agent token $\boldsymbol{A}$ as a query and perform self-attention calculations between $\boldsymbol{A}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ to aggregate the temporal agent feature $\boldsymbol{V_A}$ from all values. This step can integrate the temporal information of all segments, improve the model's ability to model the global temporal features of the video, and solve the problem of establishing long-term dependencies across segments in anomaly detection. Subsequently, we use $\boldsymbol{A}$ as the key and $\boldsymbol{V_A}$ as the value, and use the query matrix $\boldsymbol{Q}$ to perform a second self-attention calculation to broadcast the global information of the temporal agent feature to each query token, obtaining the final output $\boldsymbol{O}^{\mathrm{A}}$. This step enables the temporal agent features to be expanded synchronously on each segment, solving the problem that the model cannot balance short-term and long-term context features. This dual self-attention calculation enhances the model's ability to capture abnormal events across time and avoids misjudgments caused by lack of temporal consistency. Moreover, the stability of information transfer is maintained through residual connections and the temporal continuity of features is enhanced. The newly defined temporal agent token $\boldsymbol{A}$ essentially acts as an agent for $\boldsymbol{Q}$, aggregating global information from $\boldsymbol{K}$ and $\boldsymbol{V}$, and then broadcasting it back to $\boldsymbol{Q}$. This agent design effectively structures the flow of information across segments, allowing the model to better understand feature changes across long time spans and irregular anomalies.

In practice, a set of deformable convolutions that dynamically adjust the feature extraction receptive fields are used to generate agent markers $\boldsymbol{A}$, allowing the model to automatically focus on key areas of abnormal events and adapt to complex scene changes over a long period of time. This design solves the problem of insufficient local features caused by the fixed receptive field of traditional convolution. In the overall design of the temporal agent mechanism, the computational efficiency of the traditional visual Transformer is overcome, and the model's coverage ability of the global receptive field of the video is significantly improved. This mechanism significantly improves the anomaly detection effect of the model, enabling it to accurately distinguish between normal and abnormal segments in complex time series scenarios, ultimately improving the robustness and accuracy of detection.

3.2.3. Anomaly detection scores

To address the scenario where multiple abnormal segments exist within a video, we use the existing RTFM model for the final VAD stage. The RTFM model is theoretically driven by the MIL strategy of Top-k instances. For complex anomaly detection tasks, it makes full use of the temporal feature amplitude (i.e., L_2 norm) of video segments to distinguish abnormal and normal segments. This design enables the RTFM model to better adapt to the irregular distribution of abnormal segments in multiple abnormal scenarios, and effectively solves the problem of detecting multiple abnormal segments in complex scenarios. The model assumes that abnormal segments exhibit a higher average feature magnitude compared to normal segments, where normal segments are represented by low magnitude features and abnormal segments are represented by high magnitude features.

During model training, the RTFM model uses the k instances with the highest classification scores from abnormal and normal videos to train the classifier. Specifically, the Top-k strategy improves the model's sensitivity to local significant anomalies by aggregating the k instances with the highest feature amplitudes in abnormal videos, avoids noise interference in normal segments, and makes detection more accurate and efficient. At the same time, this Top-k MIL strategy allows the joint optimization of feature amplitude learning and MIL-based anomaly classification, so that the model can improve the discrimination on the boundary between normal and abnormal segments while learning the feature amplitude distribution, and enhance the model's recognition effect on boundary anomalies. The schematic diagram of the RTFM model is shown in Fig. 4. Through the feature amplitude drive and Top-k MIL strategy of the RTFM model, the model can adaptively distinguish abnormal and normal segments in multi- segment abnormal scenarios, effectively solving the problem of decreased detection accuracy due to segment overlap or blurred boundaries in multi-abnormal segment scenarios.

In the figure above, $\boldsymbol{x}^{+}$ denotes the abnormal segment feature magnitude and $\boldsymbol{x}^{-}$ denotes the normal segment feature magnitude, which are obtained from the abnormal ($\boldsymbol{x}^{+}$) and normal ($\boldsymbol{x}^{-}$) videos, respectively. The model achieves segment classifier learning by trying to maximize the separability between abnormal and normal video features, which is regarded as $\delta_{\mathrm{score}}(\boldsymbol{x}^{+}, \boldsymbol{x}^{-})$. This equation represents the difference between the mean of L_2 norm of Top-k features in abnormal and normal videos, where k is the number of abnormal segments in the abnormal video.

The RTFM model uses a loss function based on binary cross entropy to learn the segment classifier as shown in Eq. (2):

$$\mathrm{loss} = -\Big(y\log\big(f_{\phi}(\mathbf{x})\big) + (1-y)\log\big(1-f_{\phi}(\mathbf{x})\big)\Big) \tag{2}$$

where $\mathbf{x}$ is the Top-k feature segments with the largest L_2 norm, f_{ϕ} : $\mathbf{x}\rightarrow[0,1]^{T}$ is the snippet classifier, T is the number of video snippets (Tian et al., 2021). And y denotes the binary label actually classified as normal or abnormal, including 0 and 1 class labels, which denote normal and abnormal segments, respectively.

To effectively capture the subtle differences between video clips and maximize the distinction between normal and abnormal videos, we

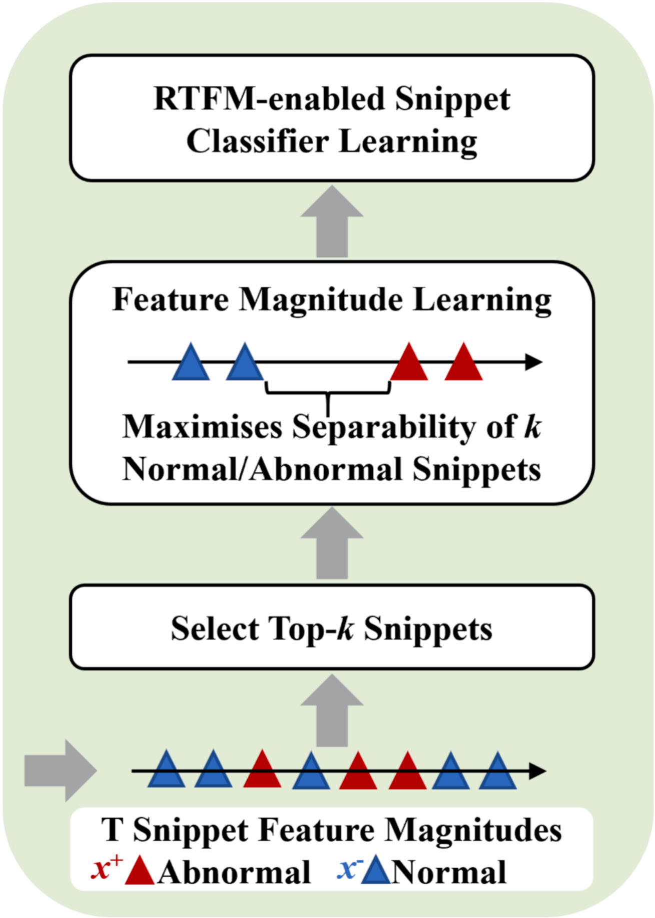

RTFM-enabled Snippet Classifier Learning
Feature Magnitude Learning
Maximises Separability of k Normal/Abnormal Snippets
Select Top-k Snippets
T Snippet Feature Magnitudes
x+ Abnormal
x- Normal

