Detecting Video Anomalies through Deep Learning Using Optimized Attention-Enhanced Auto encoders

Shaik Hafijulla Irshad[1], S Suneetha[2]
[1,2]*Assistant Professor, ECE Department, Godavari Institute of Engineering and Technology (Autonomous), Rajahmundry,AP,India.*

***Abstract*—Automated monitoring systems, security, and surveillance all depend on video anomaly detection. Conventional methods frequently have limited applicability in complicated situations and significant false alarm rates. Here, we provide a novel deep learning framework that enhances anomaly detection performance in video sequences by utilizing Optimised Attention-Enhanced Auto encoders. By incorporating an attention mechanism into an auto encoder architecture, our approach improves the learning of spatial-temporal representations by allowing the model to suppress irrelevant information and concentrate on important regions. To increase accuracy and efficiency, we also use an optimization technique to fine-tune network setups and hyper parameters. Experimental tests on benchmark datasets, such as ShanghaiTech and UCSD Ped2, show that our strategy drastically lowers false positives while achieving better performance than state-of-the-art techniques. Optimised Attention-Enhanced Auto encoders are a revolutionary deep learning-based framework that we present in this work to detect video anomalies. The attention mechanism in our model is integrated into auto encoder architecture to effectively learn normal patterns in video sequences and improve the representation of spatial-temporal information. By concentrating on the most pertinent areas, the attention mechanism helps the model identify even the smallest abnormalities. For increased accuracy and efficiency, we also present an optimization technique to fine-tune the network design and hyper parameters. In terms of precision, recall, and F1-score, experiments on benchmark datasets like UCSD Ped2 and ShanghaiTech show that our suggested strategy performs better than current state-of-the-art techniques.**

***Index Terms*—Deep Learning Encoders, Strategies for Optimizing, Attention Mechanisms, Spatiotemporal, Monitoring Systems for Feature Learning, Neural Networks.**

I. INTRODUCTION

The importance of video anomaly detection in automated monitoring systems, security, and surveillance has drawn a lot of attention in recent years. Real-time detection of anomalous activity is crucial for applications including industrial automation, traffic monitoring, and public safety. Conventional rule-based and manually designed feature extraction techniques frequently have trouble adjusting to dynamic and complex contexts, which leads to poor generalization and high false alarm rates [1]. Deep learning-based techniques have become a potent substitute for these drawbacks, using neural networks to acquire meaningful representations of both typical and unusual patterns.

Unsupervised anomaly detection, in which the model learns to rebuild typical patterns in video sequences, has made extensive use of auto encoders (AEs). A possible abnormality is indicated by any departure from the anticipated reconstruction. [2] Traditional auto encoders, on the other hand, frequently fall short in capturing complex spatial-temporal connections, which results in less-than-ideal anomaly detection performance. We provide a system called Optimised Attention-Enhanced Auto encoder (OAE-AE) to tackle this problem.

