Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Cross Attention Based Feature Fusion Network for Robust Anomaly Detection in Surveillance Videos

Dipak Ramoliya[1], Amit Ganatra[2]

[1,2]Department of Computer Science & Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Changa, Anand, India
[2]Department of Computer Science and Engineering, Faculty of Engineering and Technology, Parul University (PU), Waghodia, Vadodara, India.

E-mail: [1]dipakramoliya.ce@charusat.ac.in, [2]ganatraamitp@gmail.com

Abstract

For enhancing public safety, a surveillance system is essential. Specifically, video surveillance is the most popular way to maintain safety in public and private areas. The detection and recognition of abnormal activity is difficult due to a complex environment, video quality, and varying noise levels. Addressing the challenges of accuracy and video processing, the proposed study uses a cross-attention network with feature fusion to improve the recognition of abnormal activity in complex scenarios. Cross-attention helps to capture contextual information from different videos. The proposed model combines an innovative method of cross attention and feed-forward attention with latent space representation-based fusion, aiming to improve accuracy. The simulation of the study uses two benchmark datasets, UCF and UCSD and achieves remarkable performance with 97.1 % and 91.31 % accuracy. A simulation study has also demonstrated a comparative analysis with different convolution and attention networks for anomaly detection. This study proposes an effective video processing scheme with wide practical potential. The study also provides a new perspective and methodological basis for future research and applications in related fields.

Keywords: Anomaly Detection, Computer Vision, Video Surveillance, Multimodal Learning, Attention Network, Feature Fusion.

1. Introduction

With the growth of advanced data, information with video and images is used to improve human safety and life. Advanced AI is integrated into domains such as autonomous vehicles, medical diagnostics, and safety monitoring, where video processing is essential for analysis and the building of action-driven models [1]. The widespread use of surveillance cameras in public places like streets, parks, banks, intersections, shopping malls, schools, and public properties has greatly improved public safety. However, law enforcement agencies' ability to monitor these areas hasn't kept up, resulting in a large number of cameras being watched by very few people [2,3]. The security monitoring system needs to identify potential illegal activities or dangerous situations in real time. There is an unworkable ratio of cameras to human supervision. This problem calls for the creation of smart computer vision algorithms

DOI: https://doi.org/10.36548/jiip.2025.3.006
Received: 07.07.2025, received in revised form: 06.08.2025, accepted: 20.08.2025, published: 29.08.2025

that can automatically detect unusual events such as traffic accidents, crimes, and other suspicious activities [4].

Anomaly detection is the process of identifying rare and abnormal activities from normal patterns in real time. Traditional methods mainly focus on finding specific anomalies like violence or traffic incidents [5]. However, these approaches are not flexible enough because real-world anomalies can vary greatly and are hard to predict. That's why an effective anomaly detection system needs to work with proper supervision, using learned patterns instead of predefined event types [6,7].

Sparse-coding-based methods have shown great success by creating a dictionary of normal events and identifying anomalies as differences from this learned representation. Many computer vision methods struggle with adapting to changing environments, such as variations in visibility and different sizes of crowd density, resulting in a high number of false alarms [8]. To address these issues, we require a robust and flexible anomaly detection system that can manage complex real-world scenarios such as different areas of context (corner or middle of the frame), different lighting angles, and different human movement scenarios [9]. Human or animal movements cannot be similar in every situation under the same class label.

Despite the requirement, automated surveillance systems with computer vision and deep learning are struggling with a massive amount of data. High magnitude data requires powerful computation to handle high-dimensional data streams [10, 11]. The most complicated challenge is the variation from scene to scene. No identical or look-alike data is possible from surveillance camera footage. A large variation is present in every second video, even though the type (label) of abnormality is the same. Because of a rule-based structure, machine learning models generally fail to find useful features from such complex data [12]. In the above context, the encoder-based architecture finds compatible results compared to traditional deep learning and machine learning models.

The proposed study aims to develop anomaly detection using video data. It proposes a recognition algorithm based on a multimodal attention model to improve the accuracy and efficiency of anomaly detection. The research objective of the study includes designing and implementing an efficient vision transformer-based multimodal architecture to identify abnormal activity for video input. The major contributions of the study are: 1) The proposed study uses multi-stream feature extraction with a cross-attention encoder network, cross attention helps to improve generalisation ability. 2) Optimised feature fusion with latent space representation improves detection accuracy from complex video input.

2. Literature Review

Sultani et al. [13] introduce a deep learning approach to anomaly detection in surveillance videos from weakly labeled data. Rather than labeling anomalous segments manually, the authors suggest using a Multiple Instance Learning (MIL) scheme where whole videos are treated as labeled instances (normal or anomalous) and video segments receive anomaly scores. The Deep MIL Ranking Model learns automatically to identify anomalies by ranking video fragments according to their anomaly considering sparsity and temporal smoothness constraints to enhance localization precision. The paper also presents a large-scale dataset of 1,900 real-world surveillance videos for 13 categories of anomalies (e.g., fighting, burglary, accidents). Experiments demonstrate that their approach surpasses current methods in anomaly detection with reduced false alarm rates and better detection accuracy. The dataset

is also used as a benchmark for anomalous activity recognition, highlighting the difficulty in real-world surveillance due to high intra-class variability and untrimmed video sequences.