Fig. 4. The structure of the MIL-based RTFM anomaly scoring strategy, which maximizes the separability between abnormal and normal video features, and trains a snippet classifier using the Top-k largest magnitude feature snippets from abnormal and normal videos.

introduced a category contrast function as an optimization target for video detection. By bringing the feature representations of same-type samples (positive sample pairs) closer together, while simultaneously pushing apart the feature representations of different-type samples (negative sample pairs), we enhance the model's ability to distinguish abnormal samples. This strategy significantly improves both the stability and reliability of the model. The Eq. (3) for contrastive learning is as follows:

$$\mathscr{L}_{\text{contrastive}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp\left(\text{sim}\left(\mathbf{x}_i,\mathbf{x}_i^{+}\right)/\tau\right)}{\sum_{j=1}^{K}\exp\left(\text{sim}\left(\mathbf{x}_i,\mathbf{x}_j\right)/\tau\right)}\right] \tag{3}$$

Where $\mathbf{x}_i$ is the feature representation of segment i, which is linearly mapped to a low-dimensional space in the experiment. $\mathbf{x}_i^{+}$ is the feature representation of a positive sample from the same class as $\mathbf{x}_i$, and $\mathbf{x}_j$ is the feature representation of a negative sample (from a different class than $\mathbf{x}_i$). $\text{sim}\left(\mathbf{x}_i,\mathbf{x}_j\right)$ represents the similarity between the feature representations $\mathbf{x}_i$ and $\mathbf{x}_j$, typically computed using cosine similarity. τ is a temperature parameter that controls the sharpness of the similarity distribution, while N is the total number of segments and K is the number of negative segment.

The temporal smoothing between consecutive video segments is used to smoothly change the anomaly scores between video segments, enforcing similar anomaly scores between adjacent segments. The Eq. (4) is:

$$\mathscr{L}_{\text{Smoothness}} = \left(f_{\phi}(\mathbf{x}_i) - f_{\phi}(\mathbf{x}_{i-1})\right)^2, i \in (o, T) \tag{4}$$

Meanwhile, anomalies often occur within a short period of time, resulting in sparse anomaly scores for the segments in the anomaly bag, and a rare prior for abnormal events is imposed in each abnormal video, with the Eq. (5):

$$\mathscr{L}_{\text{Sparsity}} = \Sigma_{\text{t}=1}^{\text{T}}\left|f_{\phi}(\mathbf{x}_{\text{i}})\right| \tag{5}$$

The final loss function for VAD based on the RTFM model is:

$$\ell_f\left(f_{\phi}(\mathbf{x}), y\right) = \lambda 1^* \text{ Eq. } 2 + \lambda 2^* \text{ Eq. } 3 + \lambda 3^* \text{ Eq. } 4 + \lambda 4^* \text{ Eq. } 5 \tag{6}$$

where $\ell_f(.)$ is accompanied by the temporal smoothness and sparsity regularization, λ is the learning rate of each loss term. As defined in Deshpande et al. (2022), we set all λ values to 1 in the code.

4. Experiments

4.1. Experimental settings

4.1.1. Description of the dataset

UCSD-Peds is obtained through cameras fixed at higher positions on the campus, overlooking a sidewalk. The crowd density on the sidewalk varies, ranging from sparse to very crowded. In the normal setup, the video only contains pedestrians. Anomalous events are caused by the following two situations: non-sidewalk objects circulating on the sidewalk (e.g., skateboards, bicycles) and abnormal pedestrian movement patterns (e.g., running, crossing from the grass). Common anomalies include people riding bicycles, skateboarders, strollers, and people walking along the sidewalk or across the surrounding grass. There are also records of people in wheelchairs. All anomalies are naturally occurring, meaning they were not staged for dataset assembly. The data is divided into two subsets, each corresponding to different scenes. Video clips recorded from each scene are divided into various segments of approximately 200 frames.

Peds1: Clips of crowds moving toward and away from the camera, with some perspective distortion. It contains 34 training video samples and 36 testing video samples with a resolution of 238*158, totaling 12 anomalous events.

Peds2: Pedestrians moving parallel to the camera plane. It contains 16 training video samples and 12 testing video samples with a resolution of 360*240, totaling 40 anomalous events.

For each clip, the ground truth annotations include binary labels for each frame, indicating whether an anomaly is present in that frame. Additionally, a subset of 10 clips from Peds1 and 12 clips from Peds2 provides manually generated pixel-level binary masks identifying the regions containing anomalies. This allows for the evaluation of the algorithm's ability to locate anomalies.

CUHK Avenue is collected using a ground camera positioned parallel to the pedestrian movement path. It contains 16 training videos and 21 testing videos with a resolution of 640*360, totaling 47 anomalous events, with a total of 30,652 frames. Normal behavior is defined as pedestrians walking normally, while anomalous behaviors include running, throwing, loitering, and others. The size of people may change due to the camera's position and angle. The training videos consist of videos with normal behavior, while the testing videos include both standard and anomalous events. The dataset poses the following challenges: slight camera shake (appearing in Test Video 2, Frames 1051–1100), some outliers are present in the training data, and some normal patterns rarely appear in the training data.

The evaluation metric for the dataset (spatial location) is based not only on frame-level annotations but also considers the spatial location. The dataset uses rectangular bounding boxes to mark anomalous events and follows the VOC Pascal-style object detection criteria for anomaly detection.

ShanghaiTech is a large-scale crowd density counting dataset, released by ShanghaiTech University in 2016. It is captured by surveillance cameras fixed on the streets, containing data from different shooting angles and lighting conditions. The dataset includes anomalies caused by sudden movements, such as chases and fights that are not present in existing datasets, making it more suitable for real-world scenarios. The dataset contains 437 videos from 13 different backgrounds, with each video having a frame size of 856*480. It includes 330 training videos and 107 testing videos, totaling 130 anomalous events and over 270,000 training frames. Additionally, the dataset provides pixel-level ground truth annotations for anomalous events. Normal behavior is defined as pedestrians walking normally on campus, while anomalous behavior includes motor vehicles, fights, robberies, etc.