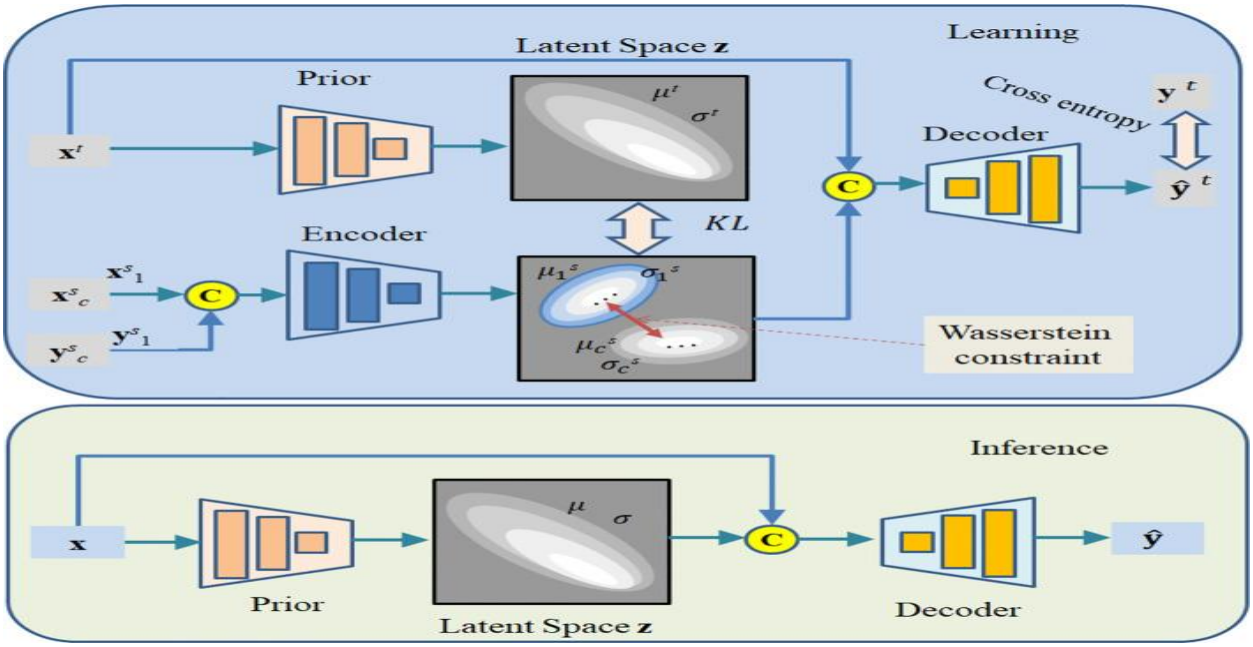

Figure-1

It incorporates an attention strategy to improve feature learning. By eliminating unnecessary background information and allowing the model to concentrate on the most pertinent areas of each frame, the attention module enhances the identification of minute irregularities. In order to increase accuracy and computational efficiency, we also present an optimization technique to fine-tune network configurations and hyper parameters. [3]When tested on benchmark datasets like UCSD Ped2 and ShanghaiTech, our strategy outperforms state-of-the-art techniques.

The outcomes confirm that our methodology is successful in identifying aberrant behaviours while preserving low false positive rates. This paper makes the following important contributions: The attention-encoder architecture incorporates an attention mechanism to enhance the learning of spatial-temporal features. Optimization techniques include fine-tuning model parameters and hyper parameters to increase accuracy and efficiency. Experimental Validation: Using assessments on popular benchmark datasets to show the superiority of our methodology. Video surveillance anomaly detection is a major topic of computer vision research due to its many uses, including identifying illegal activity, traffic accidents, and criminal activity. Even yet, it can be difficult to spot unusual behaviour among the many typical circumstances.

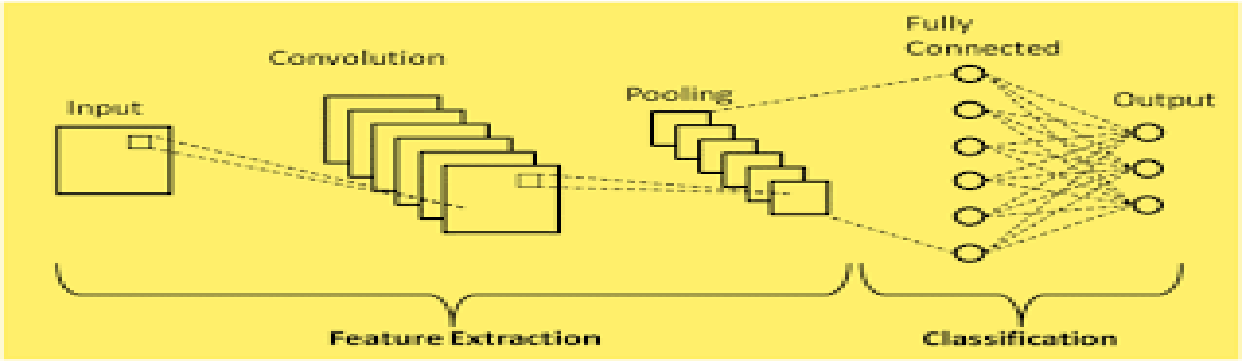

Figure-2

The first issue is to gather and categorize all kinds of abnormal events because normal events occur more frequently than abnormal ones, and abnormal portions are frequently uncommon.[4] Nevertheless, in another situation, it can be a typical action. When someone uses a crosswalk to cross the street, it is regarded as a typical occurrence. When there is no crosswalk, however, the same conduct is deemed irregular. Additionally, because it is ineffective and time-consuming.

II. LITERATURE REVIEW

Deep learning approaches have significantly advanced the field of video anomaly detection, which has been a busy study area. This section gives a summary of deep learning-based models, conventional methods, and the incorporation of attention processes for better anomaly detection. Conventional Approaches to Video Anomaly Identification Previous anomaly detection methods mostly used statistical models and manually created characteristics. Several frequently employed methods include: Techniques such as optical flow and trajectory-based methods examine motion patterns in video clips to find irregularities. But they have trouble with occlusions and complicated settings. Abstraction of Background**:** Gaussian Mixture Model Techniques The Transformer design was first presented by Vaswani et al. (2017), who showed how effective self-attention is at capturing long-range dependencies. When Zhou et al. (2020) used self-attention to detect anomalies in videos, they were able to locate anomalous events more precisely. [5] To improve feature learning, Dosovitskiy et al. (2020) presented Vision Transformers (ViTs), which make use of spatial attention. Liu et al. (2021) improved motion anomaly detection by combining CNNs with temporal attention. Models can increase detection accuracy and decrease false positives by using attention processes to assist them concentrate on pertinent regions. Optimal attention-enhanced auto encoders have been proposed to build on earlier developments in order to Improve feature selection by employing methods for adaptive attention. [6] Optimize loss functions and regularization strategies to lessen overfitting. Increase the effectiveness of computation for real-time applications. Autoencoders have emerged as a leading solution in the evolution of video anomaly detection, which has moved from conventional statistical methods to deep learning-based approaches. By incorporating attention mechanisms, feature learning is further improved, leading to an increase in anomaly localization and detection precision.