Doshi and Yilmaz [14] present an online anomaly detection model that balances training issues with real-time decision-making. Deep learning methods require huge amounts of labeled training data and cannot deal with shifting patterns. The authors propose a hybrid approach, fusing transfer learning with statistical k-nearest neighbor (kNN) decision-making, to enable anomaly detection with minimal supervision. The proposed method applies pre-trained neural networks to carry out feature extraction, combining motion information (optical flow), spatial information (bounding box coordinates), and visual information (object class probabilities). The system identifies anomalies in behavior through kNN-based anomaly detection without requiring pre-labeled anomalies. The work also introduces a new any-shot learning feature, which allows the system to continuously adapt to new nominal patterns from a few samples. Experiments on the benchmark datasets (UCSD, CUHK Avenue, Shanghai Tech) show that the proposed approach performs better than existing state-of-the-art models with significantly less training data for anomaly detection. Its real-time potential and the ease with which it adapts to new anomalies make it an implementable solution for surveillance video analysis.

Bhakat and Ramakrishnan [15] suggest an autoencoder-based framework for unsupervised anomaly detection from surveillance videos. The model is trained in a manner that reduces reconstruction loss on the assumption that normal frames are reconstructed with precision and anomalies are reconstructed with a higher error. The process is also aided by semi-supervised learning, where user ratings are utilised to enhance detection accuracy. Grad-CAM is also used to detect anomalous regions in frames, making model decision-making more interpretable. The authors compare their approach to three unrelated datasets: Avenue (CUHK dataset), Surveillance Office, and the novel Police dataset. The suggested model facilitates anomaly detection by identifying frames with high reconstruction errors, supplemented with thresholding, ranking, and a graph-theoretic formulation.

Authors Rohit Raja et al [16] discussed about the tough job of finding odd stuff in busy, ever-changing scenes. They stress the need for quick, hands-off fixes because people can't keep an eye on everything. The researchers group anomaly detection techniques into four main types: extracting key features, processing on the fly, machine learning models, and deep learning-based approaches. They weigh the pros and cons of each. The paper verifies various datasets and measures how well these model techniques perform. It demonstrates that convolutional neural networks (CNNs), autoencoders, and mixed models perform fairly well. The paper identifies some significant challenges, such as obstructions, overlapping objects, and changes in the environment.

Authors Kun Liu and Huadong Ma [17] tackle the problem of background bias in deep learning models that detect anomalies. The authors argue that most models rely on background information instead of spotting unusual behaviors. To fix this, they've added new labels to the biggest anomaly detection dataset, including exact time and location details. They've also developed a new way to measure how well models spot anomalies based on what's happening, not just the background. To reduce background bias, the authors suggest a model that focuses on unusual areas using a special loss function to make the model focus on these spots. They also use meta learning to help the model work better with limited training data and avoid overfitting. Their tests show that this approach reduces the impact of backgrounds and outperforms other methods on standard benchmarks.

Ali Khaleghi; and Mohammad Shahram Moin [18] present a new approach to identifying anomalous activity within security camera streams using advanced artificial intelligence. The approach consists of two elements: salient feature extraction and anomalous incident detection. To obtain these details, the system uses complex AI networks to analyze how things appear, how crowded areas are, how objects move, and how scenes are structured. This helps the AI distinguish between normal and strange actions. To spot unusual events, the system employs several different sorting tools and a group decision-making process to enhance its accuracy. The researchers tested their method on the UCSD dataset, and it performed better than current methods, with higher accuracy and fewer mistakes. The study highlights how helpful AI is for finding important details in videos. It also suggests ways to improve the system in the future, like pinpointing where strange events occur and providing better descriptions of what's happening.

Chen, D et al. [19] present a model for anomaly detection based on deep learning methods through bidirectional prediction networks. The method improves stability and accuracy through subnetworks for forward and backward prediction to detect the target frame, as compared to conventional models based on previous frames to predict the next frame. The core structure of the model is built upon the U-Net model, which presents a new loss function that considers both forward and backwards dependencies. To enhance the performance of anomaly detection, it adopts a sliding window strategy that focuses on foreground anomalies while efficiently suppressing background noise, Experimental results conducted on popular surveillance datasets show that this method outperforms conventional models, thus presenting an effective and efficient solution for handling anomalies.

Some important difficulties with literature reviews have been highlighted by the authors. Anomaly detection faces three main challenges, as shown in Figure 1. The most significant obstacle in anomaly identification is the ambiguity of the problem. Classifying anomalous activity is not as easy as it seems.

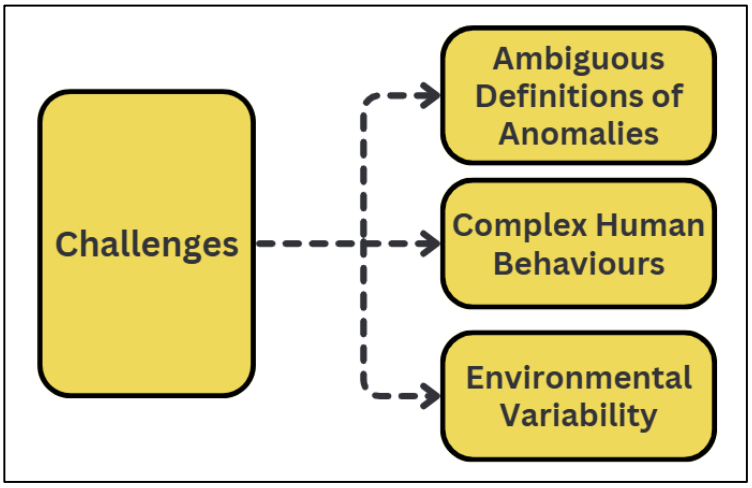

Figure 1. Different challenges in Anomaly Detection with Machine Learning

A comprehensive survey of machine learning techniques adopted in surveillance video anomaly detection has been discussed by the authors [20]. The article compares and contrasts the advantages, disadvantages, and usability of various techniques based on the classification of supervised, semi-supervised, and unsupervised learning techniques. Occlusions, real-time processing, environmental adaptability, and big data are the issues identified in the survey. It

also investigates evaluation metrics, datasets, and state-of-the-art learning models like CNNs, RNNs, autoencoders, and transfer learning-based models. The survey reflects the necessity of adopting hybrid models, efficient feature extraction, and real-time adaptability to enhance anomaly detection performance in challenging surveillance environments.