The ShanghaiTech dataset consists of Part A and Part B. Part A images are sourced from the internet, and the targets in these images are relatively dense. Part A contains 482 images, with 300 images in the training set and 182 images in the testing set, with an average resolution of 589*868. Part B images are taken from busy streets in Shanghai, and the targets in these images are relatively sparse. This part includes 716 images, with 400 images in the training set and 316 images in the testing set, with an average resolution of 768*1024. In total, there are 1,198 images with 330,165 annotated heads.

UCF-Crime is the largest publicly available video anomaly detection dataset to date, jointly released by the Center for Research in Computer Vision at the University of Central Florida (UCF) and the Information Technology University (ITU) in Pakistan. It encompasses a wide range of anomaly events captured by social surveillance systems across hundreds of diverse real-world scenes. The dataset consists of 1,900 untrimmed real-world surveillance videos, totaling approximately 13,769,300 frames with a cumulative duration of 128 h. The dataset is evenly divided into 950 anomalous and 950 normal videos. Regarding its partitioning, the training set includes 1,610 videos (800 normal and 810 anomalous videos) with weak labels indicating anomalies only at the video level, making it suitable for weakly supervised anomaly detection algorithms. The test set consists of 290 videos (150 normal and 140 anomalous) with frame-level ground truth annotations. Most videos are in 240*320 resolution, with an average length of 7,274 frames per video in MP4 format at a frame rate of 30 fps. The dataset features 13 types of real-life anomalies, including abuse, arrest, and arson, all of which are relevant to public safety. The dataset serves two purposes: general anomaly detection, where a set of anomalous activities is distinguished from normal activities, and anomaly classification, where specific types of anomalies are identified.

4.1.2. *Implementation details*

The experiments were performed on a Windows 10 system equipped with a high-performance NVIDIA GeForce RTX 3090 GPU, an Intel(R) Core (TM) i7-9700 CPU @ 3.00 GHz and 32G of RAM for model training. The programming language is Python 3.7.12 and the version of Pytorch used is 1.8.0, which is paired with CUDA version 11.1. A pre-trained VideoSwin Transformer model from the Kinetics dataset is applied to extract video segment features. Each video was divided into T = 32 snippets, each containing 16 frames. The video frames were sequentially cropped and resized to 224*224 using techniques like center cropping and horizontal flipping. The model was trained end-to-end using the Adam optimizer with a learning rate of 0.001 and no learning rate decay during training. The weight decay was set to 0.005, with a batch size of 64 and 500 training epochs. To ensure model convergence, hyperparameters were cross-validated and fine-tuned, and early stopping was used to prevent overfitting. During model training, each mini-batch consisted of 32 randomly selected samples from both normal and anomalous videos, resulting in a total of 64 samples per batch. In the loss function, we set all the λ values to 1 based on previous literature (Deshpande et al., 2022), which yielded the most effective results in the experiments.

4.1.3. *Evaluation indicators*

To comprehensively evaluate the performance of our proposed video anomaly detection method in the VAD task, we adopted the frame-level Receiver Operating Characteristic (ROC) curve and its Area Under the Curve (AUC) as the primary evaluation metrics. These metrics are widely used in anomaly detection tasks and effectively measure a model's classification performance and discriminative ability. In addition, we introduce two indicators: sensitivity and (1-specificity), also known as True Positive Rate (TPR) and False Positive Rate (FPR). These indicators allow the ROC and AUC metrics to remain unaffected by class imbalance. Since our experiments focus on positive samples, we are particularly concerned with how many negative samples are mistakenly predicted as positive. Hence, we use (1-specificity) instead of specificity.

The ROC curve is a tool for evaluating classification performance by plotting a model's performance at different thresholds. The curve is generated by traversing all thresholds, with TPR as the vertical axis and FPR as the horizontal axis. TPR measures the proportion of positive (anomalous) samples correctly predicted as positive, reflecting the model's coverage of predicted anomalies. FPR measures the proportion of negative (normal) samples incorrectly predicted as positive, reflecting the model's false alarm rate. The False Alarm Rate (FAR) measures the ability of the classifier or model to correctly predict positive samples while minimizing the misclassification of negative samples as positive. It represents the proportion of negative samples incorrectly predicted as positive out of the total negative samples. In theory, it is equivalent to the FPR. A lower FAR value indicates better model performance.

The TPR is calculated by the Eq. (7):

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

where *TP* refers to the number of positive samples correctly judged as positive, and *FN* refers to the number of positive samples incorrectly judged as negative.

The FPR and FAR are calculated by the Eq. (8) and Eq. (9):

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

$$FAR = FPR = \frac{FP}{FP + TN} \tag{9}$$

where *FP* refers to the number of negative samples incorrectly judged as positive, and *TN* refers to the number of negative samples correctly judged as negative.

TPR and FPR are calculated relative to the actual classes (positive and negative, respectively). Thus, they are unaffected by class imbalance.

AUC represents the area under the ROC curve, indicating the model's ability to separate positive and negative classes across different thresholds. To compute the AUC, we use an efficient ranking-based algorithm. The formula for AUC is calculated as shown in Eq. (10):

$$\mathrm{AUC} = \int_0^1 (TPR)\mathrm{d}(FPR) \tag{10}$$

This formula integrates TPR across various FPR values to obtain the AUC score. AUC is unaffected by class imbalance, making it an effective metric for evaluating a model's discriminative ability in imbalanced datasets. AUC values range from 0 to 1. The closer the value of AUC is to 1, the better the classification performance of the model, and vice versa. A value close to 0.5 indicates that the model is equivalent to random guessing and has no class separation ability.

4.2. *Comparison with state-of-the-art methods*

We compared our method with 28 state-of-the-art approaches, including 12 unsupervised learning methods, 12 weakly supervised learning methods, and 4 hybrid methods. The AUC results for these different methods are shown in Table 1. The results across various datasets demonstrate that our method outperforms all existing methods, achieving 99.98 % on UCSD-Ped2, 97.69 % on CUHK Avenue, 97.23 %

Table 1
Frame-Level AUC (%) Comparison of State-of-the-Art Methods ON UCSD-PED2, CUHK Avenue, Shanghaitech and UCF-Crime Dataset (best marked bold, second best underlined, and number ^{n} represents different references).