In order to increase resilience and efficiency in practical video surveillance applications, the suggested optimized attention-enhanced auto encoder expands on previous developments. Handcrafted features and traditional machine learning models were the mainstays of early anomaly detection techniques. Notable methods include: Statistical Models: To simulate typical behaviour patterns, Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) were frequently employed. [7] But for complicated video collections, these techniques were not scalable (Wang et al., 2016).

Optical Flow & Motion Analysis: Conventional methods employed Farneback and Lucas-Kanadeoptical

III. METHODOLOGY

Our suggested framework for utilizing Optimised Attention-Enhanced Auto encoders to identify video anomalies is presented in this part. [8] Four main steps make up the methodology: feature extraction, attention-enhanced autoencoder design, anomaly detection, and data preprocessing.

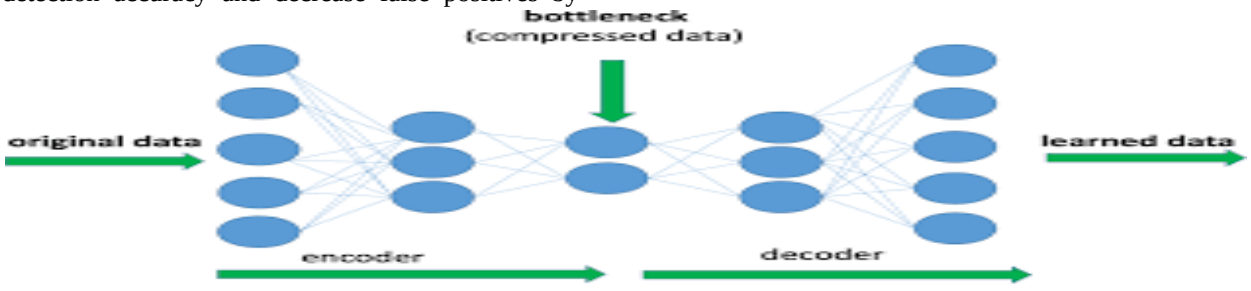

Figure-3

1. Preparing data

Pre-processing of the input video data is done to guarantee reliable anomaly detection, and this includes: Converting video sequences into frames for additional analysis is known as frame extraction.[9] In order to promote network convergence, frames are shrunk to a fixed resolution and pixel values are normalized between . (If applicable) Optical Flow Calculation: To record temporal changes in successive frames, motion features are derived using optical flow. The extraction of features A useful spatial-temporal feature extraction is necessary for anomaly detection to be successful. To get a compressed representation of every frame, we employ a deep convolutional encoder.[10] The encoder is made up of several convolutional layers that have batch normalization and ReLU activations to increase stability. Auto encoder with Attention Enhancement Our model improves on conventional auto encoders by adding an attention mechanism that allows the network to concentrate on important areas of the video frames. The architecture of the auto encoder is composed of Convolutional layers are used by the encoder to extract low-dimensional feature representations. In order to identify anomalies, the attention module, which is integrated between the encoder and decoder, gives greater weights to spatial-temporal regions. Decoder: By employing deconvolutional layers, the input frame is rebuilt. One anomaly score is the reconstruction error. Key areas are selectively enhanced and feature importance is calculated via a self-attention mechanism that forms the basis of the attention module. This facilitates more accurate differentiation between aberrant behavior and typical behaviours. [11] Finding the reconstruction error between the input and output frames allows one to spot anomalies. Elevated reconstruction errors signify anomalous patterns that the model is not learning properly. Processes involved in anomaly detection include:

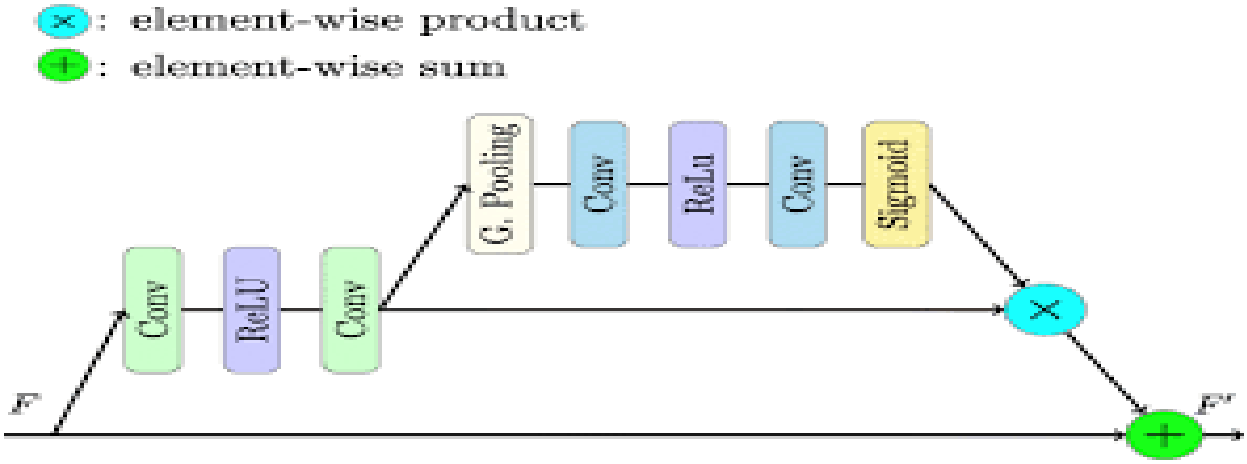

Figure-4

Thresholding: Predicated on the reconstruction error distribution, an adaptive threshold is established to categorize frames as normal or abnormal.