3. Proposed Work

This section discusses the principles of the Transformer and then discusses the proposed multimodal transformer for anomaly detection. First, the image is encoded, and in the second phase, a sequential decoder is used for context identification from videos [21]. The video consists of two data modalities: visuals and audio; both modalities have equal importance in identifying and understanding the context of the video for humans. Therefore, computational models need to consider this as well. The proposed study uses a cross-model multihead transformer to effectively and efficiently handle the complexities of abnormal activity in videos. The aim of the proposed study is to learn inter-model interaction to improve recognition accuracy.

The proposed study uses two non-aligned modalities as a sequence of them represented as α and β, which are formally represented as $X_a = \mathbb{R}^{T_\alpha * d_\alpha}$ and $X_\beta = \mathbb{R}^{T_\beta * d_\beta}$ Respectively. T represents feature length, and d represents feature dimensions. Keys, values and query are defined as $Q_\alpha = X_\alpha W Q_a, K_\beta = X_\beta W K_\beta$, and Values as $V_\beta = X_\beta W V_\beta$ The final cross model attention score can be calculated as equation 1 [22].

$$Y_\alpha = softmax\left(\frac{X_\alpha W_{Q\alpha} * X_\beta W_{K\beta}}{\sqrt{d_k}}\right) \times X_\beta W_{V\beta} \quad (1)$$

Attention score Att_c has the same length as the query of modality a and sequence representation scaled in $\sqrt{d_k}$ For the softmax activation function. Cross-modality attention uses the i-th modality of α to the j-th modality of β. This is how cross-modality attention will be performed. Final attention will be determined by the i-th row in softmax with V_β. Att_c Represents a single-head cord's attention score, while the proposed study uses multi-head cross attention with a residual neural network [23, 24].

The proposed model uses a residual network with cross attention to improve feature extraction. This hybrid approach helps to adapt the low-level feature sequence from input videos. The proposed study added a pointwise fully connected feed-forward network as a sub-layer in cross-model attention, as demonstrated in Figure 2. The proposed attention network helps to reduce dependency on self-attention. The proposed model performs adaptation of low-level features to preserve all necessary information from both modalities.

Multi-modality approach helps to identify the different modalities available in the input. The proposed study uses video and audio molality with a cross and feed-forward attention network, respectively. Proposed modality-based feature extraction helps to classify input with dual possibilities. Input can be truly classified with a single modality or both modalities, resulting in improved model performance with better learning parameters.

3.1 Self-Attention Variants in Multimodal Context

The attention mechanism is the core of encoder learning in transformers. Different attention strategies like multihead, multiple, long-range, hard/soft and many others, are available. The transformer strategy is directly aligned with modality and the amount of training data. The proposed study uses cross-attention to improve feature learning from input videos. The proposed study initially performs tokenising of the input video frames and selects an embedding space to represent the tokens. The embedding token space is highly flexible with many alternatives [25]. The proposed study uses forward attention features from token embedding with an appropriate label. From video input, a common tokenisation is to treat the non-overlapping, down-sampled windows over the video as tokens. Figure 2 illustrates the proposed multimodal transformer used in the study, where Oi represents the count of output labels available in the dataset, while ix stands for the parameter of the last input layer in the stack.

Our proposed method leverages a Multimodal Cross-Attention (MCA) framework for video classification by integrating visual and audio information. We hypothesize that cross-modal interactions between features enhance semantic understanding, which is crucial for fine-grained classification tasks in video content. The architecture consists of modality-specific encoders followed by a Cross-Attention Fusion Module, and a final Classification Head. Proposed study uniformly samples T_S frames from each video clip and extracts spatial features using a deep vision transformer; each frame is resized and normalised to conform to the input requirements of the chosen backbone. Optionally, features are extracted using a temporal encoder to retain spatiotemporal information [26].

