A Novel Video Anomaly Detection Method Based on Multi-scale Swin Transformer and Memory-augmented Attention Feature Fusion

Shuaina Huang [*,**,***], **Zhiyong Zhang** [*,***,****], **Bin Song** [*,***,****], **Yueheng Mao** [†]

[*] *Information Engineering College,Henan University of Science and Technology, Luoyang, 471023, Henan, China(e-mail: huangsn0202@126.com, xidianzzy@126.com, songbin@haust.edu.cn)*
[**] *Mechanical and Electrical Engineering College, Luoyang Polytechnic, Luoyang 471934, Henan, China*
[***] *Henan International Joint Laboratory of Cyberspace Security Applications,Henan University of Science and Technology, Luoyang, 471023, Henan, China*
[****] *Henan Intelligent Manufacturing Big Data Development Innovation Laboratory, Henan University of Science and Technology, Luoyang, 471023, Henan, China*
[†] *Sunnetech Ltd, Quzhou, 324003, ZheJiang, China*

Abstract: The purpose of Video Anomaly Detection is to automatically identify abnormal spatiotemporal patterns in surveillance systems. Traditional autoencoder frame reconstruction methods have made great progress in abnormal video detection. However, these methods often fail to adequately correlate global features with local emphasis feature and to harness feature information at various scales within the network. With the aim of addressing these issues, this study proposes a novel multi-scale Swin Transformer and memory-augmented attention feature fusion network for video anomaly detection(MSTMAF), which effectively performs multi-scale cross-fusion of global and fine-grained local features of the autoencoder. Specifically, a multi-scale Swin Transformer encoder is designed to achieve multi-layer feature fusion while extracting video frame features through a multi-branch skip connection operation. Furthermore, we design multi-attention fusion mechanism to refine attention, reducing the influence of background location in the feature map. This process heightens the high-level semantic representation of local features. Additionally, we introduce memory augmented module to better retain archetypal features of normal behavior from historical data. MSTMAF-Net achieved outstanding AUC performance on four publicly available datasets: UCSD Ped1(0.881), UCSD Ped2(0.982), CUHK Avenue (0.913), ShanghaiTech (0.764) and UCF-crime(0.831) demonstrating the effectiveness of our study.

Keywords: Anomaly Detection, Attention Mechanism ,Multi-Scale, Swin Transformer ,Feature Fusion, Autonomous Driving

1. INTRODUCTION

Video Anomaly Detection technology is a core technology in intelligent monitoring systems (Liu et al., 2023). Its objective is to automatically identify sensitive and anomalous events in video streams, such as traffic accidents, violent crimes, shootings, pornography, among others (Kumar et al., 2023)(Wang et al., 2019). Interest in VAD technology has recently increased due to its role in maintaining public security and safeguarding lives and property as a core function of intelligent video surveillance systems. VAD technology utilizes image processing and machine learning techniques to automatically identify a variety of deviant events caused by different targets within a surveillance video scene. This technology enables staff to promptly address abnormal events, reducing labor costs, improving monitoring efficiency, and minimizing false positives and negatives. However, the concept of anomaly is inherently unbounded and ambiguous. Due to their randomness and diversity in the real world, it becomes impractical to collect and model all types of anomalous data. Therefore, anomaly frame detection is frequently described as an unsupervised coarse-grained detection learn-

⋆ The work is supported by National Natural Science Foundation of China(No.61972133), Project of Leading Talents in Science and Technology Innovation in Henan Province(No.204200510021), Henan Provincial University-Enterprise Collaborative Innovation Project (No.26AXQXT029), The Henan Province Science and Technology Research Project (No. 242102210122, 252102110380), Henan Province International Science and Technology Cooperation Project(No.262102521051).
Corresponding author: Zhiyong Zhang, e-mail: xidianzzy@126.com

ing problem, focusing on learning a model that captures behavioral representations using only normal data. During inference, patterns deviating from established norms are flagged as anomalies. This method, known as unsupervised video anomaly detection (UVAD), eliminates the need for complex sample labeling and can automatically pinpoint anomalous events in video,thereby reducing inefficiencies in manual analysis and potential human error from the underutilization of existing surveillance infrastructure.

Most existing works employ reconstruction or prediction methods based on Auto Encoders (AE) for anomaly detection (Lee et al., 2022). These techniques usually rely on the autoencoder to learn the hidden space representation of normal events, with the assumption that abnormal occurrences will have a significantly different representation. The frame reconstruction method (Perera et al., 2019) uses the autoencoder's encoder to capture the features of normal events, while the decoder outputs reconstruction results to detect anomalies based on reconstruction error; the frame prediction method (Medel and Savakis, 2016) leverages previous frames to predict subsequent ones, detecting anomalies based on the discrepancy between predicted and actual frames. Although these approaches have been effective, the limited receptive field of AE's encoder (Wang et al., 2023b) hinders the comprehensive understanding of human and object behaviors within the global structure. The Swin Transformer algorithm (Liu et al., 2021), presented by Microsoft Research at ICCV in 2021, refines the Transformer architecture to address computational complexity. It employs a shifted window scheme that confines self-attention computation to non-overlapping local windows while still permitting for cross-window connectivity, capturing global context with reduced computational requirements. Jiang et al. (Jiang et al., 2022) employed the Swin Transformer network to the intelligent detection of industrial anomalies, effectively tackling the issue of limited anomaly samples and achieving excellent performance in anomaly detection and localization. Extensive research on abnormal video detection reveals that AEs struggle to capture distant and global spatial features due to their fixed receptive field size and often overlook critical local features. The effective integration of AEs and Swin Transformer for video anomaly detection remains uncharted territory. By harmonizing local and global dependencies, we can potentially enhance the learning of spatiotemporal characteristics in video frames.

Building on this foundation, this paper presents an innovative anomaly detection method. This method, referred to as the Multi-Scale Swin Transformer and Memory-augmented Attention Feature Fusion Video Anomaly Detection Network (MSTMAF-Net), significantly increases the accuracy of abnormal video detection. It features a unified architecture that spans computer vision and natural language processing, designing two cutting-edge technologies: the multi-scale Swin Transformer and Memory-augmented Attention Fusion. The intent is to effectively harness the global and local information from video data. To further enhance the model's capability to distinguish abnormal behaviors from normal ones, a memory enhancement module was designed. This module stores prototype features of normal behaviors extracted from consecutive frames, enabling the effective integration of multimodal information. By constructing multi-scale Swin Transformers and an attention-based feature fusion module, the MSTMAF model demonstrates outstanding performance in capturing both local and global representations, thereby significantly improving the recognition accuracy of abnormal behaviors. The primary contributions of our research can be presented as following:

1. This study presents an effective video anomaly detection modular architecture MSTMAF-Net. This modular architecture features an innovative multi-scale Swin Transformer encoder-decoder architecture. The multi-scale Swin Transformer extracts global motion features of video frames and accomplishes multi-scale feature cross-fusion, improving the encoder's feature characterization ability.

2. To mitigate the impact of background position on the feature map, a sophisticated attention fusion mechanism has been designed. This mechanism effectively captures the subtle attention semantics of local features, thereby yielding more precise local attention features.

3.We developed a memory-augmented module that is embedded within the AE framework to encode the normal behavioral space into prototype representations in real time. This module functions as a discriminative filter based on the stored normal prototypes, and its integration into the overall architecture is crucial for enhancing the model's capability to detect anomalies.

The remaining of this study is structured as follows: Section 2 is the review of the associated works in anomaly detection and Transformer-based approaches. Section 3 elaborates on MSTMAF-Net, including problem formulation, network architecture, and loss functions. The experimental data and comments are presented in Section 4. Section 5 is the conclusion of the study, outlines the limitations of our approach and directions for the future studies.