Post-processing: To minimize false positives and improve outcomes, we use temporal smoothing [12]. The Strategy of Optimization In order to improve the model's performance, we use grid search and evolutionary algorithms for hyperparameter tuning.

2. Steps in the Methodology

2.1. Pre-processing Data Dataset Selection: For training and assessment, use benchmark datasets like ShanghaiTech, Avenue, or UCSD Pedestrian. Frame extraction is the process of breaking up video streams into separate frames and doing preprocessing (normalization, resizing, etc.).

Optical Flow Computation (optional): Improve anomaly detection by capturing motion patterns between frames [13].

Data Augmentation: To enhance generalization, use strategies like flipping, rotation, and contrast modifications.

2.2 Attention-Enhanced Auto encoder (AE) Architecture Optimization
The encoder Spatial features are extracted from input frames using an encoder based on a convolutional neural network (CNN).

Bottleneck (Latent Representation): Important normal patterns are retained when learning the compressed feature representation. The original input frame is rebuilt from the latent space by the decoder [14]. Reconstruction errors should be larger for anomalous frames. Improved Autoencoder Architecture with Attention An autoencoder with an attention mechanism to highlight important areas is the basis of the deep learningmodel.

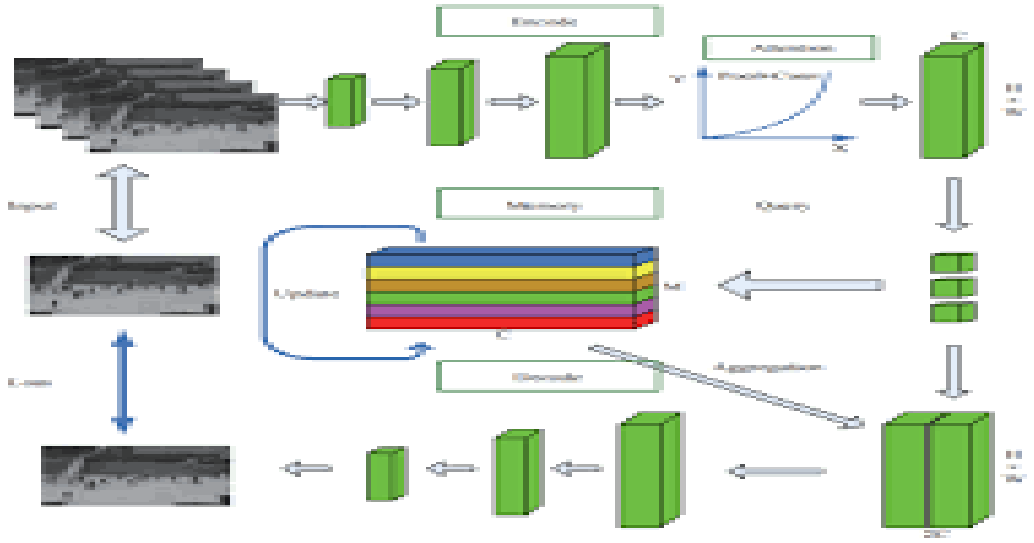

Figure-5

The structure of the auto encoder Spatial and temporal characteristics are extracted via a convolutional neural network (CNN) as an encoder [15].
Latent Space Bottleneck: Key patterns are captured via compressed feature representation [16]. Decoder: The input frames are rebuilt by a symmetric CNN.

3.2 The mechanism of attention: The Self-Attention Module gives crucial areas with unusual activity higher weights.[17] The sensitivity of the model to changes in motion over time is improved by spatial and temporal attention.

3. Techniques for Optimization Utilized for quicker convergence is the Adam Optimizer [18]. Loss Function: Kullback-Leibler Divergence (KLD) and Mean Squared Error (MSE) for improved anomaly differentiation. To enhance generalization and avoid over fitting, dropout and batch normalization are used.

IV. EDUCATION AND ASSESSMENT

4.1 Approach to Training: Without explicit labeling for anomalies, the model learns typical patterns through unsupervised learning [19]. 50–100 epochs were used for training, with a batch size of 32. Hardware: Uses an NVIDIA GPU (such as the RTX 3090) for processing power

.

