FUDMA Journal of Engineering and Technology
ISSN: 3092-9385
https://fjet.fudutsinma.edu.ng/index.php/fjet/
Volume 2, Issue 1, 2026, 41-50

A Hybrid Deep Learning Model for Efficient Anomaly Detection in Video Surveillance Systems

Abba M. BALA[1*], Abdurrauf G. SHARIFAI[2], Umar S. HARUNA[3]

[1*,2]Department of Computer Science, Faculty of Computing, Northwest University, Kano, Kano State, Nigeria
[3]Department of Cybersecurity, Faculty of Computing, Northwest University, Kano, Kano State, Nigeria

[1*]abmukhtar@nwu.edu.ng, [2]asharifaigarba@yumsuk.edu.ng, [3]shafsonharun@gmail.com

Abstract

Traditional CNNs often face challenges in capturing long-range dependencies and contextual relationships within data, which limits their effectiveness in complex tasks like anomaly detection. To overcome these limitations, we propose an innovative enhanced attention-mechanism hybrid model that combines the strengths of CNNs with Transformer architecture. This hybrid model leverages the powerful feature extraction capabilities of four distinct CNN architectures, VGG16, DenseNet121, ResNet50, and MobileNetV2. To process the training data comprehensively, the extracted features are fused and passed through Swin Transformer which integrates attention mechanisms to capture long-range dependencies within the data effectively, and focuses on the most relevant regions of the input data. The effectiveness of this approach is evaluated on the UCF-Crime benchmark dataset using performance metrics such as ROC-AUC, achieving an outstanding accuracy of 99.2% surpassing existing state-of-the-art methods. Moreover, the model's ability to handle complex video data and extract semantically rich features highlights its potential for real-time surveillance applications where timely and accurate anomaly detection is critical.

Keywords: *CNN, ViT, Swin-Transformer, VAD, Attention Mechanism, Feature Fusion.*

1.0 Introduction

In many applications, such as industrial inspection, traffic monitoring, and surveillances, video anomaly detection can be an essential application. The goal is to automatically recognize odd events or behaviors within video sequences that depart significantly from the standard normal patterns [1]. Surveillance system often produced redundant video data which took up unnecessary storage space. The convenience offered by these cameras comes at the cost of massive amounts of monitoring data that keep accumulating from the surveillance feed, which creates immense challenges in terms of storage analysis, as well as the retrieval of data [2]. Building an effective monitoring system that can identify any odd behaviors which could result in hazardous circumstances is essential to addressing these problems by lowering human error and storage expenses. Hence an intelligent system that can automatically identify unusual events and detect anomalous events in streaming videos is therefore highly needed [3].

Traditional methods for anomaly detection often rely heavily on hand-crafted features and statistical models such as SVM and Naibe Bayes which can be limited in their ability to capture complex spatio-temporal patterns in video data [4]. However, the advancements in deep learning techniques, particularly CNNs have shown promising results in video analysis. CNNs are capable of automatically learning hierarchical features directly from raw video frames, leading to improved performance in various tasks, including anomaly detection as highlighted by [5]. Despite their success, CNNs predominantly emphasize local spatial features and often fall short in capturing long-range dependencies which are essential aspect for comprehending complex and dynamic events in video data [6]. This limitation underscores the need for advanced architectures that integrate temporal awareness, where understanding the context of distant objects is very crucial in detecting anomalies paving the way for more robust and context-aware video anomaly detection solutions. Furthermore, we recently witnessed a groundbreaking transformation in the field of image processing in Computer Vision (CV) with the introduction of new method called "Vision Transformer" (ViT) by Dosovitskiy et al. [7]. ViTs introduced a paradigm shift by departing from traditional convolution-based architectures and leveraging self-attention mechanisms as their core operational principle. This innovative approach enables ViTs to effectively capture global dependencies effectively within images and provides a fresh perspective on feature representation and significantly pushes forward the progress in computer vision tasks.

To overcome the above-mentioned limitations of CNN in video anomaly detection domain, this paper introduces a novel hybrid model that combines the strengths of both CNN architectures with the capabilities of ViT (Swin Transformer in particular). By incorporating attention mechanisms into the CNN framework, the method can adaptively concentrate on the most significant areas of the video sequence improving its capability to detect anomalous events effectively. The proposed approach utilizes pre-trained CNN architectures, including VGG16, DenseNet121, ResNet50, and MobileNetV2 to extract rich and diverse spatial features from the video

Received: 18/01/2026; Revised: 29/01/2026; Accepted: 05/02/2026; Published: 02/03/2026

frames. These spatial features are subsequently fused to form a comprehensive representation which is then processed by the Swin Transformer. The Transformer effectively captures long-range temporal dependencies enabling the model to identify subtle and complex anomalies within video data with improved accuracy and robustness.