2. RELATED WORK

Video anomaly detection is traditionally defined as an unsupervised learning issue (Leyva et al., 2017) or a sparse representation problem (Sabokrou et al., 2018) because to the scarce full-size anomalous data and high annotation costs. Unsupervised learning algorithms use training datasets that only include data representing normal behavior. By learning the distribution model of normal behavior, the system identifies input data that deviates from the normal distribution as abnormal. Sparse representation methods reconstruct test samples from a dictionary constructed from normal samples and label those exceeding a preset reconstruction error threshold as anomalies. Our work primarily focuses on unsupervised learning due to its broader applicability in practical scenarios. Traditional feature extraction methods have focused on low-level visual features, including histogram of gradients, compact feature, and foreground masks (Saligrama and Chen, 2012), among others. These methods heavily depend on handcrafted features for video representation, which do not support end-to-end processing and have limited generalization capabilities.

With the advancement of deep autoencoders, various AE variants have been developed to model spatiotemporal information and reconstructing video frames. In order to

differentiate between normal and abnormal occurrences, wang et al. (Ravanbakhsh et al., 2017) created the Memory Augmented Appearance Motion Network (MAAM-Net) utilizing an autoencoder architecture. They encouraged memory modules to choose a sparse selection of essential memory items by using margin-based potential loss. Although the introduction of memory modules mitigates the overfitting problem of autoencoders, this model still lacks an effective mechanism for hierarchical multi-scale feature fusion and fails to explicitly enhance salient local features. Consequently, while it is capable of memorizing regular patterns, its feature representation capacity remains limited. The absence of multi-scale processing results in the loss of fine-grained details that are crucial for detecting subtle anomalies, whereas the lack of a dedicated local attention mechanism renders the model susceptible to interference from complex or dynamically changing backgrounds. Barbalau et al. (Barbalau et al., 2023) employed alternating 2D and 3D Convolutional Vision Transformer(CVT) blocks within a self-supervised multi-task learning(SSMTL) framework. This approach, using pseudo-anomalies for adversarial learning, pseudo-label generation techniques, and attention-enhancement modules, notably enhances the spatial anomaly localization capability of the C3D anomaly video recognition network. Thakare et al. (Thakare et al., 2023) introduced a multi autoencoder based method for anomaly detection and classification, combining latent features with pseudo-labeled I3D features of video clip. In order to optimize the multi-instance learning process, Shao et al. (Shao et al., 2023) proposed a time-convolutional network multi-instance learning model that balances feature connections between uncommon positive and negative samples. Although these methods address video anomaly recognition to some extent, they suffer from "over-generalization" due to insufficient representation of global and local features. As a result, abnormal frames can sometimes be reconstructed or predicted as reliably as normal frames.

The Swin Transformer, derived of the ViT model, introduces a hierarchical structure to mitigate the shortcomings of CNNs, which typically struggle to facilitate interaction between global and remote semantic information. Recent studies have explored the application potential of Transformer networks in classification and detection, particularly in anomaly detection. In order to identify and amplify local pattern features in video sequences, wu et al. (Wu et al., 2023) developed an unsupervised learning strategy based on vision transformation, which greatly increased the effectiveness of anomaly detection tasks. Li et al. (Li et al., 2022) used a Transformer to amalgamate the intricate relationships between different tags, with the combination of a memory module boosting the extraction of spatiotemporal features from abnormal videos. It encodes patches using Transformer operations and reduces local redundancy through a self-attention mechanism. The Swin Transformer uses a layered architecture and a shifted window strategy for feature representation computation, enabling rapid and efficient processing of visual data. This method is suitable for modeling global features and simplifying computational complexity. However, current approaches have not combined Swin Transformer with video anomaly detection to take advantage of its benefits. While Transformers excel at capturing features of abnormal videos, they struggle with local context and multi-scale spatial information.

3. PROPOSED METHODOLOGY

This section outlines the structure of the MSTMAF-Net. Figure 1 presents the overall architecture of the MSTMAF-Net framework for video anomaly detection. The proposed framework comprises three modules: multi-scale Swin Transformer, Attention Feature Fusion, and Memory-augmented module. In this study, we design a multi-scale Swin Transformer and Attention Feature Fusion, taking a sequence of video frames as input and producing multi-scale features as output. The multi-scale Swin Transformer uses multi-branch skip connections to promote the fusion of multi-layer features at various scales while concurrently extracting temporal motion features from normal events, enhancing the encoder's feature representation capability. The Attention Feature Fusion module focuses on local representation of normal motion, reduces the influence of background elements in the feature map, and optimizes the representation of normal targets. In addition, the decoder amalgamates features from the encoder and processes them through the innovative memory-augmented module, which archives prototype features of typical behavior from historical data. This storage enables the coexistence of multi-modal data and informs the prediction of future frames. The following subsections elaborate on the specifics of these modules.

3.1 Multi-scale Swin Transformer Encoder Module (MST)

Conventional AE architecture lacks global context modeling capabilities. In order to improve the encoder's feature characterisation capabilities, this study presents the Multi-scale Swin Transformer encoder module. This encoder module uses multi-scale operations for multi-layer feature fusion. This is shown in Figure 1. This study uses the feature fusion module to extract and merge multi-scale behavioral features, accepting the output of Patch merge at different stages as input. The feature fusion module utilizes straightforward and effective up-sampling, down-sampling, and concatenation techniques to refine the final feature extraction precision. During the process of training, the $H * W * 3$ RGB images fed into the model are partitioned into small blocks and sequentially pass through four stages. In this process, each stage generates feature maps of varying scales. The feature fusion module extracts and combines multi-scale features, progressively merging the output of Patch merge during the last three stages as the input. The module uses simple yet potent up-sampling, down-sampling, and concatenation methods bolsters the ultimate recognition accuracy.

The multi-scale Swin Transformer module employs a hierarchical down-sampling technique similar to convolutional structures for feature extraction. For the purpose of computing self-attention within each window, it divides feature maps into numerous uniform, non-overlapping windows. The window self-attention calculation is conducted in parallel. The following is an expression of the computation process:

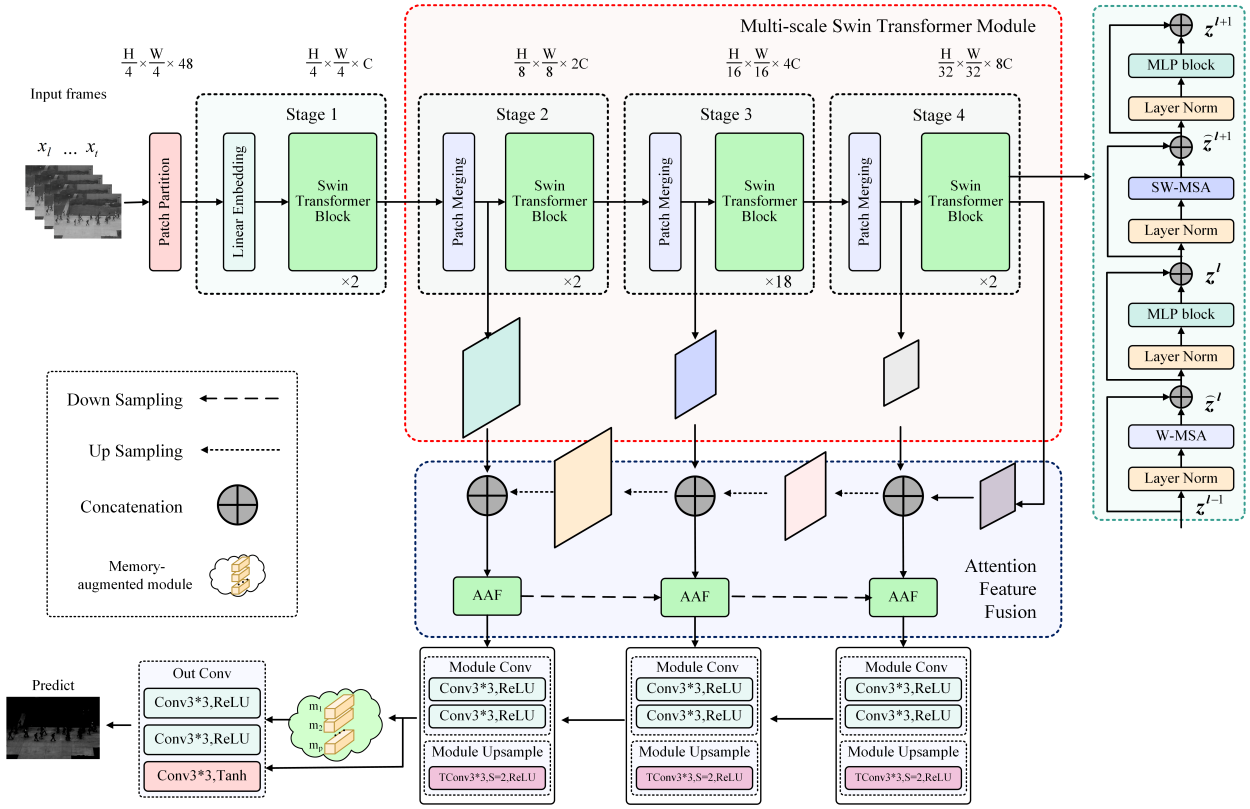

Fig. 1. Schematic Diagram of The Overall Network Framework of The MSTMAF-Net Model.

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{Q\cdot K^T}{\sqrt{d_k}} + B\right)V \quad (1)$$

Where Q indicates the query vector, K suggests the key vector, and V indicates the value vector, which is used to calculate the similarity between features. d_k indicates the dimension of the query or key. K^T represents the transpose of the K vector. B is the relative position encoding matrix.

Furthermore, multi-scale Swin-Transformer adopts the shift window multi-head self-attention module in order to improve the information interaction between self-attention Windows. The shift window attention mechanism moves the window to associate multiple adjacent windows, enabling information fusion and global modeling. The computational process for this mechanism is as follows:

$$\begin{aligned}
\hat{z}^l &= \text{W}-\text{MSA}(\text{LN}(\text{z}^{\text{l}-1})) + \text{z}^{\text{l}-1} \\
\hat{z}^l &= \text{MLP}(\text{LN}(\hat{z}^l)) + \hat{z}^l \\
\hat{z}^{l+1} &= \text{SW}-\text{MSA}(\text{LN}(\text{z}^{\text{l}})) + \text{z}^{\text{l}} \\
z^{l+1} &= \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1}
\end{aligned} \quad (2)$$

Where LN indicates layer normalization and MLP indicates multi-layer perception. The MSA module is based on shift windows followed by a two-layer MLP having a GELU nonlinear activation function. Each MSA module and every MLP module is applied before the regularization layer. Each module is followed by the residual join in the interim.

3.2 Attention Feature Fusion (AFF)

Although the multi-scale Swin Transformer excels at obtaining enhanced global features, it often overlook the finer local details. These details are crucial for anomaly detection. For instance, areas such as road surfaces, shopping malls, and pedestrian zones are typically more important than trees and houses. Recognizing that the traditional AE models failed to address the spatial information of regions of interest, we designed an Attention Feature Fusion (AFF) mechanism, illustrated in Figure 2. The attention feature module accumulates context information across all axes, enriching the pixel-wise representation of the context. This approach enables selective concentration on pivotal features within the data, resulting in enhanced recognition capabilities. The introduction of the AFF represents an obvious advancement in the domain of anomalous video recognition, demonstrating exceptional proficiency in learning context-aware feature sequences and focusing on local key features.

The feature space of an image x_i is processed by several layers of the multi-scale Swin Transformer in the encoder, and can be classified into several regions indicated by $X \in R^{3C\times H\times W}$, in which W and H are the number of rows and columns of the feature matrix, respectively, and $3C$ indicates the number of feature region channels. Divide X into 3 groups along the channel dimension to get $X = [X_1, X_2, X_3]$, where $X_i \in R^{C\times H\times W}$. The input feature map X_1 is converted to a one-dimensional feature vector after global average pooling, which makes the global perceptual information available to each channel. Next, the dimension of the one-dimensional feature vector is reduced

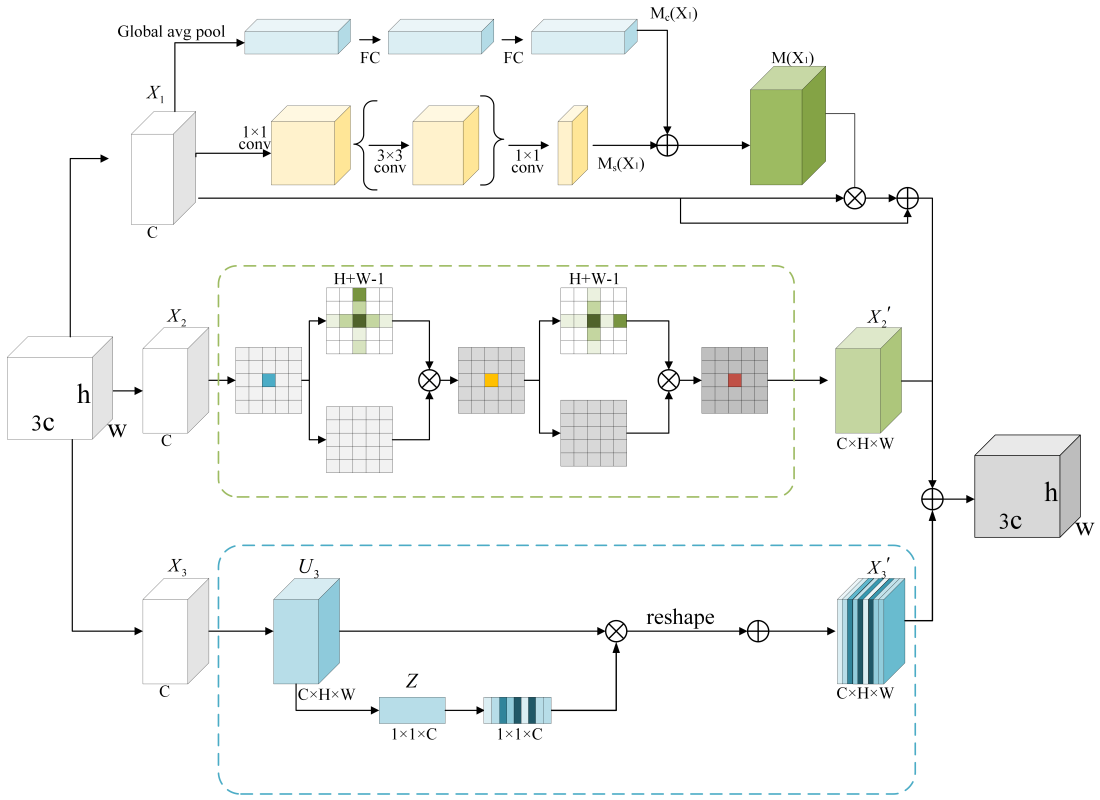

Fig. 2. Schematic Diagram of Attention Feature Fusion Model (AFF).

through the fully connected layer and nonlinearized using the ReLU activation function. Finally, the dimension is then increased through the fully connected layer and batch normalization is applied to obtain the corresponding weights $M_c(X_1)$. Thus, the computational formula is as follows:

$$M_c\left(X_1\right) = BN(W_1(W_0 AvgPool(X_1))) \tag{3}$$

the $1*1$ convolution is adopted for reducing the dimension of the input feature map X_1. Afterwards, the feature information is extracted using two dilated convolution with convolution kernel size $3*3$, which has a larger receptive field. Finally, $1*1$ convolution is applied to map the feature map to $1*w*h$, and the spatial attention map $M_s(X_1)$ is obtained. The calculation formula is as follows:

$$M_s\left(X_1\right) = BN(f_3^{1*1}(f_2^{3*3}(f_1^{3*3}(f_0^{1*1}(X_1))) \tag{4}$$

The $M_c(X_1)$ and $M_s(X_1)$ are extended to the same dimension according to the broadcast mechanism, and then the weights are summed up to finally obtain the attention vector $M(X_1)$. The input feature map X_1 is multiplied by $M(X_1)$ element by element, and then the residual structure is added to X_1, to get the feature graph X_1'.