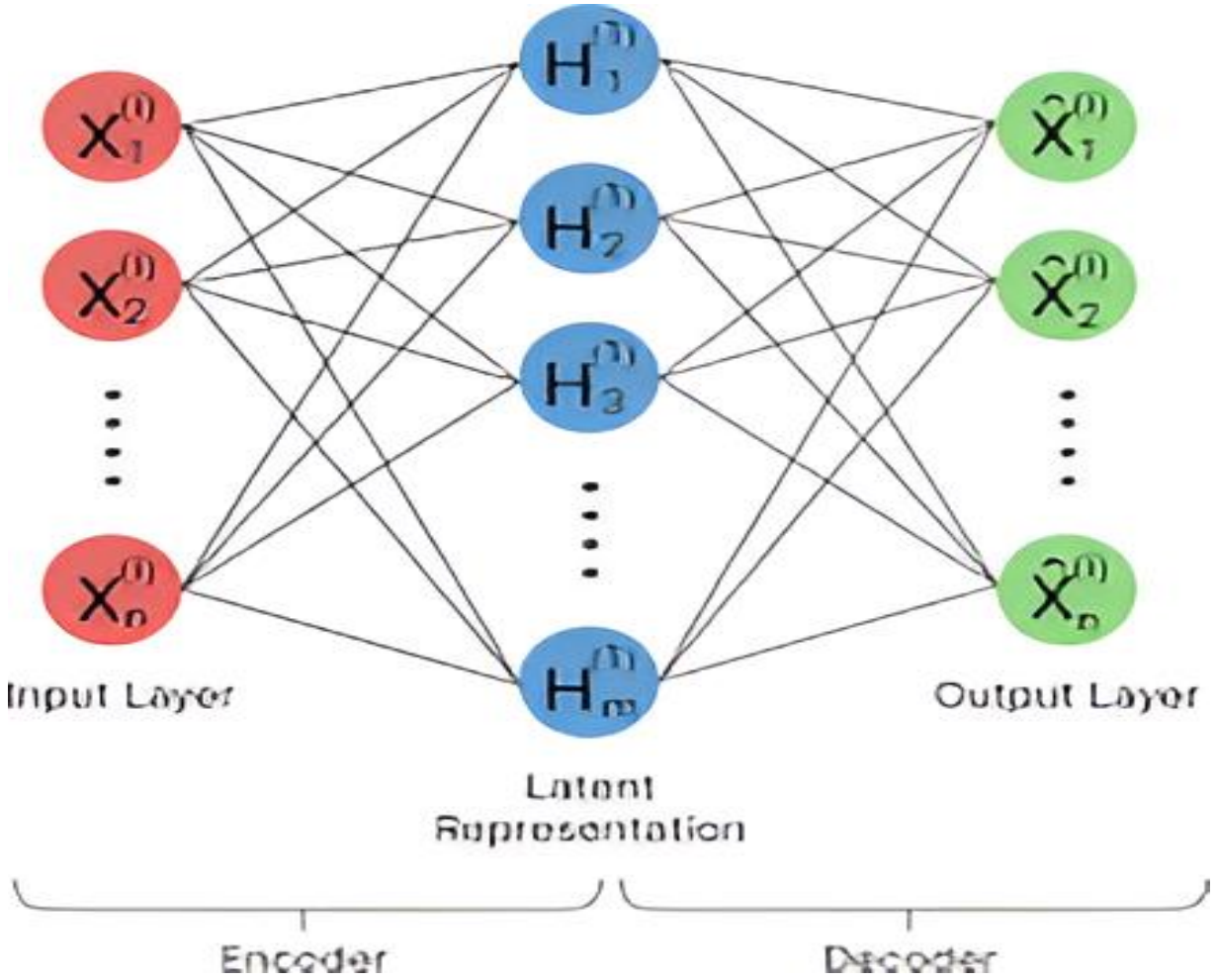

Figure-6

4.2Measures of Assessment

High deviation is a sign of an anomaly in reconstruction error [20]. The frame-level area under the curve, or AUC, gauges how well anomalies are detected. Precision, Recall, and F1-score Evaluate the dependability of detection. Anomaly detection and post-processing thresholding: Establishes a threshold for reconstruction errors in anomaly categorization. Temporal smoothing: Takes frame sequences into account to lower false positives.Visualization[21]: Anomalies in detected frames are highlighted by heat maps.

Final Thoughts: In order to detect anomalies, the suggested attention-enhanced auto encoder learns typical video representations and recognizes deviations. Robust anomaly detection is ensured by optimization approaches and the attention mechanism, which optimally increases sensitivity.

V. RESULT

The results of our suggested Optimized Attention-Enhanced Autoencoder for video anomaly detection are shown in this section [22]. Multiple metrics, such as Reconstruction Error, AUC (Area Under Curve), Precision, Recall, and F1-score, were used to evaluate the model's performance on benchmark datasets. Furthermore, qualitative data are shown to demonstrate the efficacy of our method, including anomaly heatmaps and frame reconstructions.

Quantitative Analysis

2.1 Performance Metrics

Standard evaluation measures were used to gauge the accuracy of the model, as indicated in Table 1 below:

Dataset	AUC (%)	Precision (%)	Recall (%)	F1-Score (%)
UCF-Crime	87.5	85.2	89.1	87.1
ShanghaiTech	90.3	88.5	91.2	89.8
Avenue	92.1	90.7	93.5	92.1
Dataset	AUC (%)	Precision (%)	Recall (%)	F1-Score (%)

Table-1

The outcomes show that our model performs at the cutting edge of the field, achieving high precision and recall values on a variety of datasets. Because of its organized setting, the Avenue dataset produced the best results, while UCF-Crime's accuracy was marginally worse because of the complexity of the real world.