Supervision	Method	Feature/Time	Ped2	Avenue	ShanghaiTech	UCF-Crime
Hybrid	Morais et al.	2019	–	86.3	73.4	–
	Tang et al.	2020	96.3	85.1	73.0	–
	Liu et al.	2021	99.3	91.1	76.2	–
	Chang et al.	2022	96.7	87.1	73.7	–
Unsupervised	Park et al.	2020	97.8	88.5	70.5	–
	Yu et al.	2020	97.3	90.2	74.8	–
	Lv et al.	2021	96.9	89.5	73.8	–
	Doshi and Yilmaz^{1}	2021	97.2	86.4	70.9	–
	Yang et al.	2022	97.6	89.9	74.7	–
	Wang et al.	2022	99.0	92.2	89.8	–
	Reiss and Hoshen	2022	99.1	93.3	85.9	–
	Huang et al.	2023	95.5	87.63	76.57	–
	Doshi and Yilmaz^{2}	2023	–	85.8	71.2	–
	Tur et al.	2023	–	–	77.18	68.17
	Osman and Torki	2024	–	–	93.29	78.30
	Wu et al.	2024	97.2	90.6	75.5	–
Weakly Supervised	Sultani et al.	I3D/2018	92.3	–	85.33	75.41
	Zhong et al.	TSN/2019	93.2	–	84.44	–
	Zaheer et al.	C3D/2020	–	–	89.67	–
	Wan et al.	I3D + Flow/2020	–	–	91.24	–
	Tian et al.	I3D/2021	98.6	93.7	97.21	84.30
	Gong et al.	I3D/2022	–	–	94.92	–
	Chen et al.	VST-RGB/2023	–	–	–	86.67
	Zhou et al.	I3D/2023	–	–	–	87.24
	Zhang et al.^{1}	I3D/2023	–	–	–	86.22
	Karim et al.	Uniformer-32-RGB/2024	–	–	–	86.97
	Yun et al.	GNN/2024	–	–	–	84.48
	Zhang et al.^{2}	VST-RGB/2024	–	–	95.32	87.16
	Ours	I3D	98.43	95.21	94.36	85.52
	Ours	VideoSwin	**99.98**	**97.69**	**97.23**	**89.27**

on ShanghaiTech, and 89.27 % on UCF-Crime. These results validate the effectiveness of our approach.

As shown in the table, our method achieves the state-of-the-art performance across all three learning paradigms. Compared to hybrid methods, our approach achieves a 0.68 % improvement on Ped2, a 6.59 % lead on Avenue and a 21.03 % increase on ShanghaiTech. In comparison with unsupervised methods, our method shows a 0.88 % improvement on Ped2, a 4.39 % increase on Avenue, a 3.94 % boost on ShanghaiTech, and a 10.97 % gain on UCF-Crime, highlighting its advantages. Similarly, compared to weakly supervised methods, our approach exhibits notable improvements: a 1.38 % increase on Ped2, a 3.99 % gain on Avenue, a 0.02 % boost on ShanghaiTech, and a modest 2.03 % gain on UCF-Crime. These performance gains are commendable, demonstrating significant competitiveness. Additionally, using the VideoSwin Transformer for video feature extraction achieves better performance compared to the I3D network. Specifically, AUC scores improve by 1.55 %, 2.48 %, 2.87 %, and 3.75 % on Ped2, Avenue, ShanghaiTech, and UCF-Crime datasets, respectively. The VideoSwin Transformer demonstrates superior capability in understanding video content compared to the traditional I3D network. Notably, when combined with the same MIL anomaly scoring mechanism as enhanced RTFM, the VideoSwin Transformer achieves superior anomaly detection performance, surpassing the RTFM model. However, the RTFM model performs better with I3D features compared to MIL. Detection performance also varies across datasets. On Ped2, Avenue, ShanghaiTech, and UCF-Crime datasets, AUC scores increase by 0.68 %, 3.99 %, 0.02 %, and 2.03 %, respectively, with the most significant improvement observed on the Avenue and UCF-Crime dataset. These results indicate that our model achieves optimal performance in temporal feature fusion and demonstrates strong generalization capability.

4.3. Ablation study

In this section, we further validate the effectiveness of each design component. The experiments were conducted on the UCSD-Ped2, CUHK Avenue, ShanghaiTech, and UCF-Crime datasets to enhance the credibility of our model design.

Our network primarily includes three design modules: multi-scale attention, time agent mechanism, and enhanced RTFM. To assess their effectiveness, we conducted an ablation study focusing on feature aggregation using the VideoSwin Transformer, with the anomaly scoring strategy based on the MIL mechanism. Initially, we performed anomaly detection solely using the video feature network. Subsequently, we incorporated the multi-scale attention and time agent mechanism modules separately to evaluate their contributions. Finally, we assessed the performance of the complete design model. The experimental results are presented in Table 2.

We controlled for other design modules and evaluated the performance by incorporating only the multi-scale temporal attention module. This addition led to AUC improvements of 6.2 %, 7.76 %, 2.9 %, and 2.79 % on the Ped2, Avenue, ShanghaiTech, and UCF-Crime datasets, respectively. This indicates that the multi-scale temporal attention module enhances detection performance for anomalies of varying lengths by effectively capturing both long-term and short-term temporal relationships, making it suitable for VAD with irregular anomaly time. Next, integrating only the time agent mechanism module resulted in AUC improvements of 6.94 %, 8.08 %, 6.6 %, and 6.52 % on the Ped2, Avenue, ShanghaiTech, and UCF-Crime datasets, respectively. This highlights the importance of the time agent mechanism in capturing temporal correlations throughout the video, effectively distinguishing between abnormal and normal segments. Finally, combining both mechanisms further improved performance, with AUC increases of 8.39 %, 8.1 %, 8.51 %, and 8.91 % on the four datasets, respectively. This demonstrates the critical role of both modules in overall anomaly detection, significantly enhancing the integration of multi-scale temporal context information and improving anomaly detection capability.

To assess the importance of the RTFM anomaly scoring mechanism, we compared the original RTFM with the MIL scoring mechanism in the ablation study framework. The experimental results are shown in Table 3.

Table 2
The Contribution of Each Component on UCSD-Ped2, CUHK Avenue, ShanghaiTech and UCF-Crime DATASET (best marked bold).

VideoSwin	Multi-Scale	Time Agent	Ped2	Avenue	ShanghaiTech	UCF-Crime
√	–	–	91.51	83.14	86.12	77.35
√	√	–	97.71	90.9	89.02	80.14
√	–	√	98.45	91.22	92.72	83.87
√	√	√	**99.9**	**91.24**	**94.63**	**86.26**

Table 3
The Contribution of the RTFM Mechanism on UCSD-Ped2, CUHK Avenue, ShanghaiTech and UCF-Crime DATASET (best marked bold).