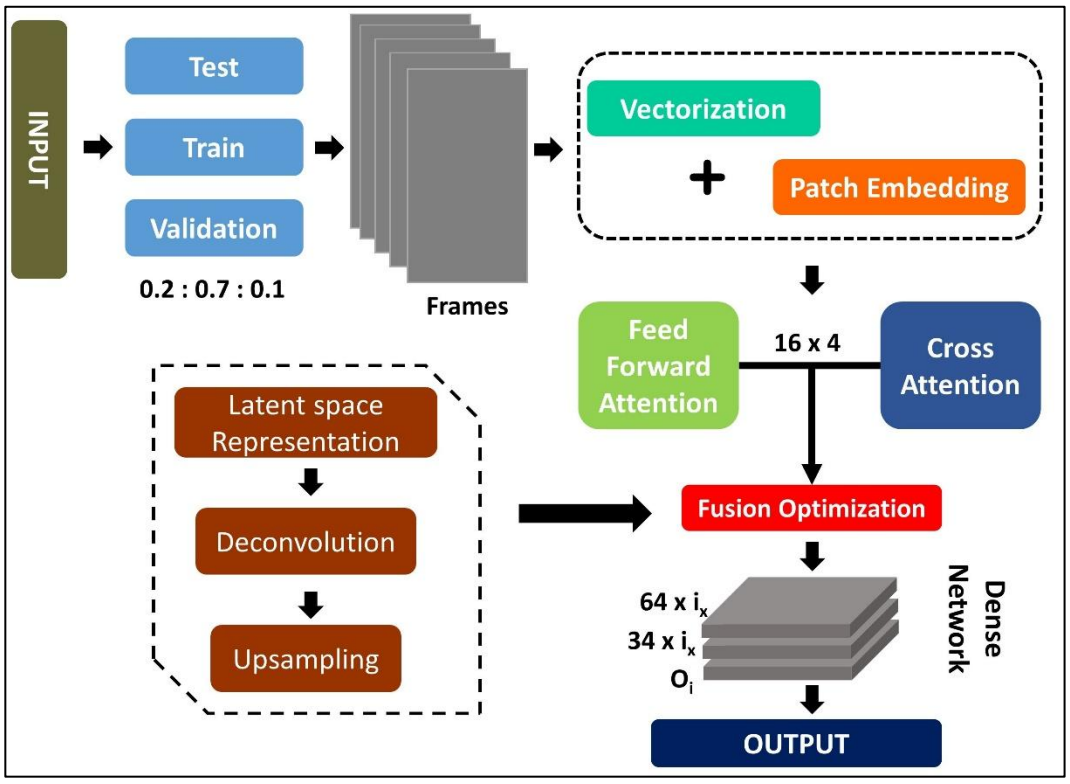

Figure 2. Proposed Multimodal Transformer-based Architecture

The raw audio track is extracted from each video and converted into a log-Mel spectrogram. We use a feed-forward attention mechanism to encode the spectrogram into a sequence of audio feature vectors. Temporal alignment with visual frames is performed to match the number of audio segments to the number of video frames, using interpolation if necessary [27].

3.2 Modality-Specific Encoders

Each modality is passed through a dedicated encoder, as visual and audio can be represented separately. Visual Encoder can be represented as $V \in R^{T*D_v}$, while Audio Encoder as $A \in R^{T*D_a}$ then projected into a shared dimensionality DD using linear transformations as equations 2, 3, where W_v and $W_a \in R^{D \times D_{v,a}}$ They are learned projection matrices.

$$V` = W_v * V \tag{2}$$

$$A` = W_a * A \tag{3}$$

3.3 Cross-Attention Fusion

The proposed study uses cross-attention modules to enable each modality to attend to features from the other modalities. The attention module uses a 24 x 1 input value, having 24fps video frames. The key idea is that the representation of each modality is enhanced by incorporating information from other modalities. Given query QQ, key KK, and value VV, the scaled dot-product attention as equation 4, as conventional attention score was calculated [28].

$$Attention_{Score} = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4}$$

Cross-attention network enable a more sophisticated and targeted fusion of information from multiple sources or across different dimensions of data. Rather than concatenation or simple averaging, cross-attention computes attention scores between two different sequences, while self-attention calculates attention scores within a single input sequence.

3.3.1 Attention Blocks

We construct multiple cross-attention heads. Visual-Audio Attention for visual queries attends to audio keys/values, as in equations 5 and 6, and Audio-Visual Attention for audio queries attends to visual keys/values.

$$V_{Att} = Cross_{Att}(Q = V` \; and \; K, V = A`) \tag{5}$$

$$A_{Att} = Cross_{Att}(Q = A` \; and \; K, V = V`) \tag{6}$$

Each updated modality feature is then passed through a residual connection and layer normalisation as formulated in equations 7 and 8.

$$V_{fused} = LayerNorm(V` + V_{Att}) \tag{7}$$

$$A_{fused} = LayerNorm(A` + A_{Att}) \tag{8}$$

These fused representations are optionally concatenated or further refined via a transformer encoder.

3.3.2 Temporal Aggregation

To produce a fixed-size video representation from the time-aligned fused features, we apply a temporal pooling strategy, as Attention Pooling over the time dimension can be formulated as equation 9 [29]. Bayesian inference-based fusion aggregation is used to improve the recognition of contextual or dependent information. While voting and averaging, the fusion may not be able to find a dependent feature over a given input video frame.

$$F_{Attention} = Aggregate\left(V_{fused}, A_{fused}\right) \tag{9}$$

3.3.3 Classification Head

The final feature vector FF is passed through a fully connected layer followed by softmax activation to yield the predicted class distribution, as calculated in equation (10), where c is the number of targeted classes.

$$Y = softmax(W_C F + b_c) \tag{10}$$

The cross-entropy loss is calculated during model training, formulated in equation 11, where y_i is the ground truth label (one-hot encoded) and $y^i\hat{y}_i$ is the predicted probability for class i.

$$C_{Loss} = -\sum_{i=1}^{C} y_i \log\left(\widehat{y_i}\right) \tag{11}$$

The final classification vector for the attention fusion will be passed to the fully connected neural network. A final four-layer dense network identifies the final class of the input video. Aggregation fusion using vector operations has been performed to combine the features of the audio and video modalities.

4. Dataset, Results and Discussion

The proposed study uses two benchmark datasets: UCF [30] Crime and UCSD [31]. The UCF Crime dataset has a size of 12.5 GB and includes various crime footage such as fighting, theft, and arson, providing video data in a more complex urban environment. The UCSD dataset also contains more realistic outdoor scenes, including the appearance of non-pedestrian entities such as cyclists and scooters. The properties of these benchmark datasets are demonstrated in Table 1.