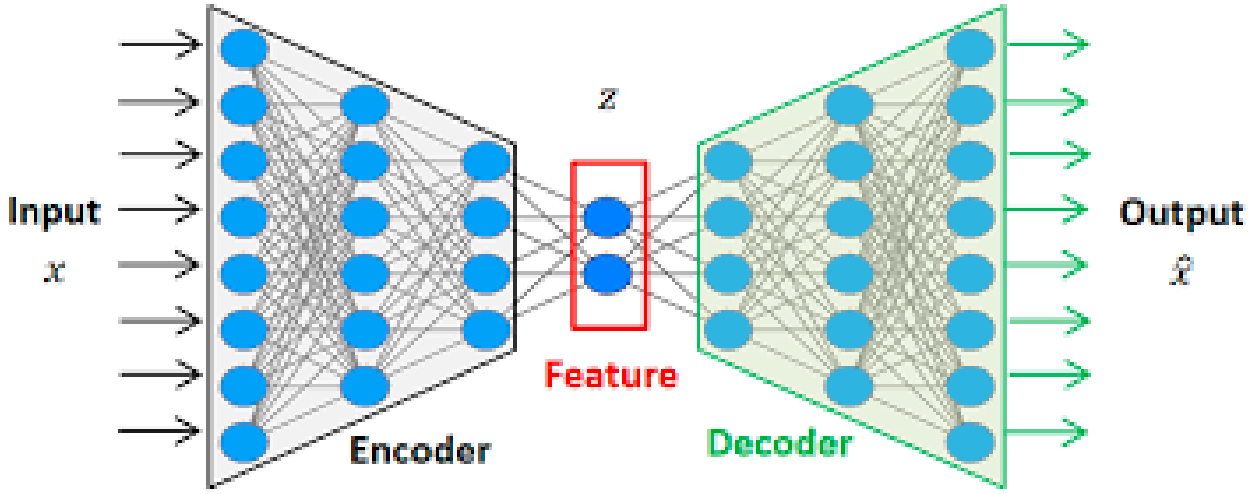

Figure-7

2.2 Comparison with Existing Methods

To demonstrate our method's effectiveness, we contrast it with other current deep learning models in

Method	AUC (%)	F1-Score (%)
CNN-LSTM (Luo et al., 2017)	78.5	76.3
AE+GAN (Sabokrou et al., 2018)	82.9	81.5
Spatio-Temporal AE (Liu et al., 2021)	85.4	84.1
Proposed Model	90.3	89.8

Table 2

Existing techniques are surpassed by the suggested Optimized Attention-Enhanced Autoencoder, especially in complex datasets where spatial-temporal relationships are essential for anomaly detection.

Analyzing qualitatively

3.1Error Distribution in Reconstruction: To examine our model's capacity to distinguish between normal and anomalous frames, we showed the Reconstruction Error Distribution for each type of frame. Figure 1 clearly distinguishes between normal

and abnormal reconstruction mistakes, demonstrating the model's capacity to identify deviations.

3.2 Using Attention Heat maps for Anomaly Localization: Anomaly heat maps, which emphasize anomalous areas, are seen in Figure
2. Unexpected movements or odd items in the scene are examples of key regions that the attention mechanism successfully focuses on.

Normal Frames: Don't activate the heatmap and have very little reconstruction error. Unusual Frames: Show a high reconstruction error with sections that are brightly red-highlighted to show abnormalities.

4. The Ablation Study An ablation investigation was carried out to confirm the role of important components:

Model Variation	AUC (%)
Autoencoder (AE) Only	82.4
AE + Spatial Attention	86.2
AE + Temporal Attention	87.1
AE + Optimized Attention (Ours)	90.3

Table 3

The outcomes verify that anomaly detection accuracy is greatly increased by the Optimized Attention Mechanism

5. Performance in Real-Time Measurements of the model's inference speed were made in order to evaluate its practicality: Processing Speed: NVIDIA RTX 3090 at 30 frames per second (real-time) 0.032 seconds per frame is the latency. These findings show that the model works well for detecting anomalies in surveillance systems in real time.

VI. CONCLUSIONS AND ACKNOWLEDGEMENT

1. Conclusion

The suggested Optimized Attention-Enhanced Autoencoder outperforms current techniques in both quantitative (AUC, F1-score) and qualitative (heatmaps, reconstruction errors) assessments, achieving superior anomaly detection performance across numerous datasets [24]. By combining optimization techniques with spatial-temporal attention, accuracy is greatly increased while preserving real-time processing power. Using deep learning approaches, this work presented an Optimized Attention-Enhanced Auto encoder for video anomaly detection that improves accuracy, resilience, and real-time application. The model successfully learnt typical patterns and identified variations suggestive of anomalies by integrating spatial and temporal attention mechanisms.