The feature map X_2 first undergoes two $1*1$ convolution to generate Q and K, $\{Q, K\} \in R^{C' \times W \times H}$, where the resulting channel dimension C′ is smaller than C. For each feature point u within the spatial dimension in $Q^{C' \times W \times H}$, a vector $Q_u \in R^{C'}$ can be derived. In this region, features are collected from the same horizontal and vertical directions corresponding to the position of the feature point u in $K^{C \times W \times H}$, resulting in a set $\Omega_u \in R^{(H+W-1) \times C'}$, where $\Omega_{t,u} \in R^{C'}$ refers to the i-th element of Ω_u. The affinity operation is expressed as:

$$\mathrm{d_{i,u}} = \mathrm{Q_u} \times \Omega_{\mathrm{i,u}} \tag{5}$$

Here, $d_{i,u} \in D^{(H+W-1) \times (W \times H)}$ represents the degree of correlation between Ω_u and $\Omega_{i,u}$, where $i \in \{1, ..., H+W-1\}$, After performing a softmax operation over the channel dimension, the attention feature map A is obtained. Then the feature map H is then convolved with $1*1$ to produce feature graph $V \in R^{C \times W \times H}$. At the spatial position of each feature point u within V, a vector $V_u \in R^C$ is obtained, and features can also be collected from the same horizontal and vertical directions corresponding to the position u to form a set $\Phi_u \in R^{(H+W-1) \times C}$. An aggregation process is used to collect context information:

$$\mathrm{X_2'} = \sum_{\mathrm{i} \in |\Phi_\mathrm{u}|} \mathrm{A_{i,u}} \Phi_{\mathrm{i,u}} + \mathrm{X_2} \tag{6}$$

The feature map X_3 is arranged based on the channel sequence as $U_3 \in R^{D \times C}$, where $D = W \cdot H$ and C represents the number of characteristic channels. In our proposed model, the convolution compresses the global spatial information to generate channel statistics, where the global spatial features of each channel are represented as:

$$z_c = F_{sq}\left(u_c\right) = \frac{1}{H*W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{7}$$

Here, u_c indicates a channel in feature map u_3, z_c represent the compressed vector Z. From the formula, each feature map channel of $H \times W$ has only one pixel left after global averaging. Thus, an $H \times W \times C$ feature map is squeezed into an $1 \times 1 \times C$ vector.

The outcomes of the aforementioned operations are used to determine statistical characteristics of each channel. Then, find out the relationship between each channel. To get channel features with adaptive relevance, these dependencies are fused with feature mappings of other feature channels. The process is as follows:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma\left(W_2 \delta\left(W_1 z\right)\right) \tag{8}$$

where σ represent ReLU activation function, and $W_1 \in R^{\frac{C}{r} * C}$, $W_2 \in R^{C * \frac{C}{r}}$ represents the weighted matrix. Specifically, a fully connected layer is used to transform the $1 \times 1 \times C$ vector Z into a $1 \times 1 \times \frac{C}{r}$ vector, followed by a ReLU activation layer. Then, a second fully connected layer transforms the $1 \times 1 \times \frac{C}{r}$ vector back into a $1 \times 1 \times C$ vector, followed by a Sigmoid activation layer.

Finally, the obtained weights are merged with the original feature map u_3, and the feature map is recalibrated. Each element in s is taken out one by one and multiplied with the corresponding channel of the feature map. This operation is essentially a scalar multiplied by a matrix.

$$X_3{}' = F_{\text{scale}}\ (u_c, s_c) = s_c u_c \tag{9}$$

To obtain high-level semantic information of local regions, we designed an attention feature fusion mechanism that structurally integrates the spacial and channel attention mechanisms. First, the tensor is split into three groups, each processed by the Attention Unit. Then, the information within each group is fused by concatenation. Finally, the Channel Shuffle and convolution operation rearrange the groups, allowing for information exchange between different groups. The AFF structure is shown in Figure 2. Moreover, AFF is embedded in the skip connection layers of the multi-scale Swin Transformer to enhance feature expression capability. This fusion method efficiently utilizes local and global features at various degrees, allowing them to work in concert and improve the model's anomaly detection capabilities.

3.3 Memory-augmented module

With the generalization ability of the prediction network being too strong, the model has a strong prediction ability for the abnormal samples that have not been seen before. As a result, the prediction network can not only predict normal samples well, but also accurately predict abnormal samples, which makes the model unable to effectively distinguish abnormal samples from normal samples. To solve this problem, We design memory-augmented modules. The memory-augmented modules first process the input features through the criss-cross and channel attention unit. They then compresses the normal features in real time through various methods, constructing a dynamic memory pattern pool. Normal pattern features are retrieved and reconstructed from this pool. Then, to produce the output of the memory unit, the input feature and the normal pattern feature are concatenated.

The encoder feature $X_t \in R^{h \times u \times c}$ passing through the double attention mechanism is divided along the feature channel dimension into $h \times w$ query items $q_t^k \in R^{\mathrm{l} \times \mathrm{l} \times \mathrm{c}}(k = 1, 2, ..., K; K = h \times w)$. Each items represents a specific spatial position. For each query item x_t^k, P attention mapping functions are applied to assign normal weights, resulting in P attention maps. These maps indicate the weights of the feature vectors for each spatial relationship and generate associated similarity and pattern vectors. The process is described mathematically as follows:

$$\begin{aligned} &\omega_t^{k,p} \in W_t^p = \psi_p(x_t); \left\{\psi_p : R^{h \times w \times c} \rightarrow R^{h \times w \times 1}\right\} \\ &m_t^p = \sum_{k=1}^{K} \frac{\omega_t^{k,p}}{\sum_{k'=1}^{K} \omega_t^{k',p}} x_t^k \end{aligned} \tag{10}$$

Therefore, a set of N coding vectors can be obtained. The output vector X_t^k, after passing through the decoder and double attention mechanism, is treated as the query value. The pattern vector $M_t = \{m_t^p\}_{p=1}^{P}$ in the memory-augmented unit is used as the key value to reconstruct the normal memory encoding $\hat{X}_t \in R^{h \times w \times c}$. The calculation process is as follows:

$$\begin{aligned} \widehat{x}_t^k &= \sum_{p=1}^{P} \beta_t^{k,p} m_t^p \\ \beta_t^{k,p} &= \frac{x_t^k m_t^p}{\sum_{p=1}^{P} x_t^k m_t^p} \end{aligned} \tag{11}$$

When a matching pattern $\hat{x}_t^k$ is found for each query X_t^k, all K pattern vectors $\hat{x}_t^k$ form a feature tensor $\hat{X}_t \in R^{h \times w \times c}$ of the same size as X_t^k. This tensor is concatenated alongside the feature channel dimension to acquire the feature tensor $f_t = \left[X_t; \hat{X}_t\right] \in R^{h \times w \times 2c}$ as the output of the decoded network, producing the prediction result.

3.4 Loss function

In training process, the prediction loss function is carefully crafted considering the differences and associations between the Multi-scale Swin Transformer, Attention Feature Fusion and memory-augmented module. A feature reconstruction loss is specifically designed for the attention feature fusion and dynamic memory module.