Table 1. Properties of the Dataset Used in the Simulation of the Study

Dataset	Size (GB)	Total Frames	Training Frames	Validation Frames
UCF Crime	12.5	7000	5000	2000
UCSD	1.72	20000	16000	4000
ShanghaiTech	4.3	35000	25000	1000

Simulation of the proposed study has a training model with several hyperparameters listed in Table 2. The simulation of the proposed study is implemented on a system with an NVIDIA GeForce RTX 3080, 32GB RAM, and Python version 3.10. The simulation setup uses many standard parameter configurations, such as a train-test split of 0.2 and Adam as an optimiser in every hidden layer of the network. The authors have used a traditional approach for activation selection; every hidden layer uses ReLU, and the final layer uses softmax activation for multi-label classification. For the comparative analysis, the study uses standard evaluation metrics for accuracy, precision, and recall. Figure 3 demonstrates the accuracy and loss curve for the UCF Crime dataset, while Figure 4 demonstrates it for the UCSD dataset. The simulation of the proposed multimodal attention network achieves remarkable accuracy of 97.10% on the UCF Ccrime dataset and 91.31% accuracy on the UCSD dataset.

Table 2. Hyperparameter Settings Used for Model Training

Hyperparameters	Range	Best Value
Split	0.2, 0.3	0.2
Batch Size	8, 12, 16, 32	16
Optimizer	Adam / AdamW	Adam
Epoch	10 – 50	30
Dropout	0.2 – 0.5	0.3
Learning Rate	0.1 – 0.0001	0.01