The principal contributions of this study are:

1. We construct a robust spatial feature extraction pipeline by integrating multiple CNN architectures (VGG16, DenseNet121, ResNet50, and MobileNetV2). This ensemble method exploits the complementary representational strengths of heterogeneous convolutional backbones, resulting in richer local feature characterization and improved predictive stability.
2. Introduced a hybrid architecture that unifies fused CNN-derived spatial descriptors with a Swin Transformer, thereby incorporating self-attention mechanisms capable of modeling long-range temporal dependencies and global contextual relations. This design directly addresses the limitations of conventional CNN baselines in anomaly understanding and yields a more context-aware detection framework.
3. Perform an extensive assessment using standard benchmark datasets, demonstrating that the proposed model outperforms recent state-of-the-art approaches in accuracy and robustness for video anomaly detection.

The subsequent sections of this paper are organized as follows: Section 2 discusses the relevant literature. Section 3 presents the principles of CNN architectures and Swin transformers. Section 4 encompasses the results, discussion, and analysis. Finally, Section 5 outlines the conclusion and suggestions for future research.

2.0 Related Work

The field of video anomaly detection has evolved from traditional methods relying on handcrafted features to modern approaches leveraging deep learning techniques. Early methods used statistical techniques and motion patterns to identify anomalies. The advent of deep learning introduced CNNs and RNNs enabling automated feature extraction and temporal modeling, in this section we take a look at those previous approaches that had been used for video anomaly detection.

2.1 Traditional Methods: Early research in video anomaly detection primarily relied on traditional methods that employed hand-crafted features and statistical modeling to identify unusual patterns in surveillance videos. These approaches were shaped by the technological constraints and limited computational resources available at the time, as well as the early understanding of the complexities inherent in video data. One of the most commonly used techniques was optical flow introduced by [8], which analyzes the displacement of pixels between consecutive frames to capture motion information. Optical flow provided a foundational understanding of movement dynamics, such as the speed and direction of objects. Moreover, researchers explored motion trajectories such as [9], which involved tracking the movement of objects over time. This method allowed for the identification of irregularities in motion patterns, such as abrupt changes in direction or speed, which could indicate anomalous behavior. Additionally, bag-of-words representations were employed to encode video sequences as histograms of visual words, summarizing motion and appearance features into a compact representation [10]. These features were typically paired with statistical models to establish a baseline for normal activities. Hidden Markov Models (HMMs) by [11], were employed to capture temporal dependencies, making them suitable for modeling sequential activities like walking or running. Support Vector Machines (SVMs) were also popular due to their strong performance in high-dimensional spaces, often used for classifying whether a sequence of frames contained normal or anomalous behavior as often used in the study of [12].

While these early methods demonstrated promise in controlled environments, they faced several critical limitations that hindered their effectiveness in real-world applications. First, the design of effective hand-crafted features required extensive domain expertise and an intimate understanding of the underlying video dynamics. The manual nature of this process meant that feature sets were often tailored to specific scenarios, making it challenging to generalize across diverse and complex environments. Furthermore, hand-crafted features often struggled to capture the rich spatio-temporal dynamics present in real-world video data [13]. For example, subtle variations in object appearance, interactions, or contextual relationships might go unnoticed. These approaches were also found to be susceptible to environmental influences including variations in lighting conditions, camera angles, and background clutter, which could introduce noise or distortions into the feature extraction process [14]. Additionally, these approaches lacked the flexibility to adapt to new or unforeseen scenarios, as their reliance on predefined features limited their capacity for learning from data.

2.2 CNN-based Methods: The deep learning revolution driven by the success of CNNs in image recognition has significantly impacted the field of video analysis. CNNs have demonstrated outstanding capabilities in automatically learning hierarchical representations from raw image and video data, eradicating the need for labor-

Received: 18/01/2026; Revised: 29/01/2026; Accepted: 05/02/2026; Published: 02/03/2026

intensive feature engineering. Several studies have explored the use of CNNs for feature extraction from individual video frames, followed by RNNs such as LSTMs or GRUs to model temporal dependencies. For instance [15], proposed method that designed to segment a video into distinct parts using a shot boundary detection algorithm. Once segmented, a selected sequence of frames is fed into a Convolutional Neural Network to extract crucial spatiotemporal features. These extracted features are then enriched with valuable contextual information enhancing the model's ability to identify abnormal events. Finally, LSTM cells are employed to analyze the spatiotemporal patterns within each anomalous event sample, learning from the frame sequences to enable precise anomaly detection. [16] also introduces an intelligent anomaly detection framework leveraging deep feature extraction, specifically optimized for efficient operation within surveillance networks while reducing time complexity. In their approach, consecutive frames are first processed through a pre-trained CNN to extract essential spatiotemporal features. These deep features are then passed into a multilayer Bi-Directional Long Short-Term Memory (BD-LSTM) which effectively classifies the ongoing events as either anomalous or normal within the complex surveillance environments of smart cities. Study of [17] adopt simple and effective strategy for capturing spatiotemporal features by utilizing deep three-dimensional convolutional networks (3D ConvNets) trained on the UCF Crime video dataset. The process involves extracting 3D features by incorporating frame-level information and enhancing spatial representation through augmentation techniques. [18] also present lightweight convolutional neural network based on anomaly detection framework designed to operate efficiently within surveillance settings, minimizing computational time. CNN is being used for spatial features extraction and then inputted to residual attention-based long short-term memory (LSTM) network. This LSTM network is adapted at accurately identifying anomalous activities within surveillance footage. The fusion of representative CNN features with the concept of residual blocks in LSTM for sequence.