The prediction loss function L_{pre} ensures the network accurately predicts the current frame F_t in accordance with the sequence of previous video frames. It is formally referred to as the L_2 distance between the predicted result $\hat{F}_t$ and the actual video frame F_t:

$$L_{pre} = \sum_{t=l+1}^{T} \left\| F_t - \hat{F}_t \right\|_2 \tag{12}$$

The feature reconstruction loss comprises a compact term loss L_{com} and a feature separation loss L_{sep}:

$$\begin{aligned} L_{com} &= \sum_{t=l+1}^{T} \sum_{k=1}^{K} \left\| x_t^k - m_t^* \right\| \\ L_{sep} &= \frac{2}{P(P-1)} \sum_{p=1}^{P} \sum_{p'=1}^{P} \left[\left\| -(m_t^p - m_t^{p'}) \right\| \right. \\ &\quad \left. + cov \left\| m_t^p - m_t^{p'} \right\| \right] \end{aligned} \tag{13}$$

Where m_t^* is the memory feature with the greatest similarity to the query value x_t^k in the memory module,

$* = \arg\max_{p \in P} w_t^{k,p}$. The definition of m_t^p, $\omega_t^{k,p}$ has been provided in section 3.3 and Equation 10. $cov\|\cdot\|$ is the covariance of the corresponding vector. The compactness loss and separation loss promote diversity between prototype projects by pushing learned prototypes away from each other. Covariance can be used to push all learned prototypes to distant places, so that the learned normal prototypes are maximum distance and independent. The overall loss function is defined as:

$$L_{all} = L_{pre} + \lambda_1 L_{com} + \lambda_2 L_{sep} \tag{14}$$

4. EXPERIMENTS

The current section presents extensive experiments on the proposed MSTMAF algorithm using four widely recognized benchmark datasets. We conducted quantitative and qualitative comparisons with existing Video Anomaly Detection (VAD) methods to demonstrate the superiority of the proposed model.

4.1 Datasets

UCSD Ped1 and UCSD Ped2 are the two sub-datasets that make up the UCSD Ped datasets. 34 training and 36 testing movies are included in Ped1. Twelve testing movies and sixteen training videos make up Ped2. A sizable dataset of 37 movies for a single scenario is known as the CUHK Avenue dataset. It features 14 types of abnormal events, containing abnormal running and throwing foreign objects. The datasets include totally of 30,652 frames. The ShanghaiTech datasets addresse the lack of diversity in scenes and perspectives found in existing datasets. This dataset comprises 437 surveillance videos, including 130 abnormal videos across 13 scenarios under complex lighting conditions. The UCF-crime refers to the largest dataset for abnormal event detection. The dataset includes traffic accidents, street robberies, explosions, fights, shootings, normal events, and so on. There were 1900 videos in all, clocking in at 128 hours with an average frame rate of 7274 frames per second. Furthermore, the majority of the video size is 240x320.

4.2 Implementation details

The experiments were carried out with four NVIDIA RTX A5000 GPUs and implemented using the Pytorch development framework. Video frames were resized to $256 * 256$ pixels, and the input sample consisted of sequences of video frames. The Adaptive Moment Estimation (Adam)optimizer with a momentum component was used to optimize the model. Besides, the initial learning rate was defined to 0.0001. The feature dimension in the memory unit was set to 512 with a total quantity of 10 prototype vectors . Training epochs were set to 1000, with models saved every 30 epochs.

4.3 Comparisons with existing methods

As presented in Table 1, Our MSTMAF-Net achieves 98.2%, 91.3%, and 76.4% frame level AUCs on the UCSD Ped2, CUHK Avenue and ShanghaiTech datasets, respectively. These results outperform both the deep learning-based and the manual feature-based baselines. The success of the MSTM-AF-Net model can be attributed to the multi-scale Swin Transformer and memory augmented attention feature fusion mechanism, which process encoder features in both global and local contexts, respectively. The memory-augmented module can store a rich array of normal event prototypes, leading to significant improvements in video anomaly detection performance. To evaluate the robustness and stability of the proposed MSTMAF-Net, we conducted five independent training trials on the UCSD Ped2 datasets. The model achieved a mean frame-level AUC of 98.2% with a standard deviation of 0.3%. This remarkably low deviation indicates the high reliability and training stability of our method.

Relative to IPR, the AUC of our proposed method on the three data sets is 2%-5% higher. Although the network structure used in IPR is based on encoder-decoder structure, it lacks the processing of multi-scale and global-local features. When evaluating the larger Ped1 datasets, as depicted in Figure 4, the AUC of the MSTMAF-Net model is 6.3%, 1.8%, and 3.0% higher than that of MDT, rGAN, and MPM, respectively. This indicates that the global-local enhanced attention mechanism and memory-augmented module can significantly improve anomaly video detection performance. Both the methods in this paper and the Stacked RNN are stacked with attention mechanism, but in terms of AUC, our method shows good performance. Mainly because MSTMAF-Net has proposed an advanced video processing architecture multi-scale Swin-Transformer, which combines innovative attention feature fusion mechanism in multi-scale feature extraction. Moreover, residuals are integrated into it, which can better transfer and utilize information and greatly improve its performance. On the Avenue datasets, the AUC of our MSTMAF-Net algorithm reaches 91.3%, which is better than that of the MAAM-Net algorithm's 90.9%. The overall structure of MAAM-Net algorithm is an autoencoder network with feature memory reconstruction. The disadvantage of the MAAM-NET algorithm is that the global feature representation ability of abnormal events is insufficient, so the accuracy rate of the other two data sets is not optimal. And its processing speed is slow, as presented in Table 2. Relative to the memory method MemAE, MSTMAF-Net designed new memory-augmented attention feature fusion modules to store rich prototypes of normal motion, obtaining AUC gains of 5.8%, 7.5% and 5.9%, respectively. Furthermore, the AUC gap on the challenging ShanghaiTech multi-scene datasets show that our MSTMAF-Net can learn complex global spatio-temporal patterns, effectively improving the performance of the unsupervised VAD method in realistic scenarios. In Figure 3, On the UCF-Crime datasets, the proposed method can identify the exception frames very well. The detection accuracy is improved by 1.0% compared to the current best anomaly detection method. In summary, with the support of multi-scale Swin Transformer, attention feature fusion module and memory-augmented module, our method can outperform all competing methods on four datasets in terms of AUC, which verifies the effectiveness of the MSTMAF-Net constructed in the present study.

The ROC curve is presented in Figure 4. The curve of the proposed MSTMAF-Net is closer to point (0,1). This reflects good performance and validates the effectiveness of the proposed algorithm. ShanghaiTech datasets covers a

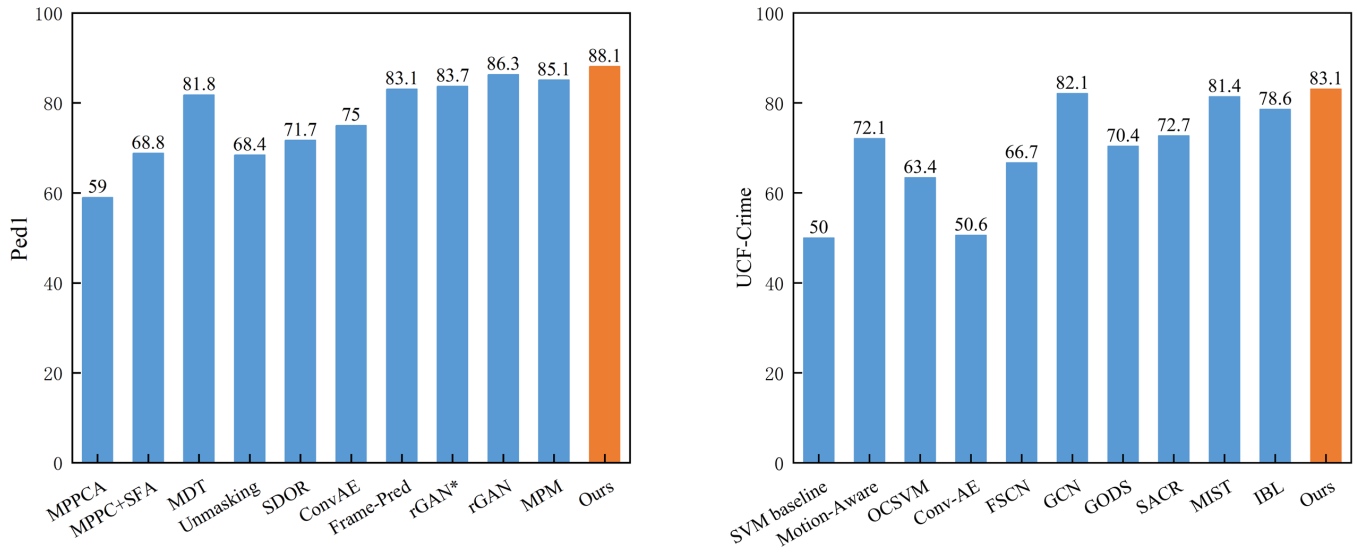

Fig. 3. Results of Quantitative Frame-level AUC Comparison on Ped1 and UCF-Crime.