VideoSwin	Multi-Scale	Time Agent	RTFM	Ped2	Avenue	ShanghaiTech	UCF-Crime
√	–	–	√	92.61	85.16	87.16	78.42
√	√	–	√	97.84	92.16	91.60	81.13
√	–	√	√	98.68	94.75	93.77	84.02
√	√	√	√	**99.95**	**96.07**	**95.64**	**87.34**

In both scoring mechanisms, we incorporated temporal smoothing and sparsity constraint losses to enhance model training performance while ensuring consistency in the optimization approach. As shown in the Table 3, our model demonstrated varying levels of improvement across different combinations. Specifically, the overall model achieved AUC increases of 0.05 %, 4.83 %, 1.01 %, and 1.08 % on the Ped2, Avenue, ShanghaiTech, and UCF-Crime datasets, respectively, with the most notable improvement on the Avenue dataset. Across all datasets, the RTFM mechanism consistently outperformed the MIL mechanism, further validating the effectiveness of RTFM in leveraging the strength of abnormal features relative to normal features.

Finally, to evaluate the performance of category contrast loss in the RTFM model, we incorporated video segment feature amplitude contrast loss along with temporal smoothing and sparsity constraint loss. We then compared this optimized method with the baseline methods, one without the original RTFM (with MIL strategy) and one with the original RTFM. The experimental results, as shown in Table 4, clearly demonstrate that by adding contrast loss, the model's detection accuracy across various datasets is significantly improved, and the false alarm rate is reduced. This confirms the reliability and effectiveness of the improved RTFM approach.

From the table above, it is evident that our improved RTFM scoring method has delivered superior experimental results. Compared to the method without RTFM (which uses the MIL mechanism), our model with integrated contrast loss has achieved improvements of 0.08 %, 6.45 %, 2.60 %, and 3.01 % on the Ped2, Avenue, ShanghaiTech, and UCF-Crime datasets, respectively. Additionally, when the anomaly threshold is set to 0.8, the false detection rate is reduced by 0.0017 %, 0.0085 %, 0.0068 %, and 0.0102 % for the same datasets, significantly lowering the model's false detection rate. Furthermore, when compared to the method with the original RTFM, our model has achieved improvements of 0.03 %, 1.62 %, 1.59 %, and 1.93 % on the four datasets, with a reduction in the false detection rate by 0.0009 %, 0.0021 %, 0.0033 %, and 0.0047 %, respectively. These experiments demonstrate that our enhanced RTFM anomaly scoring strategy effectively improves the anomaly detection accuracy, reduces the false alarm rate, and significantly enhances the robustness and reliability of the model.