The work of [19] utilizes pre-trained models for high-level feature extraction, which are then employed to train a denoising autoencoder (DAE). Remarkably, this process requires minimal training time, achieving satisfactory detection performance comparable to top-performing methods within a short timeframe (specifically, within 10 seconds on UCSD Pedestrian datasets). Also, [20] proposed a concept termed Aggregation of Ensembles (AOE) aimed at detecting anomalies in video footage of crowded scenes. The approach capitalizes on the existing capabilities of pre-trained ConvNets alongside a diverse set of classifiers. By leveraging an ensemble of fine-tuned ConvNets, they hypothesize that different CNN architectures capture varying levels of semantic representation from crowd-scene videos leading to the extraction of more comprehensive feature sets. [21] introduce an approach called the Deep Spatiotemporal Translation Network (DSTN), it leverages Generative Adversarial Network (GAN) and Edge Wrapping (EW) techniques. During training, the system only learns from videos showing normal activity. It calculates how objects move within each video (optical flow) to understand typical motion patterns. When the system encounters new videos during testing, it compares the motion patterns to what it learned during training. If the motion patterns in the new video are significantly different from the learned normal patterns, the system flags the event as an anomaly.

Despite significant advancements over traditional methods, CNNs and RNNs often face limitations in capturing very long-range dependencies, due to challenges like vanishing gradients and restricted receptive fields. This limitation becomes particularly pronounced in tasks requiring a holistic understanding of sequences such as video anomaly detection or document summarization. For example, in video data, detecting anomalous events often requires understanding context spread across several minutes. However, standard CNNs struggle to effectively maintain this context leading to suboptimal performance in such scenarios.

2.3 Transformer-based Methods: Transformers, particularly Vision Transformer (ViT) and its variants, have revolutionized computer vision. This Transformer technique, initially introduced to natural language processing, have demonstrated remarkable capabilities in capturing long-range dependencies. With regard to video surveillance systems, some researchers try to adapt the power of transformers in the field of VAD too. Like the study of [22] which introduce a baseline model named Anomaly Detection with Transformers (ANDT). This model treats a series of video frames as a sequence of short tubelets. It uses a Transformer to analyze these clips and extract meaningful features. Then, it uses another part of the network (the decoder) to predict what the next frame in the sequence should look like. During the training phase the network learns normality, while in the testing phase it indicates clips that contains unexpected and unpredictable movements suggesting an anomaly. [23] Introduce approach called AnoViT, a vision transformer-based encoder. The model is trained to recognize normal patterns while understanding global relationships between different image patches, by passing these patches through multiple self-attention layers. AnoViT generates a feature map that preserves their original spatial information allowing for more accurate anomaly detection and localization. Furthermore, the research of [24] presents an interpretable anomaly detection approach by employing a ViT-based Deep Support Vector Data Description (SVDD). The technique applies SVDD to address the issue encountered in Multilayer Perceptron (MLP) where the outcome varies due to weight adjustments affected by the patch sequence data utilized in the vision transformer.

Received: 18/01/2026; Revised: 29/01/2026; Accepted: 05/02/2026; Published: 02/03/2026

[25] designed the transformer model with three distinct contextual prediction streams; masked, whole and partial. In order to capture different normalcy patterns in a video, the model learns to anticipate missing frames between successive normal frames. When encountering anomalous occurrences that do not fit the learnt context, this leads to a significant reconstruction error. To assess the effectiveness of their approach, the model is tested on publicly available benchmark datasets, including UCSD Pedestrian2, CUHK Avenue, and ShanghaiTech. [26] Introduce the novel Dual-attention Transformer and Discriminative Flow (DADF). Their approach obtains embeddings with multi-scale priors by utilizing pre-trained networks, followed by ascertain dual attention mechanisms namely self-attention and memorial-attention. This enables a two-tier reconstruction of the prior embeddings, taking into account both sequential dependencies and normality correlations. [27] propose a prediction-based VAD approach named Trans Anomaly. The method integrates U-Net and Video Vision Transformer (ViViT) to enhance the detention of extensive temporal information and broader global contexts. To fully utilize ViViT for prediction, they modified the model to enhance its capability for video prediction. Table 1 below shows the contemporary advantage of ViTs over CNNs and vice versa.