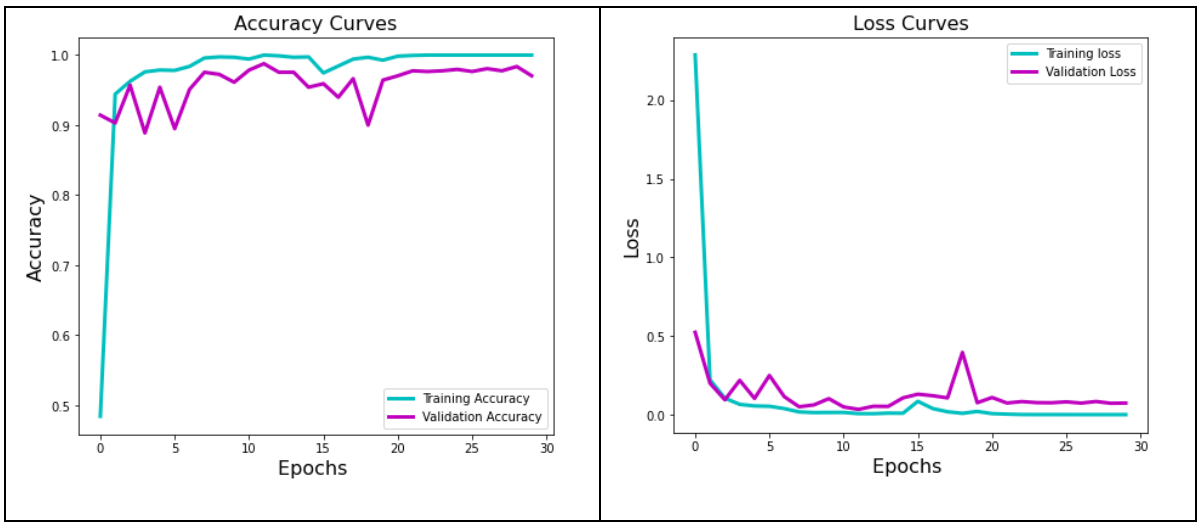

Figure 3. Accuracy and Loss Curve for the UCF Crime Dataset

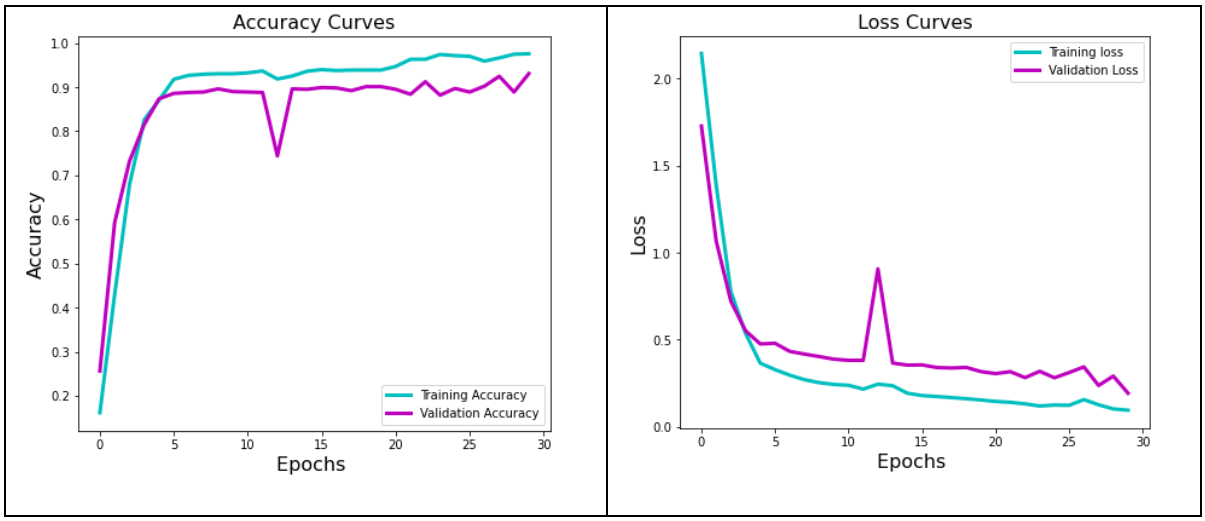

Figure 4. Accuracy and Loss Curve for the UCSD Dataset

Simulation of the proposed study also compares with SOTA deep learning models such as 3D CNN [32], RNN [33], LSTM [34] and GRU [35]. Anomaly detection uses video as a data modality, so a recurrent architecture is more suitable for extracting the features compared to a convolutional network. The output of the input data frame depends on the previous data frame and the possible condition of the next data frame. An architecture with timely backpropagation is more effective in finding relevant features. Figure 5 demonstrates a comparative analysis of the proposed study with the SOTA recurrent architecture having hyperparameters as listed in Table 2.

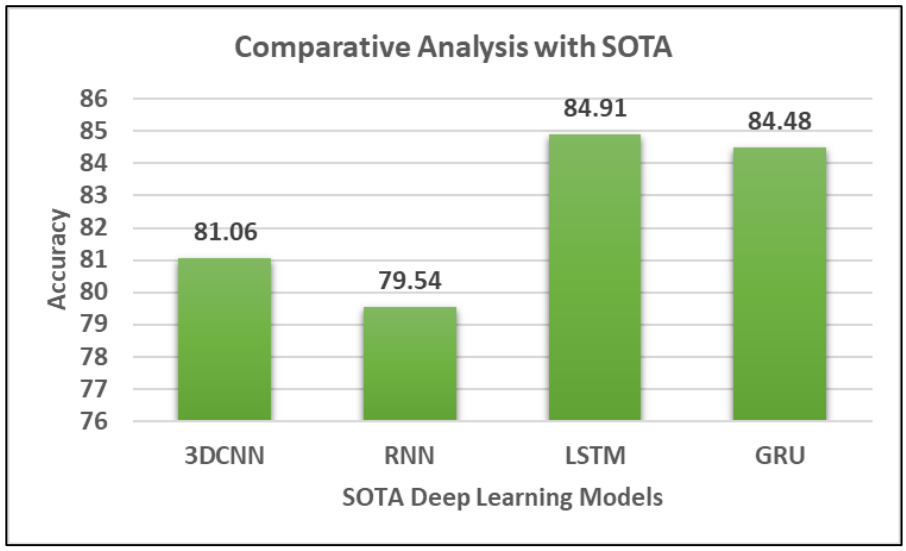

Figure 5. Comparative Analysis with SOTA

Simulation of the proposed multi-head attention also demonstrates the study with different possible encoders and attention mechanisms, such as multi-head attention, cross-attention and feed-forward attention. Table 4 presents a comparative study of the encoder model with different attention strategies. The results found in table 3, which have used multi-head attention [36], cross-attention [37] and feed-forward attention [38] have been simulated on the dual modality of input frames, to justify the performance of individual learning with a multimodal approach. Table 4 demonstrates the complexity comparison with performance in terms of AUC value [39]. Generally, FLOPs are used to measure the computational complexity

of the model. The floating-point operations required for each forward propagation, with the unit being a floating-point operation (FLOP) [40].

Table 3. Comparison of the Encoder Model with Different Attention Strategies

Attention	UCF Crime		UCSD	
	Accuracy	Avg. F1 Value	Accuracy	Avg. F1 Value
Multi-head Attention	94.24	94	81.01	80
Cross attention	93.55	93	81.55	81
FF Attention	79.62	79	70.47	70
Proposed Model	97.10	97	91.31	92

Table 4. Complexity Comparison with Performance in terms of AUC Value

Model	Parameter (M)	FLOPs	AUC (UCF Crime)	AUC (UCSD)
RNN	4.1	120	0.74	0.71
LSTM	6.7	120	0.81	0.83
GRU	5.2	140	0.79	0.80
Multihead Attention	4.9	150	0.83	0.84
Cross Attention	7.6	140	0.88	0.84
Proposed Model	8.2	180	0.93	0.95

The higher the AUC value, the better the anomaly detection performance of the model. To identify the sensitivity of the proposed model with different window sizes, the authors experimented with short and long window sizes, as illustrated in Figure 6. The model exhibited good performance with average window sizes of 16 and 24, while a higher value of 32 decreased the model's performance. The proposed study was simulated with a higher number of frames in the ShanghaiTech [41] dataset to analyse the effectiveness of the proposed study across different volumes of anomaly datasets. Figure 7 illustrates the comparative performance of the proposed network with three benchmark datasets.