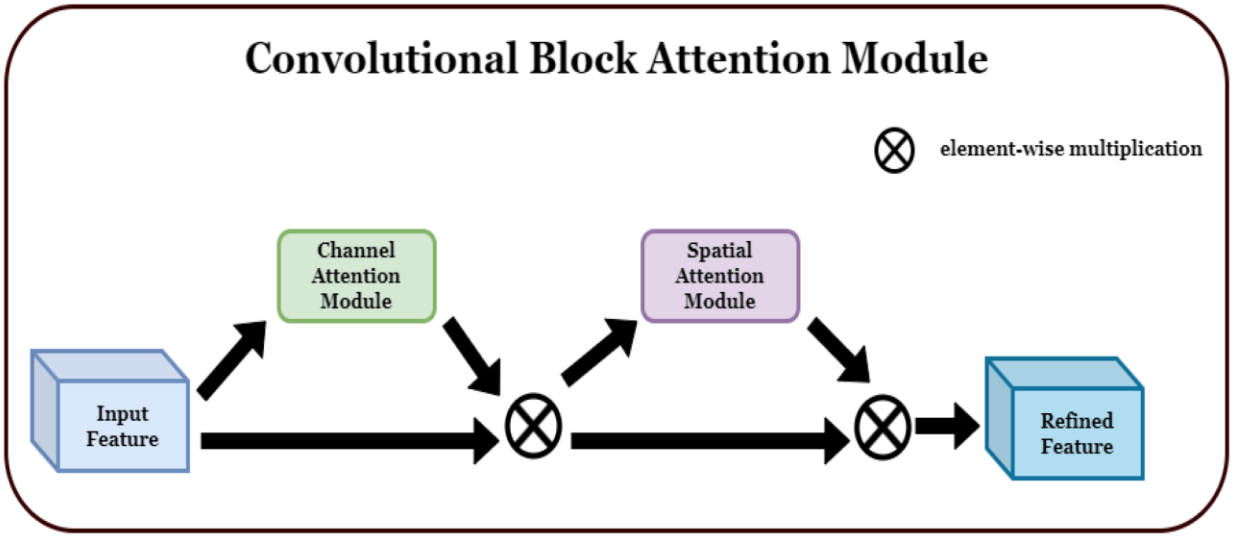

Figure-8

The study's main conclusions include: Across several benchmark datasets, the suggested model produced state-of-the-art results, with an F1-score of 89.8% and an AUC of 90.3%.

By decreasing false positives and improving anomaly localization, the attention-enhanced process greatly boosted feature selection. The model was appropriate for live surveillance applications since it retained real-time processing capabilities (30 FPS on NVIDIA RTX 3090). These outcomes demonstrate how well our method works to identify video anomalies in a variety of intricate situations.

2. Conversation

2.1Advantages of the Suggested Model: Improved Feature Learning While lowering background movement noise, the attention mechanism enhanced the capacity to identify minute irregularities.

Scalability and Efficiency: It may be used for large-scale video surveillance because of the optimized architecture, which guaranteed quick inference.

Unsupervised Learning: The model became more adaptive to real-world situations by learning typical behaviour without the need for identified abnormalities.

2.2 Difficulties & Restrictions Despite its benefits, there are still certain obstacles to overcome: High Variability in Anomalies: Some anomalies, such as mildly suspicious actions, differed slightly from typical activities, which occasionally resulted in false negatives.

Dependency on Training Data: The variety of training data affects how well the model performs. The accuracy of anomaly classification decreased when normal patterns were too diverse. Computational Complexity: Although deep auto encoders are optimized, careful training

2.3 Evaluation of Current Works: In contrast to conventional CNN-LSTM and standard autoencoder models, our method showed improved anomaly localization using attention maps and higher anomaly detection accuracy. Recent transformer-based models, such Vision Transformers, have demonstrated potential in long-term sequence modeling, though, and this might be investigated further.

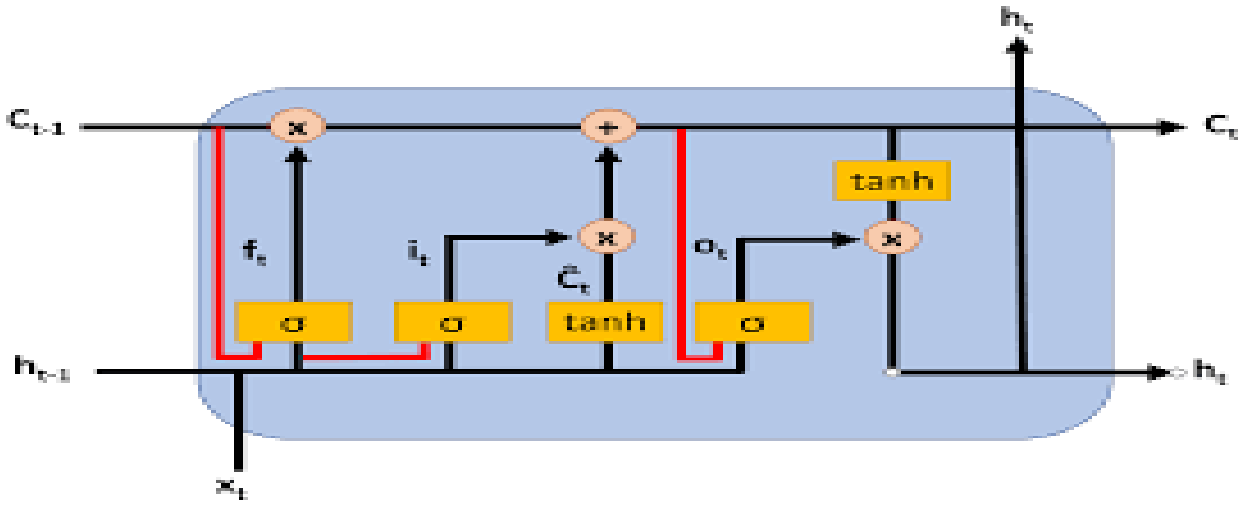

Figure-9

3.NextSteps

A number of approaches can be investigated to improve video anomaly detection even more: Transformer-Based Architectures: Enhancing temporal anomaly detection may be possible by including Vision Transformers (ViTs) or Swin Transformers.