Despite their strong ability to capture global context and long-range dependencies, transformers face few notable limitations. Their self-attention mechanism incurs high computational and memory complexity, particularly for long video sequences and high-resolution. In addition, transformers are highly data-hungry and often struggle to generalize well in low data settings. Unlike CNNs, Vision Transformers lack inherent inductive biases for locality and translation invariance to learn spatial relationships entirely from data, this makes transformers less effective at capturing fine-grained local motion cues like subtle anomalies unless hybrid architectures or additional constraints are introduced.

Table 1: Strength of CNN and ViT

ViTs are good for capturing long-range dependencies	**CNNs have good strong inductive bias and are good for capturing spatial and local receptive fields**
Long-range dependencies refer to the need for a model to understand relationships or patterns that span significant portions of a data sequence. In video anomaly detection, such dependencies are crucial, as anomalous events often unfold over extended timeframes or require contextual understanding across multiple segments of a video. For instance: • Detecting a person behaving suspiciously in a crowd may require analyzing their movement patterns over several minutes. • Understanding gradual changes, such as an escalation in crowd density, often demands a broader temporal context.	CNNs, with their convolutional layers, excel at capturing local patterns and spatial relationships within an image. Their filters, which slide across the image, are designed to detect features such as edges, corners, and textures within small receptive fields. This inherent structure allows CNNs to effectively learn hierarchical representations, from basic features to more complex patterns.

3.0 Materials and Methods

The proposed framework synergistically combines the powerful feature extraction abilities of CNNs with the sophisticated attention mechanisms of Swin Transformer to create a robust methodology for video anomaly detection. The architecture consists of three core components: feature extraction, feature fusion, and attention integration

Received: 18/01/2026; Revised: 29/01/2026; Accepted: 05/02/2026; Published: 02/03/2026

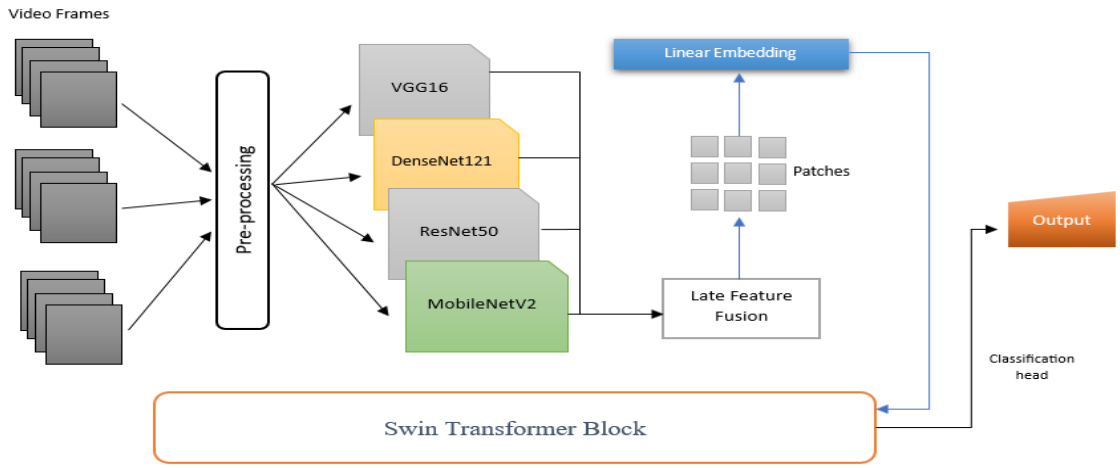

Figure 1: Our proposed model architecture

3.1 Feature Extraction: To extract spatial features from input video frames, we leverage a combination of this powerful CNN architectures: VGG16, DenseNet121, ResNet50, and MobileNetV2. Each of these models offers unique strengths, capturing diverse and complementary feature representations essential for robust video analysis. VGG16, known for its simplicity and depth, excels in capturing fine-grained details. DenseNet121 ensures efficient feature reuse through its dense connectivity, enhancing representational power while reducing computational redundancy. ResNet50 employs residual connections to alleviate the vanishing gradient problem, enabling the extraction of deeper and more meaningful features. MobileNetV2, with its lightweight structure and inverted residuals, provides computational efficiency while preserving high-quality spatial features. The integration of these complementary architectures ensures a rich multi-perspective understanding of the input data, laying a solid foundation for subsequent processing and analysis.

3.2 Feature Fusion: The features extracted from the CNN models are combined through a concatenation process, creating a unified and comprehensive representation of the input data. To address the challenge of high-dimensionality resulting from this fusion, the combined features are passed through a dimensionality reduction layer. This step is critical as it not only enhances computational efficiency by reducing the complexity of the feature space but also strategically retains the most relevant and informative aspects of the data. By preserving essential information while eliminating redundancy, the dimensionality reduction process ensures the model remains both resource-efficient and highly effective in capturing the nuances necessary for robust anomaly detection.