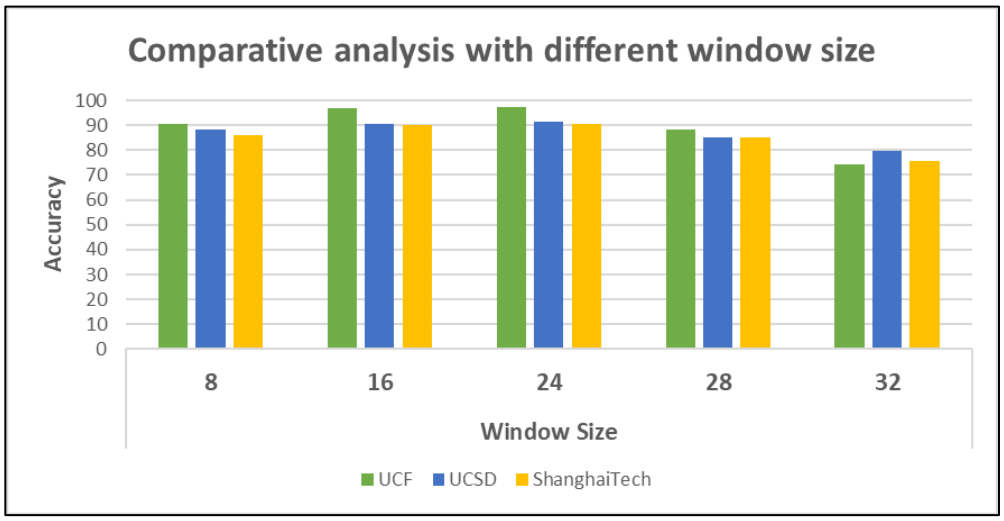

Figure 6. Comparative Study with Different Input Window Sizes

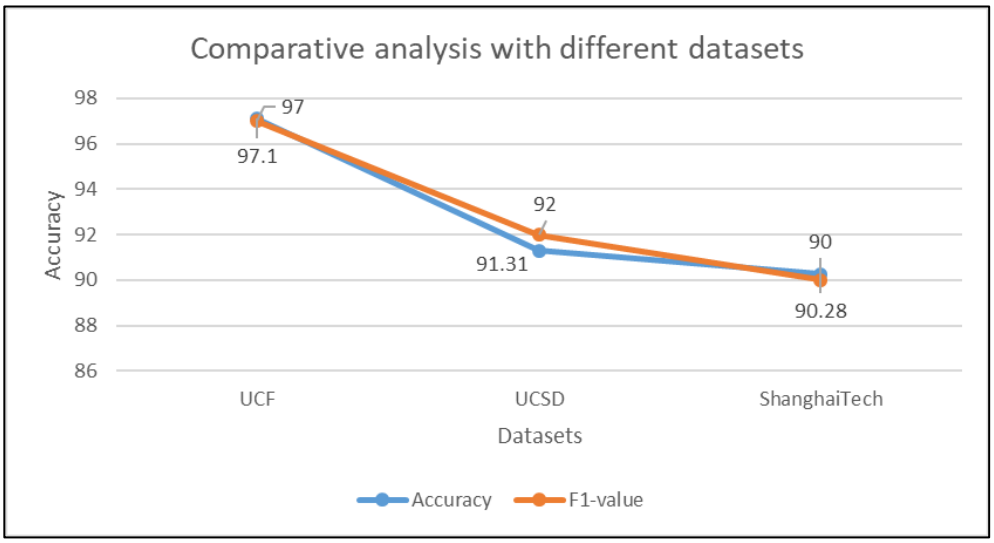

Figure 7. Comparative Analysis with Different Anomaly Benchmark Datasets

4.1 Discussion

The proposed study also demonstrates a comparative analysis with SOTA deep learning models such as 3D CNN, LSTM, and GRU. The simulation of the proposed study extends experiments with different benchmark datasets (illustrated in Figure 7). The experiments of the study also examine the effects of varying input window sizes, as illustrated in Figure 6. However, the proposed study may be extended with occlusion-based anomaly detection. The proposed study has used benchmark and preprocessed datasets, so no external filters or augmentations are required, which also helps to improve performance in a real scenario of abnormal activity.

5. Conclusion

Storage of surveillance footage is not enough to maintain security and rapid response support. Identification of abnormal activity from the surveillance camera is necessary. The proposed study used a multimodal attention model to recognize abnormal activity from video input. A multimodal approach enhances feature learning from two different input types: video and audio. The use of a multimodal approach improves the true positive rate in co-evolution learning for recognition and classification tasks. The proposed study uses a cross-attention network for both modalities, while final fusion will be performed with forward attention followed by classification layers. The simulation of the study is able to achieve 97.01% accuracy with the proposed multimodal cross-attention model. The proposed work was simulated using two benchmark datasets: UCF Crime and UCSD. Feature extension of the work can be possible with real-time integration with a surveillance system to improve rapid support in reducing loss of life from abnormal activities. Using dual modality with two different attentions results in a higher number of parameters and FLOPs count. The proposed study can be extended with the optimization of parameters, which can be more suitable for a portable device with fewer computations. A lighter version of the proposed model can be utilized for real-time surveillance because lower parameter recognition can result in faster processing.

References

[1] Handola, V., Banerjee, A. and Kumar, V., 2009. Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3),1-58.

[2] Pang, Guansong, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. "Deep learning for anomaly detection: A review." ACM computing surveys (CSUR) 54, no. 2 (2021): 1-38.

[3] Nassif, Ali Bou, Manar Abu Talib, Qassim Nasir, and Fatima Mohamad Dakalbab. "Machine learning for anomaly detection: A systematic review." Ieee Access 9 (2021): 78658-78700.

[4] Hao, Yanbin, Shuo Wang, Pei Cao, Xinjian Gao, Tong Xu, Jinmeng Wu, and Xiangnan He. "Attention in attention: Modeling context correlation for efficient video classification." IEEE Transactions on Circuits and Systems for Video Technology 32, no. 10 (2022): 7120-7132.

[5] Samariya, Durgesh, and Amit Thakkar. "A comprehensive survey of anomaly detection algorithms." Annals of Data Science 10, no. 3 (2023): 829-850.

[6] Zamanzadeh Darban, Zahra, Geoffrey I. Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi. "Deep learning for time series anomaly detection: A survey." ACM Computing Surveys 57, no. 1 (2024): 1-42.