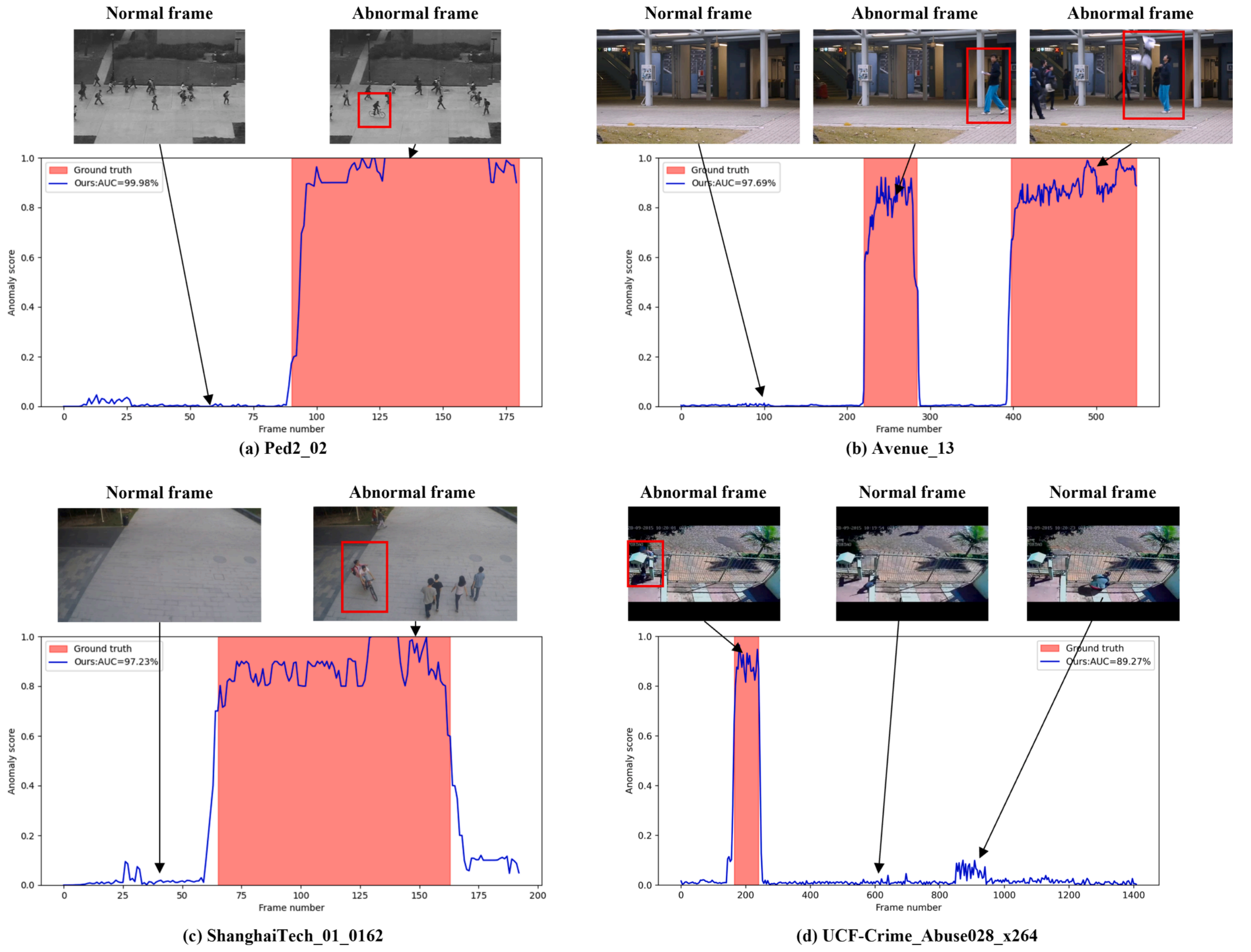
4.4. Qualitative analysis

In Fig. 5, we present several qualitative results of our video anomaly detector on the UCSD-Ped2, CUHK Avenue, ShanghaiTech, and UCF-Crime test datasets. The figure includes abnormal videos from all four datasets. As illustrated by the curves (blue line) in Fig. 5, our method effectively assigns low anomaly scores to normal segments and high anomaly scores to abnormal ones, ensuring a significant margin between normal and abnormal segments. Our model accurately detects abnormal events across the four datasets. Specifically, Fig. 5(a) shows an abnormal video from Ped2, Fig. 5(b) from Avenue, Fig. 5(c) from ShanghaiTech, and Fig. 5(d) from UCF-Crime. In Ped2, the model effectively distinguishes cycling from pedestrians, assigning a higher anomaly score to cycling. Avenue contains two abnormal segments—loitering and cycling—both of which are assigned high anomaly scores. In ShanghaiTech, cycling also results in a higher anomaly score. Also in UCF-Crime, the model successfully detects the act of abusing a puppy, assigning it a higher anomaly score compared to ambiguous behavior, such as holding a dog. Additionally, Fig. 5(b) demonstrates that our method can effectively detect multiple abnormal events within a single video, while Fig. 5(a) and 5(c) highlight that short-lived abnormal events are given higher anomaly scores. In summary, our approach consistently provides accurate anomaly scores for different types of anomalies, demonstrating its robustness and effectiveness.

We also provide t-SNE-based (Van der Maaten & Hinton, 2008; Zhou et al., 2024) visualizations of feature distributions to illustrate the feature separability on the UCF-Crime benchmark dataset. As shown in Fig. 6, normal segments from normal videos are represented as green dots, normal segments from anomalous videos as light red dots, and anomalous segments as red triangles. From the visualization, it can be observed that, before training, the normal and anomalous features exhibit a mixed distribution. However, after training, the clustering of normal and anomalous segments becomes significantly more distinct, with the distances between unrelated features effectively increased. This result demonstrates that our proposed temporal contextual feature mining network effectively distinguishes anomalous instances. The visualization further validates the effectiveness of our model in video anomaly detection tasks.

Table 4
Comparison of Improved RTFM Results (best marked bold).

Method	Ped2		Avenue		ShanghaiTech		UCF-Crime	
	AUC	FAR(%)	AUC	FAR(%)	AUC	FAR(%)	AUC	FAR(%)
Ours (no)	99.9	0.0021	91.24	0.0098	94.63	0.0082	86.26	0.0119
Ours (with)	99.95	0.0013	96.07	0.0034	95.64	0.0047	87.34	0.0064
Ours (improved)	**99.98**	**0.0004**	**97.69**	**0.0013**	**97.23**	**0.0014**	**89.27**	**0.0017**



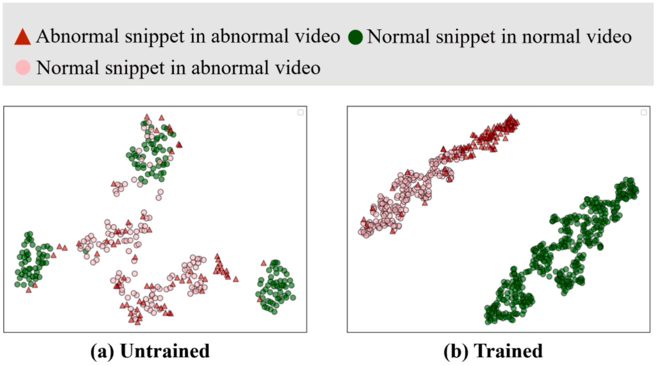
(a) Ped2_02

(b) Avenue_13

(c) ShanghaiTech_01_0162

(d) UCF-Crime_Abuse028_x264

Fig. 5. Qualitative results from the UCSD-Ped2, CUHK Avenue, ShanghaiTech and UCF-Crime test datasets. In subfigure (a), the abnormal event in Ped2 is identified as riding a bicycle. Subfigure (b) shows an abnormal video in Avenue, composed of multiple abnormal events, including wandering and cycling. Subfigure (c) illustrates an anomalous video from ShanghaiTech, where cycling is detected as an anomalous event. Subfigure (d) highlights the abuse of a dog as a clear anomaly in UCF-Crime. In subfigures (a) and (c), brief anomalous events are assigned higher anomaly scores. The blue curve represents frame-level anomaly score, with abnormal events highlighted by red boxes.



(a) Untrained

(b) Trained

Fig. 6. Results of t-SNE feature visualization on the UCF-Crime benchmark dataset. Before training, normal and anomalous segments exhibit a random distribution. However, after training, the clustering of normal and anomalous segments becomes more distinct, with the distance between unrelated features effectively increasing.

5. Conclusion

This study proposes a VAD method based on weakly supervised learning, which aims to efficiently detect abnormal events of different types and durations in videos, and achieves the best performance in this field. Firstly, the VideoSwin Transformer is used to extract spatial semantic features from video segments to capture local and global feature information of the video, which solves the problem of inaccurate video feature capture by traditional methods. Secondly, the contextual information at different time scales is captured by the multi-scale temporal attention module. Abnormal events often have non-fixed temporal features, so extracting context information at different time scales can help the model cover a wider range of temporal changes and ensure the complete detection and feature modeling of abnormal events. In addition, the time agent mechanism is used to enhance the model's representation ability of abnormal segments. This mechanism models temporal relationships at a global level, which not only improves the discriminative expression of abnormal segments, but also reveals the temporal dependencies between different segments, which helps to more clearly distinguish abnormal and normal segments in the temporal feature space, especially when dealing with complex abnormal scenes across segments. Finally, the improved RTFM model is introduced as an anomaly scoring mechanism to further improve the accuracy and robustness of anomaly detection. This mechanism enhances the model's ability to distinguish the boundaries between abnormal and normal segments by jointly optimizing feature amplitude learning and multi-instance anomaly classification, and improves the model's sensitivity to discrete abnormal segments in the video, so that the model still has strong detection performance in multi-segment anomaly scenarios. Experimental results demonstrate that the improved RTFM anomaly detection model, with its multi-scale temporal attention and time agent mechanism, significantly enhances model performance and facilitates the detection of anomalies in videos. In future work, we plan to further explore more advanced anomaly representation models and scoring strategies, and continue to optimize the performance of video anomaly detection to cope with more complex multi-class and multi-scale abnormal scenes. This method will provide new technical support for fields such as intelligent monitoring and automated video analysis, and promote the development and application of VAD in practical applications.