3.3 Attention Integration: To integrate attention, we harness the power of the Swin Transformer, the model divides the fused data into non-overlapping patches and processes them through multi-head self-attention layers by using sliding window attention and shifting windows across layers, The hierarchical structure of the Swin Transformer allows it to capture fine-grained features at different scales, enabling it to focus on both global and local relationships within the data, improving its ability to understand complex patterns across long sequence of the data.

A typical Swin Transformer block comprises a window-based multi-head self-attention (W-MSA) module and a shifted window multi-head self-attention (SW-MSA) module, these are followed by a two-layer MLP incorporating GELU nonlinearity. Additionally, Layer Normalization (LN) is applied before each MSA module and MLP, while residual connections are introduced after each module to enhance stability and performance.

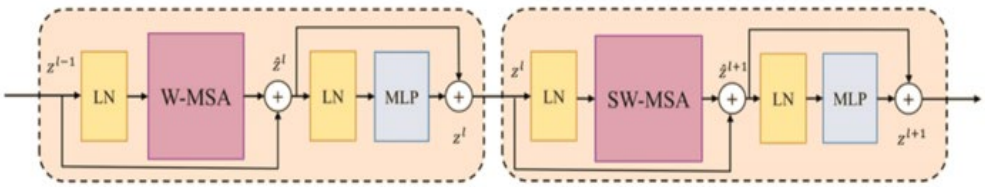

Figure 2: Self-Attention and Shifted Window Self-Attention mechanism for Swin Transformer Block

Received: 18/01/2026; Revised: 29/01/2026; Accepted: 05/02/2026; Published: 02/03/2026

$$\hat{z}^l = W - MSA\left(LN(Z^{L-1})\right) + Z^{L-1},$$
$$z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l,$$
$$\hat{z}^{l+1} = SW - MSA\left(LN(z^l)\right) + z^l,$$
$$\hat{z}^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1}, \quad (1)$$

Here, $\hat{z}^l$ and z^l represent the (S)W-MSA module and the MLP module for block l, respectively. W-MSA and SW-MSA refer to window-based multi-head self-attention mechanisms utilizing standard and shifted window partitioning configurations, respectively.

Shifted window-based self-attention is a modification of the standard self-attention mechanism used in transformer-based models, designed to efficiently capture local dependencies in the input sequence. This technique is particularly relevant in vision transformer architectures, where capturing both local and global context is crucial for effective feature extraction from images. In standard self-attention mechanisms, each token (or patch in the case of images) attends to all other tokens, leading to a computational complexity that scales quadratically with the sequence length. To address this issue and enhance computational efficiency, shifted window-based self-attention introduces a mechanism where tokens only attend to a local neighborhood of other tokens, significantly reducing the computational cost.

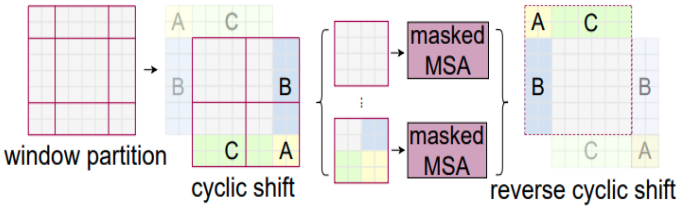

Figure 3: Shifted window cycle

4.0 Experiment and Result

4.1 Description of Datasets

The experiments were conducted using the UCF-Crime dataset, a widely used benchmark for video anomaly detection. This dataset features lengthy unedited surveillance footage showcasing 13 types of real-world anomalies. As shown in Table 2, it's a substantial dataset comprising 950 videos. Given that all videos in this dataset originate from surveillance cameras, it closely mirrors the conditions found in real-world city monitoring systems.

Table 2: Statistical Information of UCF-Crime Dataset

Types of anomalies	The number of videos	Training data	Testing data
Abuse.	50.	48.	02
Arrest.	50.	45.	05
Assault	50.	47	03
Arson	50.	41	09
Shooting	50.	27	23
Fighting	50	45	5
Explosion	50	29	21
Vandalism	50	45	05
Shoplifting	50	29	21
Stealing	100	95	05
Burglary	100	87	13
Robbery	150	145	05
Accident	150	127	23
Sum	950	810	140

4.2 Experimental Setup

The components of experimental environment are presented in Table 3. Due to resource constraints, we adapt the use of online platform specifically "Kaggle" that provide us with free GPU Nvidia P100 which accelerates the process of our training. It has 16GB high bandwidth memory with 3584 CUDA cores. Python was selected as the programming language for model training, given its widespread use and rich ecosystem in the area of deep learning. For the application programming interface, TensorFlow and Keras were chosen as the primary development frameworks due to their robust support, extensive community resources, and flexibility in designing and implementing deep learning models. These tools collectively provided an efficient and effective environment for developing and testing the proposed anomaly detection framework.