Table 1. Results of Quantitative Frame-level AUC Comparison With The Current Unsupervised Video Anomaly Detection Methods. Bold Numbers Denote Our Method's Performance

Type	**Method**	**Ped2(%)**	**Avenue(%)**	**ShanghaiTech(%)**
Traditional	MPPCA (Kim and Grauman, 2009)	69.3	–	–
	MDT (Wang and Miao, 2010)	82.9	–	–
	AMDN (Yang et al., 2016)	90.8	–	–
	Unmasking (Tudor Ionescu et al., 2017)	82.2	80.6	–
	AED (Lu et al., 2013)	–	80.9	–
	DF (Del Giorno et al., 2016)	–	78.3	–
Deep Learning-based	HybridAE (Nguyen and Meunier, 2019)	92.2	81.7	68.0
	ConvAE (Hasan et al., 2016)	90.0	70.2	60.9
	MNAD (Medel and Savakis, 2016)	97.0	88.5	70.5
	Stacked RNN (Luo et al., 2017)	92.2	81.7	68.0
	MemAE (Gong et al., 2019)	94.1	83.8	71.2
	CDDAE (Chang et al., 2020)	96.5	86.0	73.3
	IPR (Tang et al., 2020)	96.2	83.7	71.5
	MAAM-Net (Wang et al., 2023a)	97.7	90.9	71.3
	MPM (Lv et al., 2021)	96.9	89.5	73.8
	LMN (Park et al., 2020)	97.0	88.5	70.5
	MSTMAF	**98.2**	**91.3**	**76.4**

Table 2. Modeling Inference Speed vs Baseline Methods.

Methods	Processor	FPS
Unmasking (Tudor Ionescu et al., 2017)	GPU	20.0
Frame-Pred (Liu et al., 2018)	GPU	25.0
AMMC-Net (Cai et al., 2021)	GPU	40.1
MAAM-Net (Wang et al., 2023a)	GPU	79.7
MemAE (Gong et al., 2019)	GPU	86.7
MSTMAF	**GPU**	**90.6**

Table 3. Results of Ablation Study on The Models

Methods	**Methods Combination**	**Ped2 (%)**
Backbone Resnet50 (1)	(1)	95.8
Backbone Swin Transformer (2)	(2)	96.4
Multi-Scale Feature (3)	(2) + (3)	97.4
Dual Attention Feature Fusion (4)	(2) + (3) + (4)	97.6
Tri-Attention Feature Fusion (5)	(2) + (3) + (5)	98.0
Memory-Augmented Module (6)	(2) + (3) + (5) + (6)	98.2

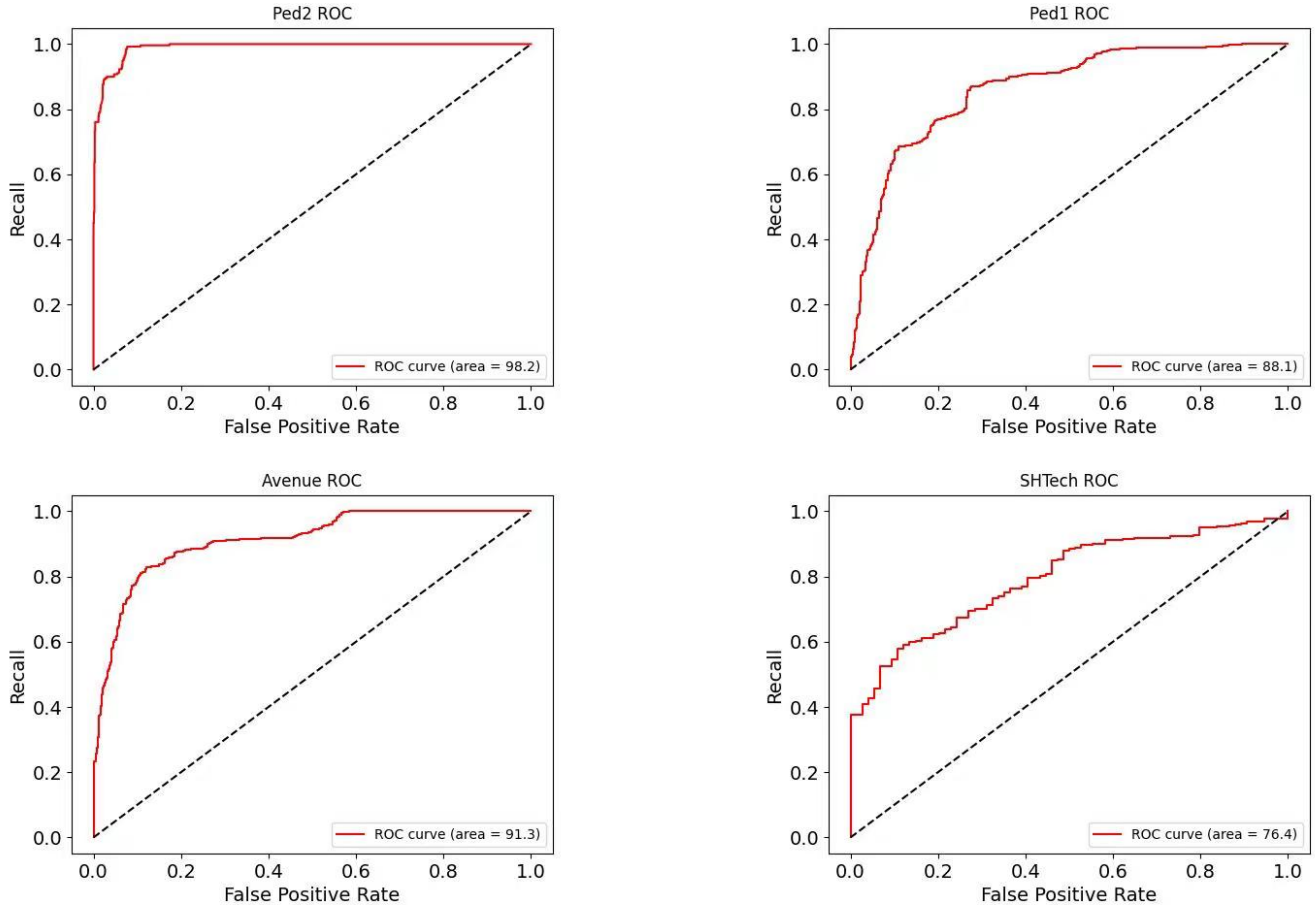

Fig. 4. ROC Curve of MSTMAF-Net on Four Data Sets.