CRediT authorship contribution statement

Shili Zhao: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization. **Ran Zhao:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition. **Yan Meng:** Data curation, Validation, Writing – review & editing. **Xiaoyi Gu:** Data curation, Validation, Writing – review & editing. **Chen Shi:** Conceptualization, Validation, Supervision. **Daoliang Li:** Conceptualization, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This paper was supported by The Key Research and Development Plan of the Ministry of Science and Technology (NO.2022YFD2001700), The National Natural Science Foundation of China (NO.32273188) and Hainan Seed Industry Laboratory (NO.B23H10004).

Data availability

The authors do not have permission to share data.

References

Al-Lahham, A., Tastan, N., Zaheer, M. Z., & Nandakumar, K. (2024). A Coarse-to-Fine Pseudo-Labeling (C2FPL) Framework for Unsupervised Video Anomaly Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 6779–6788). https://doi.org/10.1109/WACV57701.2024.00665

Bai, Y., Wang, L., Tao, Z., Li, S., & Fu, Y., (2021). Correlative channel-aware fusion for multi-view time series classification. Proceedings of the. AAAI Conference On Artificial Intelligence 35, 6714-6722. https://doi.org/10.1609/aaai.v35i8.16830.

Chang, Y., Tu, Z., Xie, W., Luo, B., Zhang, S., Sui, H., & Yuan, J. (2022). Video anomaly detection with spatio-temporal dissociation. *Pattern Recognition, 122*, Article 108213. https://doi.org/10.1016/j.patcog.2021.108213

Chang, Y., Tu, Z., Xie, W., & Yuan, J. (2020). Clustering driven deep autoencoder for video anomaly detection. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020. *Proceedings, Part XV, 16*, 329–345. https://doi.org/10.1007/978-3-030-58555-6_20

Chen, H., Mei, X., Ma, Z., Wu, X., & Wei, Y. (2023). Spatial-temporal graph attention network for video anomaly detection. *Image and Vision Computing, 131*, Article 104629.

Chen, Y., Liu, Z., Zhang, B., Fok, W., Qi, X., & Wu, Y. (2023). Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 387–395).

Cong, Y., Yuan, J., & Liu, J. (2011). Sparse reconstruction cost for abnormal event detection. *CVPR, 2011*, 3449–3456. https://doi.org/10.1109/CVPR.2011.5995434

Deshpande, K., Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2022). Anomaly detection in surveillance videos using transformer based attention model. *International Conference on Neural Information Processing*, 199–211.

Doshi, K., & Yilmaz, Y. (2021). Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition, 114*, Article 107865. https://doi.org/10.1016/j.patcog.2021.107865

Doshi, K., & Yilmaz, Y. (2023). Towards Interpretable Video Anomaly Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2654–2663). https://doi.org/10.1109/WACV56688.2023.00268

Fan, Y., Yu, Y., Lu, W., & Han, Y. (2024). Weakly-supervised video anomaly detection with snippet anomalous attention. *IEEE Transactions on Circuits and Systems for Video Technology, 34*, 5480–5492. https://doi.org/10.1109/TCSVT.2024.3350084

Feng, J., Hong, F., & Zheng, W. (2021). Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14009–14018).

Georgescu, M., Barbalau, A., Ionescu, R. T., Khan, F. S., Popescu, M., & Shah, M. (2021). Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12742–12752).

Gong, Y., Wang, C., Dai, X., Yu, S., Xiang, L., & Wu, J. (2022). Multi-Scale Continuity-Aware Refinement Network for Weakly Supervised Video Anomaly Detection. In *2022 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). https://doi.org/10.1109/ICME52920.2022.9860012

Huang, C., Liu, C., Wen, J., Wu, L., Xu, Y., Jiang, Q., & Wang, Y. (2022). Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. *IEEE T Cybernetics, 54*, 3197–3210.

Huang, X., Zhao, C., Gao, C., Chen, L., & Wu, Z. (2023). Synthetic Pseudo Anomalies for Unsupervised Video Anomaly Detection: A Simple Yet Efficient Framework Based on Masked Autoencoder. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1–5).

Karim, H., Doshi, K., & Yilmaz, Y. (2024). Real-time weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 6848–6856).

Le, V., & Kim, Y. (2023). Attention-based residual autoencoder for video anomaly detection. *Applied Intelligence, 53*, 3240–3254.

Li, G., He, P., Li, H., & Zhang, F. (2023). Adversarial composite prediction of normal video dynamics for anomaly detection. *Computer Vision and Image Understanding, 232*, Article 103686. https://doi.org/10.1016/j.cviu.2023.103686

Li, N., Chang, F., & Liu, C. (2020). Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. *IEEE T Multimedia, 23*, 203–215. https://doi.org/10.1109/TMM.2020.2984093

Li, S., Liu, F., & Jiao, L. (2022). Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 1395–1403).

Liu, W., Luo, W., Lian, D., & Gao, S. (2018). Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6536–6545).

Liu, Z., Nie, Y., Long, C., Zhang, Q., & Li, G. (2021). A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 13588–13597).

Lv, H., Chen, C., Cui, Z., Xu, C., Li, Y., & Yang, J. (2021). Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15425–15434).

Morais, R., Le, V., Tran, T., Saha, B., Mansour, M., & Venkatesh, S. (2019). Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11996–12004).

Osman, N., & Torki, M. (2025). Temporal divide-and-conquer anomaly actions localization in semi-supervised videos with hierarchical transformer. *International Conference on Pattern Recognition*, 229–244.

Park, S., Kim, H., Kim, M., Kim, D., & Sohn, K. (2023). Normality Guided Multiple Instance Learning for Weakly Supervised Video Anomaly Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2664–2673). https://doi.org/10.1109/WACV56688.2023.00269

Pu, Y., & Wu, X. (2022). Locality-aware attention network with discriminative dynamics learning for weakly supervised anomaly detection. In *2022 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6).

Ramachandra, B., Jones, M., & Vatsavai, R. (2020). Learning a distance function with a Siamese network to localize anomalies in videos. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2598–2607).

Reiss, T., & Hoshen, Y. (2022). Attribute-based Representations for Accurate and Interpretable Video Anomaly Detection. ArXiv Preprint ArXiv:2212.00789. https://doi.org/10.48550/arxiv.2212.00789.

Sapkota, H., & Yu, Q. (2022). Bayesian nonparametric submodular video partition for robust anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3212–3221).

Sultani, W., Chen, C., & Shah, M. (2018). Real-World Anomaly Detection in Surveillance Videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6479–6488). https://doi.org/10.1109/CVPR.2018.00678

Sun, S., & Gong, X. (2023). Hierarchical semantic contrast for scene-aware video anomaly detection. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition* (pp. 22846–22856).

Sun, S., & Gong, X. (2024). Multi-scale bottleneck transformer for weakly supervised multimodal violence detection. *ArXiv Preprint*. ArXiv:2405.05130.