Received: 18/01/2026; Revised: 29/01/2026; Accepted: 05/02/2026; Published: 02/03/2026

Table 3: Experimental Environment

Component	Name
GPU Card	Nvidia P100
Platform	Tensorflow with Keras
IDE	Kaggle Notebook
Language	Python
RAM	29GB

4.3 Evaluation Metrics

The proposed system is evaluated using Area Under Curve (AUC). It measures the total two-dimensional area beneath the ROC (Receiver Operating Characteristic) curve, spanning from (0,0) to (1,1). The Equal Error Rate (EER) determines the threshold where the false acceptance rate and false rejection rate are identical. The point at which these rates converge is known as the equal error rate. A higher AUC indicates better performance, whereas a lower EER signifies improved accuracy. The relationship between AUC and EER is depicted in the figure 4 below.

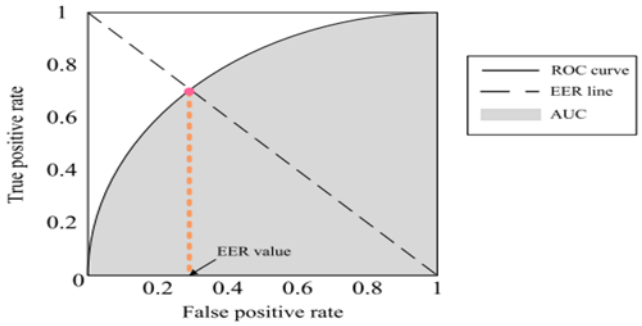

Figure 4: The relationship between AUC and EER

4.4 Result and Discussion

The table 4 below presents a frame-based AUC-comparison of accuracies between the proposed method and existing state-of-the-art approaches using the UCF-Crime benchmark dataset. The result shows that our method attains superior performance, with a remarkable accuracy of 99.2%, surpassing all previously reported methods. This highlights the effectiveness of our model in detecting anomalies with high precision and reliability, establishing it as a significant advancement in the domain of VAD. The exceptional performance underscores the robustness of the model's architecture and its capability to capture nuanced spatiotemporal patterns in complex surveillance scenarios.

Figures 5 and 6 below provide an extensive view of the training process by illustrating the accuracy and loss curves plotted against the number of epochs. These graphs are crucial for understanding the model's learning trajectory and evaluating its convergence during training. Upon examining the accuracy graph, it is evident that the model begins to show promising performance early on, but it reaches optimal accuracy after approximately 13th epochs. This shows that the model successfully learned the necessary features and patterns required for effective anomaly detection within the initial phase of training. As we move further along the training process, the accuracy curve plateaus, indicating that the model has achieved a high level of performance and is no longer significantly improving after the 15th epoch. This suggests that further training beyond this point may not yield substantial gains in accuracy, which is often a sign of overfitting or the model having reached its maximum capacity given the current data and architecture.

Table 4: AUC comparison of the proposed method with state-of-the-art techniques on the UCF Crime dataset

Methods	**AUC%**
Raza et al. [17]	45.0
Ullah et al. [28]	51.0
Waseem et al. [16]	78.43
Muhammad et al. [29]	79.54
Tian et al. [30]	84.30
Haq, et al. [31]	85.53
Majhi et al. [32]	86.37
Sharif et al. [33]	88.0

Received: 18/01/2026; Revised: 29/01/2026; Accepted: 05/02/2026; Published: 02/03/2026

Methods	AUC%
Mangai et al. [34]	95.6
Kim et al. [35]	97.0
Amin et al. [15]	98.0
Our proposed Method	99.2

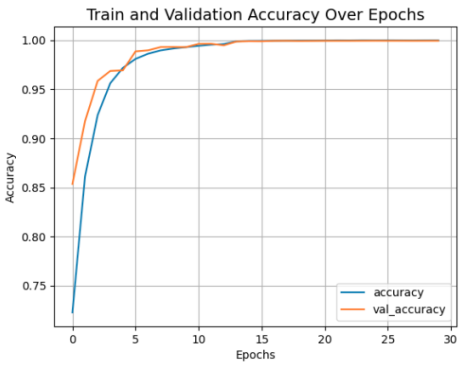

Figure 5. Accuracy

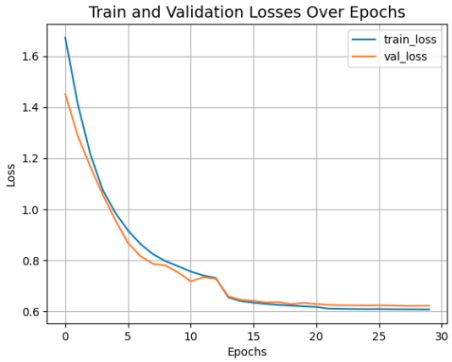

Figure 6. Loss

4.0 Conclusion and Future Work

We presented a hybrid CNN-Swin Transformer framework for video anomaly detection leveraging the strengths of multiple CNN architectures and attention mechanisms, and our experimental evaluations demonstrate that the hybridized architecture outperforms traditional CNN-based approaches in capturing complex spatial-temporal patterns and long-range dependencies in video data. The model exhibits superior accuracy in identifying anomalies particularly in scenarios with subtle or dispersed irregularities. The hierarchical and attention-driven architecture of the hybridized architecture proves to be instrumental in addressing the limitations of CNNs offering a more significant understanding of video content.