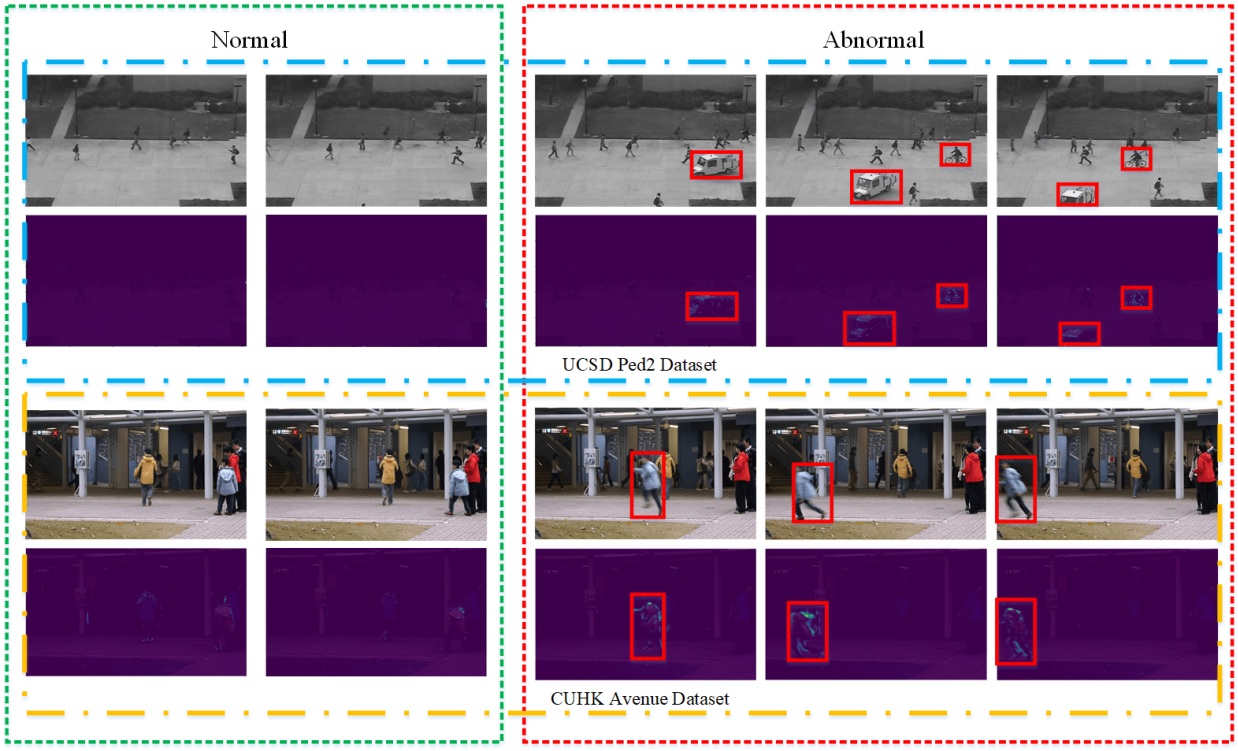

Fig. 5. Temporal Localization of Abnormal Events.

variety of scenarios and has more types of abnormal behaviors than UCSD Ped2 datasets, so it is very challenging. Based on the specific algorithmic performance on the two datasets in Figure 4, The ROC curve of the UCSD Ped2 datasets is significantly closer to point (0,1) than that of the ShanghaiTech datasets, so there is still much room for improving for the datasets with complex scenes.

4.4 Model inference speed

We have listed the processor and FPS (Frames Per Second) numbers for Unmasking, MemAE, MAAM-Net, AMSC-NET, and our MSTMAF-Net, as shown in Table 2, to provide insight into the computational performance of the suggested network. When compared to the baseline, MSTMAF-Net's FPS number shows an increase that is greater than 3.0. This suggests that our model's training and prediction processes don't require an excessive amount of time. This efficiency allows for practical implementation and operation within real-world enviroments. Despite the noteworthy performance of the Augmented Appearance-Motion Network (MAAM-Net) on the Avenue datasets, its reliance on the Patch-based Stride Convolutional Detection (PSCD) algorithm and the incorporation of multiple decoder connections result in a reduced processing velocity. Conversely, our model MSTMAF-Net achieves operational speeds of 90.6 fps. When evaluating both performance and computational efficiency, it is evident that MSTMAF-Net achieves an optimal balance. This prompted us to consider the deployment of real-time inspection in resource-limited end devices.

4.5 Ablation Study

In the current work, we quantitatively analyzed the main model components and demonstrated the performance of these components on the UCSD Ped2 datasets in Table 3. Relative to ResNet50, the addition of the Swin Transformer module resulted in a slight improvement in the AUC, increasing by 0.6% on UCSD Ped2. This enhancement is mainly attributed to the Swin Transformer's focus on processing global features. In relative to the original Swin Transformer structure, our multi-scale Swin Transformer structure shows a 1.0% improvement in detection accuracy. This suggests that multi-scale feature fusion can improve the quality of the encoder. In terms of local attention, the current attention feature fusion emphasize information processing on main characteristics and weaken the background information, thus enhancing feature extraction and significantly improving model performance, including a 0.4% improvement on the UCSD Ped2 dataset. In addition, the memory-augmented module stores prototype features of normal behavior from historical data, facilitating the coexistence of multi-modal data, and contributes to a 0.2% increase in AUC. The AUC values of the MSTMAF-Net surpass those of other models, indicating that the combination of multi-scale Swin Transformer, attention feature fusion, and enhanced memory module greatly improves the performance of abnormal video detection.

4.6 Visualization Analysis

We visualized and analyzed the test results of the model on the UCSD Ped2 and Avenue datasets, as illustrated in Figure 5. Boxes label the target objects exhibiting abnormal behavior. Columns 1 and 2 depict normal events, but columns 3 to 5 showcase abnormal events, including riding a bicycle,

driving on a sidewalk, and running anomalously. The MSTMAF network outlines the edges of abnormal areas with precision and accurately locates them. This precision indicates that the MSTMAF network captures enhanced global-local information, enabling better recording of normal motion information and thereby increasing the differentiation of abnormal frames. The prediction errors in Figure 6 are highly concentrated on the depicted cyclists, vehicle, and those running abnormally. Moreover, the network accurately predicts both the abnormal motion frames of subjects entering the scene and those about to exit. This demonstrates that the proposed memory-augmented attention feature fusion module can improve the prediction accuracy for abnormal behavior by memorizing and learning normal features. Multi-scale swin transformer provides fine-grained extraction of global and local features.

Under normal frame conditions, no anomaly is present, and the anomaly score curve remains low, as with normal pedestrian traffic in a crosswalk. When an anomaly occurs, the affected area exhibits a high anomaly score, marked by a box in Figure 6, signifying an individual riding a bicycle or driving a car on the crosswalk. Figure 6 also displays the outlier scores for the Avenue datasets test videos. A pedestrian walking normally corresponds to a low anomaly score, whereas the sudden running of a person suggests an abnormal event, leading to a sharp increase in the anomaly score. The more pronounced the abnormal behavior, the higher the anomaly score, indicating the model's effectiveness in detecting abnormal events.

5. CONCLUSION

To conclude, this study proposes an innovative video anomaly detection architecture based on multi-scale Swin Transformer and memory-augmented attention feature fusion. The architecture uses multi-scale Swin Transformer as the backbone coding network, and designs a local feature attention fusion module and memory-augmented module. These enhance the global-local processing capability of the model for information such as the motion and appearance of objects in video sequences. The model realizes the interactive fusion of fine-grained local key characteristics with global features. Additionally, the prototype features of normal behavior in historical data are stored, and the coexistence of multi-mode data is realized. Through ablation and comparison experiments on four mainstream publicly available anomaly detection datasets, we have confirmed the effectiveness and performance of the proposed model. Furthermore, the deployment of our method in real-world surveillance systems will have a significant impact on the development of crime prevention systems. The main limitations of MSTMAF-Net lie in its large number of parameters and complex network architecture, which impose substantial demands on computing resources