Tang, Y., Zhao, L., Zhang, S., Gong, C., Li, G., & Yang, J. (2020). Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters, 129*, 123–130. https://doi.org/10.1016/j.patrec.2019.11.024

Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J. W., & Carneiro, G. (2021). Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4975–4986).

Tian, Y., Pang, G., Liu, F., Liu, Y., Wang, C., Chen, Y., Verjans, J., & Carneiro, G. (2022). Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 88–98.

Tur, A. O., Dall'Asen, N., Beyan, C., & Ricci, E. (2023). In *Unsupervised Video Anomaly Detection with Diffusion Models Conditioned on Compact Motion Representations* (pp. 49–62). Ithaca: Cornell University Library, arXiv.org.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*.

Wan, B., Fang, Y., Xia, X., & Mei, J. (2020). Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE international conference on multimedia and expo (ICME)* (pp. 1–6). https://doi.org/10.1109/ICME46284.2020.9102722

Wang, G., Wang, Y., Qin, J., Zhang, D., Bao, X., & Huang, D. (2022). Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. *European Conference on Computer Vision*, 494–511.

Wang, S., & Miao, Z. (2010). Anomaly detection in crowd scene. In *IEEE 10th International Conference on Signal Processing Proceedings* (pp. 1220–1223).

Wang, Y., Qin, C., Bai, Y., Xu, Y., Ma, X., & Fu, Y. (2022). Making Reconstruction-based Method Great Again for Video Anomaly Detection. In *2022 IEEE International Conference on Data Mining (ICDM)* (pp. 1215–1220). https://doi.org/10.1109/ICDM54844.2022.00157

Wang, Y., Qin, C., Wei, R., Xu, Y., Bai, Y., & Fu, Y. (2022a). Self-supervision meets adversarial perturbation: A novel framework for anomaly detection. Proceedings of the 31st ACM International Conference on Information & Knowledge Management, New York, NY, USA, pp. 4555-4559. https://doi.org/10.1145/3511808.3557697.

Wang, Z., Zou, Y., & Zhang, Z. (2020). Cluster attention contrast for video anomaly detection. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2463–2471).

Wu, J., Zhang, W., Li, G., Wu, W., Tan, X., Li, Y., Ding, E., & Lin, L. (2021). *Weakly-Supervised Spatio-Temporal Anomaly Detection in Surveillance Video*. arXiv.org, Ithaca: Cornell University Library.

Wu, P., Wang, W., Chang, F., Liu, C., & Wang, B. (2023). Dss-net: Dynamic self-supervised network for video anomaly detection. *IEEE T Multimedia*. https://doi.org/10.1109/TMM.2023.3292596

Yang, Y., Fu, Z., & Naqvi, S. M. (2023). Abnormal event detection for video surveillance using an enhanced two-stream fusion method. *Neurocomputing (Amsterdam), 553*, Article 126561. https://doi.org/10.1016/j.neucom.2023.126561

Yang, Z., Liu, J., Wu, Z., Wu, P., & Liu, X. (2023). Video event restoration based on keyframes for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14592–14601).

Ye, M., Peng, X., Gan, W., Wu, W., & Qiao, Y. (2019). Anopcn: Video anomaly detection via deep predictive coding network. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 1805–1813). https://doi.org/10.1145/3343031.3350899

Yu, G., Wang, S., Cai, Z., Zhu, E., Xu, C., Yin, J., & Kloft, M. (2020). In *Cloze Test Helps: Effective Video Anomaly Detection via Learning to Complete Video Events* (pp. 583–591). Ithaca: Cornell University Library, arXiv.org.

Yu, S., Wang, C., Mao, Q., Li, Y., & Wu, J. (2021). Cross-epoch learning for weakly supervised anomaly detection in surveillance videos. *IEEE Signal Processing Letters, 28*, 2137–2141.

Yun, S., Masukawa, R., Na, M., & Imani, M. (2024). Missiongnn: Hierarchical multimodal gnn-based weakly supervised video anomaly recognition with mission-specific knowledge graph generation. *ArXiv Preprint*. ArXiv:2406.18815.

Zaheer, M. Z., Mahmood, A., Astrid, M., & Lee, S. (2020). Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020. *Proceedings, Part XXII, 16*, 358–376. https://doi.org/10.48550/arXiv.2011.12077

Zaheer, M. Z., Mahmood, A., Khan, M. H., Segu, M., Yu, F., & Lee, S. (2022). Generative Cooperative Learning for Unsupervised Video Anomaly Detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14724–14734). https://doi.org/10.1109/CVPR52688.2022.01433

Zhang, C., Li, G., Qi, Y., Ye, H., Qing, L., Yang, M., & Huang, Q. (2023). *Dynamic Erasing Network Based on Multi-Scale Temporal Features for Weakly Supervised Video Anomaly Detection*. Ithaca: Cornell University Library, arXiv.org.

Zhang, J., Qing, L., & Miao, J. (2019). Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *2019 IEEE International Conference on Image Processing (ICIP)* (pp. 4030–4034). https://doi.org/10.1109/ICIP.2019.8803657

Zhang, M., Wang, J., Qi, Q., Sun, H., Zhuang, Z., Ren, P., Ma, R., & Liao, J. (2024). Multi-Scale Video Anomaly Detection by Multi-Grained Spatio-Temporal Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 17385–17394).

Zhang, Y., Akdag, E., Bondarev, E., & De With, P. H. (2024). MTFL: multi-timescale feature learning for weakly-supervised anomaly detection in surveillance videos. *ArXiv Preprint*. ArXiv:2410.05900.

Zhao, M., Liu, Y., Liu, J., & Zeng, X. (2022). Exploiting spatial-temporal correlations for video anomaly detection. In *2022 26th International Conference on Pattern Recognition (ICPR)* (pp. 1727–1733).

Zhong, J., Li, N., Kong, W., Liu, S., Li, T. H., & Li, G. (2019). Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1237–1246).

Zhong, Y., Chen, X., Hu, Y., Tang, P., & Ren, F. (2022). Bidirectional spatio-temporal feature learning with multiscale evaluation for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology, 32*, 8285–8296.

Zhou, H., Yu, J., & Yang, W. (2023). Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 3769–3777).

Zhou, Y., Qu, Y., Xu, X., Shen, F., Song, J., & Shen, H. T. (2024). BatchNorm-based weakly supervised video anomaly detection. *Ieee T Circ Syst Vid, 1*. https://doi.org/10.1109/TCSVT.2024.3450734

Zhu, Y., & Newsam, S. (2019). *Motion-Aware Feature for Improved Video Anomaly Detection*. Ithaca: Cornell University Library, arXiv.org.