Future research could focus on integrating multimodal data such as audio signals and thermal imaging to provide a richer and more comprehensive understanding of scenes, thereby enhancing anomaly detection accuracy. Additionally, developing unsupervised or semi-supervised frameworks would help reduce reliance on extensive labeled datasets, enabling the detection of previously unseen anomalies in diverse and large-scale environments. Improving the model's interpretability through explainable AI (XAI) techniques, such as attention visualizations or heatmaps could foster greater trust and usability in critical surveillance applications. Lastly, integrating hybrid CNN-Transformer models with large language models (LLMs) and other cutting-edge architectures presents an exciting avenue to significantly enhance precision, adaptability, and contextual reasoning in anomaly detection systems paving the way for robust real-world applications.

References

[1] Y. Lu and Y. Wang, "Anomaly Detection in Surveillance Videos using Deep Learning," 2020.

[2] C. S. Soumya, L. Manjula, N. Pallavi, and D. N. Disha, "Enhanced Supervision of Indoor Surveillance Video Using Deep Learning," Eur. Chem. Bull. 2023, vol. 12, no. 10, pp. 12716–12725, 2023, doi: 10.48047/ecb/2023.12.10.9072023.24/08/2023.

[3] H. T. Duong, V. T. Le, and V. T. Hoang, "Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey," Sensors, vol. 23, no. 11, 2023, doi: 10.3390/s23115024.

[4] A. Berroukham, K. Housni, M. Lahraichi, and I. Boulfrifi, "Deep learning-based methods for anomaly detection in video surveillance: a review," Bull. Electr. Eng. Informatics, vol. 12, no. 1, pp. 314–327, 2023, doi: 10.11591/eei.v12i1.3944.

[5] U. Waseem, T. Hussain, F. U. M. Ullah, M. Y. Lee, and S. W. Baik, "TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection," Eng. Appl. Artif. Intell., vol. 123, 2023, doi: 10.1016/j.engappai.2023.106173.

[6] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, "ViTAE: Vision Transformer Advanced by Exploring Intrinsic Inductive Bias," NeurIPS, 2021.

[7] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR 2021, Oct. 2021. [Online]. Available: http://arxiv.org/abs/2010.11929

[8] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Textures of optical flow for real-time anomaly detection in crowds," in *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2011*, 2011, pp. 230–235. doi: 10.1109/AVSS.2011.6027327.

Received: 18/01/2026; Revised: 29/01/2026; Accepted: 05/02/2026; Published: 02/03/2026

[9] N. Madan, A. Farkhondeh, K. Nasrollahi, S. Escalera, and T. B. Moeslund, "Temporal Cues from Socially Unacceptable Trajectories for Anomaly Detection," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[10] W. Sultani, C. Chen, and M. Shah, "Real-world Anomaly Detection in Surveillance Videos," 2018. [Online]. Available: http://arxiv.org/abs/1801.04264

[11] P. Sok, "Activity Recognition for Incomplete Spinal Cord Injury Subjects Using Hidden Markov Models," 2018. [Online]. Available: https://ecommons.luc.edu/luc_theses

[12] B. M'hamed Abidine, L. Fergani, B. Fergani, and M. Oussalah, "The Joint Use of Sequence Features Combination and Modified Weighted SVM for Improving Daily Activity Recognition," 2018.

[13] H. Wang and L. Wang, "Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks," 2017.

[14] T. Bouwmans, C. Silva, C. Marghes, M. S. Zitouni, H. Bhaskar, and C. Frelicot, "On the role and the importance of features for background modeling and foreground detection," Comput. Sci. Rev., vol. 28, pp. 26–91, 2018, doi: 10.1016/j.cosrev.2018.01.004.

[15] S. Amin et al., "EADN: An Efficient Deep Learning Model for Anomaly Detection in Videos," Mathematics, vol. 10, no. 9, 2022, doi: 10.3390/math10091555.

[16] U. Waseem, U. Amin, I. U. Haq, K. Muhammad, M. Sajjad, and S. W. Baik, "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," Multimed. Tools Appl., vol. 80, no. 11, pp. 16979–16995, 2021, doi: 10.1007/s11042-020-09406-3.

[17] R. H. Raza, I. R. Road, K. Muhammad 5, and W. Anwar, "Anomaly Recognition from Surveillance Videos using 3D Convolution Neural Network," 2021. [Online]. Available: https://Arxiv.Org/Abs/2101.01073.