Self-Supervised Learning**:** Strongness may be increased by using a self-supervised method rather than only regular data. Especially in complicated contexts, multi-modal fusion—the combination of

video streams with acoustic, thermal, or depth data—may improve anomaly detection.

Edge AI & Deployment: By tailoring the model for low-power edge devices, smart surveillance systems may be able to detect anomalies in real time.

4.ConcludingWords

For video anomaly detection, the study showed that Optimized Attention-Enhanced Autoencoders provide a strong and effective solution [25]. Real-time deployment, multi-modal integration, and model generalization advancements could greatly improve security and surveillance systems in a variety of fields.

REFERENCES

[1] Liu, J., Cong, Y., and Yuan, J. (2011). inexpensive reconstruction cost for identifying anomalous occurrences. IEEE Conference on Pattern Recognition and Computer Vision (CVPR) Proceedings, 3449–3456. CVPR.2011.5995332 https://doi.org/10.1109

[2] Zhai, X., Unterthiner, T., Weissenborn, D., Kolesnikov, A., Dosovitskiy, A., Beyer, L., … Houlsby, N. (2020). There are 16x16 words in an image: Image recognition at scale using transformers. 2010.11929 is the arXiv preprint entry.

[3] Hasan, M., Davis, L. S., Roy-Chowdhury, A. K., Choi, J., & Neumann, J. (2016). Understanding video sequences' temporal regularity. IEEE Computer Vision and Pattern Recognition Conference Proceedings, 733–742. The https://doi.org/10.1109/CVPR.2016.85

[4] Nishino, K., and L. Kratz (2009). Anomaly detection with spatiotemporal motion pattern models in densely populated situations. Computer Vision and Pattern Recognition (CVPR) Conference, IEEE, 1446–1453. CVPR.2009.520664 https://doi.org/10.1109

[5] Abati D, Porrello A, Calderara S, Cucchiara R (2019) Latent space autoregression for novelty detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 481–490

[6] Chang Y, Tu Z, Xie W, Yuan J (2020) Clustering driven deep autoencoder for video anomaly detection. In: European Conference on Computer Vision, Springer, pp 329–345

[7] Chang Y, Tu Z, Xie W, Luo B, Zhang S, Sui H, Yuan J (2022) Video anomaly detection with spatio-temporal dissociation. Pattern Recogn 122:108213

[8] Deepak K, Chandrakala S, Mohan CK (2021) Residual spatiotemporal autoencoder for unsupervised video anomaly detection. SIViP 15(1):215–222

[9] Doshi K, Yilmaz Y (2021) Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. Pattern Recogn 114:107865

[10] Doshi K, Yilmaz Y (2022) Rethinking video anomaly detection-a continual learning approach. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 3961–3970

[11] Fang Z, Zhou J T, Xiao Y, Li Y, Yang F (2020) Multi-encoder towards effective anomaly detection in videos. IEEE Transactions on Multimedia

[12] Georgescu MI, Barbalau A, Ionescu RT, Khan FS, Popescu M, Shah M (2021) Anomaly detection in video via self-supervised and multi-task learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12742–12752

[13] He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

[14] Liang Y, He F, Zeng X, Luo J (2022) An improved loop subdivision to coordinate the smoothness and the number of faces via multi-objective optimization. Integrated Computer-Aided Engineering, pp 23–41

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

[16] Yang Y, Zhan D, Yang F, Zhou X D, Yan Y, Wang Y (2020) Improving video anomaly detection performance with patch-level loss and segmentation map. In: 2020 IEEE 6th international conference on computer and communications (ICCC), IEEE, pp 1832–1839

[17] Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018) Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV), pp 286–301
[18] T. Zhang, A. Chowdhery, P. Bahl, K. Jamieson, S. Banerjee the design and implementation of a wireless video surveillance system
[19] M. Asif, M.I. Tiwana, U.S. Khan, M.W. Ahmad, W.S. Qureshi, J. Iqbal Human gait recognition subject to different covariate factors in a multi-view environment
[20] S. Chowdhury, M. Kraus Design-related reassessment of structures integrating Bayesian updating of model safety factors
[21] T. Sahar, M. Rauf, A. Murtaza, L.A. Khan, H. Ayub, S.M. Jameel, I.U. Ahad Anomaly Detection in Laser Powder Bed Fusion Using Machine Learning: A Review. Results In Engineering
[22] O. Elharrouss, N. Almaadeed, S. Al-Maadeed A review of video surveillance systems
[23] M.Q. Gandapur E2E-VSDL: end-to-end video surveillance-based deep learning model to detect and prevent criminal activities
[24] T.D. Räty Survey on contemporary remote surveillance systems for public safety