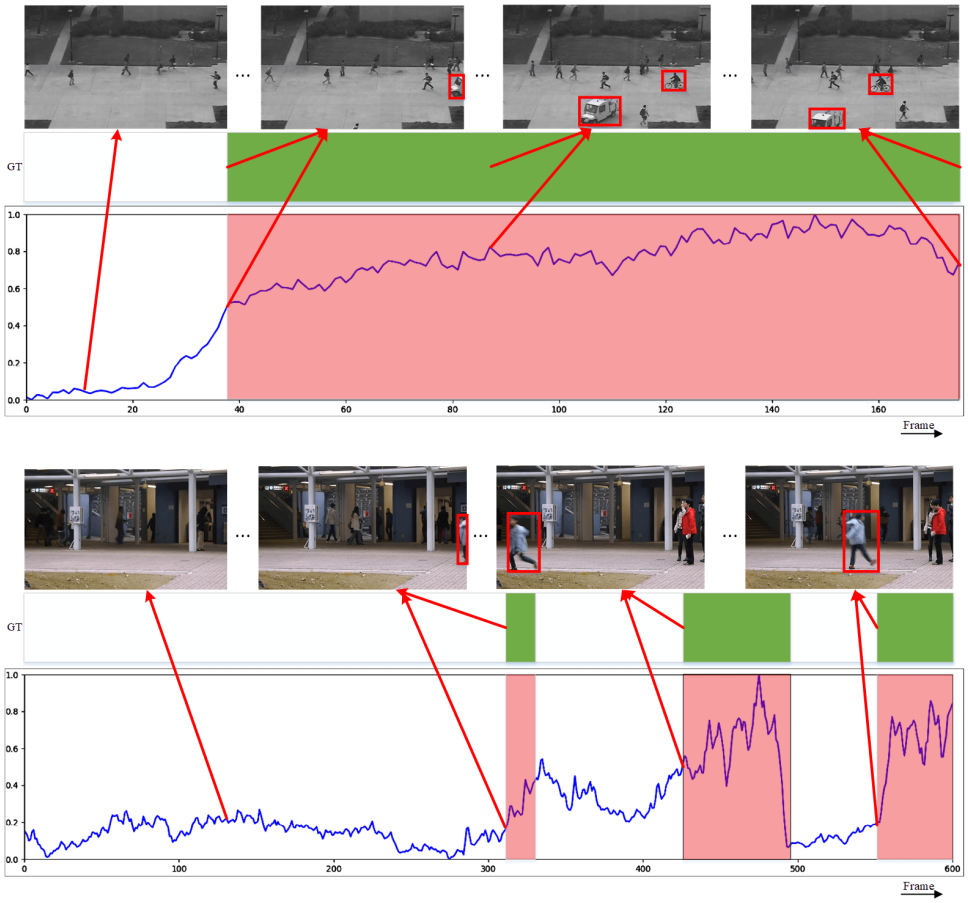

Fig. 6. Results of Frame Localization of Abnormal Events.

and memory during model deployment. These factors not only increase the training and inference costs but also limit its practicality in resource-constrained environments. Future research could focus on model compression, knowledge distillation, lightweight architecture design, and the optimization of efficient inference algorithms to substantially reduce the model's complexity and computational burden while maintaining detection accuracy, thereby enhancing its usability and generalization capability in real-world applications.

ACKNOWLEDGEMENTS

The work is supported by National Natural Science Foundation of China, Grant No.61972133, Project of Leading Talents in Science and Technology Innovation in Henan Province, Grant No.204200510021, Program for Henan Province Key Science and Technology, Grant No.222102210177, 242102211077 and Henan Province University Key Scientific Research Project No.23A520008.

REFERENCES

Barbalau, A., Ionescu, R.T., Georgescu, M.I., Dueholm, J., Ramachandra, B., Nasrollahi, K., Khan, F.S., Moeslund, T.B., and Shah, M. (2023). Ssmtl++: Revisiting self-supervised multi-task learning for video anomaly detection. *Computer Vision and Image Understanding*, 229, 103656.

Cai, R., Zhang, H., Liu, W., Gao, S., and Hao, Z. (2021). Appearance-motion memory consistency network for video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 938–946.

Chang, Y., Tu, Z., Xie, W., and Yuan, J. (2020). Clustering driven deep autoencoder for video anomaly detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, 329–345. Springer.

Del Giorno, A., Bagnell, J.A., and Hebert, M. (2016). A discriminative framework for anomaly detection in large videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 334–349. Springer.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., and Hengel, A.v.d. (2019). Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1705–1714.

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., and Davis, L.S. (2016). Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 733–742.

Jiang, J., Zhu, J., Bilal, M., Cui, Y., Kumar, N., Dou, R., Su, F., and Xu, X. (2022). Masked swin transformer unet for industrial anomaly detection. *IEEE Transactions on Industrial Informatics*, 19(2), 2200–2209.

Kim, J. and Grauman, K. (2009). Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *2009 IEEE conference on computer vision and pattern recognition*, 2921–2928.

IEEE.

Kumar, M., Patel, A.K., Biswas, M., and Shitharth, S. (2023). Attention-based bidirectional-long short-term memory for abnormal human activity detection. *Scientific Reports*, 13(1), 14442.

Lee, J., Nam, W.J., and Lee, S.W. (2022). Multi-contextual predictions with vision transformer for video anomaly detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 1012–1018. IEEE.

Leyva, R., Sanchez, V., and Li, C.T. (2017). Video anomaly detection with compact feature sets for online performance. *IEEE Transactions on Image Processing*, 26(7), 3463–3478.

Li, Y., Song, X., Xu, T., and Feng, Z. (2022). Memory-token transformer for unsupervised video anomaly detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 3325–3332. IEEE.

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

Liu, Y., Yang, D., Fang, G., Wang, Y., Wei, D., Zhao, M., Cheng, K., Liu, J., and Song, L. (2023). Stochastic video normality network for abnormal event detection in surveillance videos. *Knowledge-Based Systems*, 280, 110986.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Lu, C., Shi, J., and Jia, J. (2013). Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, 2720–2727.

Luo, W., Liu, W., and Gao, S. (2017). A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, 341–349.

Lv, H., Chen, C., Cui, Z., Xu, C., Li, Y., and 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*, 15425–15434.

Medel, J.R. and Savakis, A. (2016). Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv:1612.00390*.

Nguyen, T.N. and Meunier, J. (2019). Hybrid deep network for anomaly detection. *arXiv preprint arXiv:1908.06347*.

Park, H., Noh, J., and Ham, B. (2020). Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14372–14381.

Perera, P., Nallapati, R., and Xiang, B. (2019). Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2898–2906.

Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C., and Sebe, N. (2017). Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE international conference on image processing (ICIP)*, 1577–1581. IEEE.

Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Z., and Klette, R. (2018). Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Computer Vision and Image Understanding*, 172, 88–97.

Saligrama, V. and Chen, Z. (2012). Video anomaly detection based on local statistical aggregates. In *2012 IEEE Conference on computer vision and pattern recognition*, 2112–2119. IEEE.

Shao, W., Xiao, R., Rajapaksha, P., Wang, M., Crespi, N., Luo, Z., and Minerva, R. (2023). Video anomaly detection with ntcn-ml: A novel tcn for multi-instance learning. *Pattern Recognition*, 143, 109765.

Tang, Y., Zhao, L., Zhang, S., Gong, C., Li, G., and Yang, J. (2020). Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129, 123–130.

Thakare, K.V., Dogra, D.P., Choi, H., Kim, H., and Kim, I.J. (2023). Rareanom: A benchmark video dataset for rare type anomalies. *Pattern Recognition*, 140, 109567.

Tudor Ionescu, R., Smeureanu, S., Alexe, B., and Popescu, M. (2017). Unmasking the abnormal events in video. In *Proceedings of the IEEE international conference on computer vision*, 2895–2903.

Wang, L., Tian, J., Zhou, S., Shi, H., and Hua, G. (2023a). Memory-augmented appearance-motion network for video anomaly detection. *Pattern Recognition*, 138, 109335.

Wang, L., Zhang, J., Tian, Q., Li, C., and Zhuo, L. (2019). Porn streamer recognition in live video streaming via attention-gated multimodal deep features. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12), 4876–4886.

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

Wang, Y., Liu, T., Zhou, J., and Guan, J. (2023b). Video anomaly detection based on spatio-temporal relationships among objects. *Neurocomputing*, 532, 141–151.

Wu, J., Shu, P., Hong, H., Li, X., Ma, L., Zhang, Y., Zhu, Y., and Wang, L. (2023). Unsupervised encoder-decoder model for anomaly prediction task. In *International Conference on Multimedia Modeling*, 549–561. Springer.

Yang, J., Parikh, D., and Batra, D. (2016). Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5147–5156.