[18] U. Waseem and T. Hussain, "An efficient anomaly recognition framework using an attention residual lstm in surveillance videos," Sensors, vol. 21, no. 8, 2021, doi: 10.3390/s21082811.

[19] C. Wu, S. Shao, C. Tunc, P. Satam, and S. Hariri, "An explainable and efficient deep learning framework for video anomaly detection," Cluster Comput., vol. 25, no. 4, pp. 2715–2737, 2022, doi: 10.1007/s10586-021-03439-5.

[20] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, and G. S. Walia, "Crowd anomaly detection using Aggregation of Ensembles of fine-tuned ConvNets," Neurocomputing, vol. 371, pp. 188–198, 2020, doi: 10.1016/j.neucom.2019.08.059.

[21] T. Ganokratanaa, S. Aramvith, and N. Sebe, "Unsupervised Anomaly Detection and Localization Based on Deep Spatiotemporal Translation Network," IEEE Access, vol. 8, pp. 50312–50329, 2020, doi: 10.1109/ACCESS.2020.2979869.

[22] P. Jin, L. Mou, G.-S. Xia, and X. X. Zhu, "Anomaly Detection in Aerial Videos with Transformers," IEEE TGRS, 2022, doi: 10.1109/TGRS.2022.3198130.

[23] Y. Lee and P. Kang, "AnoViT: Unsupervised Anomaly Detection and Localization with Vision Transformer-based Encoder-Decoder," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9592–9600. [Online]. Available: http://arxiv.org/abs/2203.10808

[24] J. W. Baek and K. Chung, "Explainable Anomaly Detection Using Vision Transformer Based SVDD," Comput. Mater. Contin., vol. 74, no. 3, pp. 6573–6586, 2023, doi: 10.32604/cmc.2023.035246.

[25] J.-Y. Lee, W.-J. Nam, and S.-W. Lee, "Multi-Contextual Predictions with Vision Transformer for Video Anomaly Detection," Jun. 2022. [Online]. Available: http://arxiv.org/abs/2206.08568

[26] H. Yao, W. Luo, and W. Yu, "Visual Anomaly Detection via Dual-Attention Transformer and Discriminative Flow," IEEE Trans. Ind. Informatics, 2023. [Online]. Available: http://arxiv.org/abs/2303.17882

[27] H. Yuan, Z. Cai, H. Zhou, Y. Wang, and X. Chen, "TransAnomaly: Video Anomaly Detection Using Video Vision Transformer," IEEE Access, vol. 9, pp. 123977–123986, 2021, doi: 10.1109/ACCESS.2021.3109102.

[28] W. Ullah, T. Hussain, and S. W. Baik, "Vision transformer attention with multi-reservoir echo state network for anomaly recognition," Inf. Process. Manag., vol. 60, no. 3, 2023, doi: 10.1016/j.ipm.2023.103289.

[29] Z. Z. Muhammad, M. Arif, S. Hochul, and and Seung-Ik Lee, "A Self-Reasoning Framework for Anomaly Detection Using Video-Level Labels," IEEE Signal Process. Lett., vol. X, no. Y, 2021.

[30] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12150–12158.

[31] I. U. Haq et al., "Efficient attention-based CNN-LSTM model for video anomaly detection," IEEE Access, vol. 9, pp. 140104–140116, 2021, doi: 10.1109/ACCESS.2021.3116895.

[32] S. Majhi, R. Dai, Q. Kong, L. Garattoni, G. Francesca, and F. Brémond, "OE-CTST: Outlier-Embedded Cross Temporal Scale Transformer for Weakly-supervised Video Anomaly Detection," in IEEE/CVF

Received: 18/01/2026; Revised: 29/01/2026; Accepted: 05/02/2026; Published: 02/03/2026

Conference on Computer Vision and Pattern Recognition (CVPR), 2024. [Online]. Available: https://github.com/snehashismajhi/OECTST

[33] M. H. Sharif, L. Jiao, and C. W. Omlin, "CNN-ViT Supported Weakly-Supervised Video Segment Level Anomaly Detection," Sensors, vol. 23, no. 18, 2023, doi: 10.3390/s23187734.

[34] P. Mangai, M. Kalaiselvi Geetha, and G. Kumaravelan, "Two Stream Spatial-Temporal Feature Extraction And Classification Model For Anomaly Event Detection Using Hybrid Deep Learning Architectures," J. Theor. Appl. Inf. Technol., vol. 101, no. 18, 2023. [Online]. Available: www.jatit.org

[35] J. Kim et al., "VT-ADL: A Vision Transformer Network for Image Anomaly Detection and Localization," in IEEE 30th International Symposium on Industrial Electronics (ISIE), 2023, pp. 1–6, doi: 10.1109/ISIE51358.2023.10227928.

Received: 18/01/2026; Revised: 29/01/2026; Accepted: 05/02/2026; Published: 02/03/2026