[7] Ma, Xiaoxiao, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z. Sheng, Hui Xiong, and Leman Akoglu. "A comprehensive survey on graph anomaly detection with deep learning." IEEE transactions on knowledge and data engineering 35, no. 12 (2021): 12012-12038.

[8] Li, Zhong, Yuxuan Zhu, and Matthijs Van Leeuwen. "A survey on explainable anomaly detection." ACM Transactions on Knowledge Discovery from Data 18, no. 1 (2023): 1-54.

[9] Zhang, Ximiao, Min Xu, and Xiuzhuang Zhou. "Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 16699-16708. 2024.

[10] Liu, Zhikang, Yiming Zhou, Yuansheng Xu, and Zilei Wang. "Simplenet: A simple network for image anomaly detection and localization." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 20402-20411. 2023.

[11] Xu, Hongzuo, Guansong Pang, Yijie Wang, and Yongjun Wang. "Deep isolation forest for anomaly detection." IEEE Transactions on Knowledge and Data Engineering 35, no. 12 (2023): 12591-12604.

[12] Yan, Shen, Haidong Shao, Zhishan Min, Jiangji Peng, Baoping Cai, and Bin Liu. "FGDAE: A new machinery anomaly detection method towards complex operating conditions." Reliability Engineering & System Safety 236 (2023): 109319.

[13] W. Sultani, C. Chen and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, 6479-6488.

[14] Doshi, Keval & Yilmaz, Yasin. (2020). Any-Shot Sequential Anomaly Detection in Surveillance Videos. 4037-4042.

[15] S. Chandrakala, K. Deepak, and G. Revathy. 2022. Anomaly detection in surveillance videos: a thematic taxonomy of deep models, review and performance analysis. Artif. Intell. Rev. 56, 4 (Apr 2023), 3319–3368.

[16] Raja, Rohit, Prakash Chandra Sharma, Md Rashid Mahmood, and Dinesh Kumar Saini. "Analysis of anomaly detection in surveillance video: recent trends and future vision." Multimedia Tools and Applications 82, no. 8 (2023): 12635-12651.

[17] Kun Liu and Huadong Ma. 2019. Exploring Background-bias for Anomaly Detection in Surveillance Videos. In Proceedings of the 27th ACM International Conference on Multimedia (MM '19). Association for Computing Machinery, New York, NY, USA, 1490–1499.

[18] Khaleghi and M. S. Moin, "Improved anomaly detection in surveillance videos based on a deep learning method," 2018 8th Conference of AI & Robotics and 10th RoboCup Iranopen International Symposium (IRANOPEN), Qazvin, Iran, 2018, 73-81.

[19] Chen, Dongyue, Pengtao Wang, Lingyi Yue, Yuxin Zhang, and Tong Jia. "Anomaly detection in surveillance video based on bidirectional prediction." Image and Vision Computing 98 (2020): 103915.

[20] Choudhry, Nomica, Jemal Abawajy, Shamsul Huda, and Imran Rao. "A comprehensive survey of machine learning methods for surveillance videos anomaly detection." IEEE Access 11 (2023): 114680-114713.

[21] Xu, Jiehui, Haixu Wu, Jianmin Wang, and Mingsheng Long. "Anomaly transformer: Time series anomaly detection with association discrepancy." arXiv preprint arXiv:2110.02642 (2021).

[22] Kothadiya, Deep R., Chintan Bhatt, Aayushi Chaudhari, and Nilkumar Sinojiya. "GujFormer: A vision transformer-based architecture for Gujarati handwritten character recognition." In International Conference on Advances in Data-driven Computing and Intelligent Systems, Singapore: Springer Nature Singapore, 2023, 89-101.

[23] Ahn, Dasom, Sangwon Kim, Hyunsu Hong, and Byoung Chul Ko. "Star-transformer: a spatio-temporal cross attention transformer for human action recognition." In Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2023, 3330-3339.

[24] Kim, Hannah Halin, Shuzhi Yu, Shuai Yuan, and Carlo Tomasi. "Cross-attention transformer for video interpolation." In Proceedings of the Asian conference on computer vision, 320-337. 2022.

[25] Zhang, Haokui, Wenze Hu, and Xiaoyu Wang. "Fcaformer: Forward cross attention in hybrid vision transformer." In Proceedings of the IEEE/CVF International Conference on Computer Vision, 6060-6069. 2023.

[26] Gajjar, Dulari B., Prisha Faldu, Deep Rameshbhai Kothadiya, Aayushi Pushpakant Chaudhari, and Nikita M. Bhatt. "DeViTC: Deep-Vision Transformer to Recognize Originality of Currency." Computer 58, no. 5 (2025): 48-56.

[27] Chen, Liyang, Zhiyuan You, Nian Zhang, Juntong Xi, and Xinyi Le. "UTRAD: Anomaly detection and localization with U-transformer." Neural Networks 147 (2022): 53-62.

[28] Xu, Peng, Xiatian Zhu, and David A. Clifton. "Multimodal learning with transformers: A survey." IEEE Transactions on Pattern Analysis and Machine Intelligence 45, no. 10 (2023): 12113-12132.

[29] Hu, Ronghang, and Amanpreet Singh. "Unit: Multimodal multitask learning with a unified transformer." In Proceedings of the IEEE/CVF international conference on computer vision, 1439-1449. 2021.

[30] https://www.kaggle.com/datasets/odins0n/ucf-crime-dataset

[31] Ivan Nikolov. (2024). Reactive Anomaly Synthetic Data [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/7948157

[32] Vrskova, Roberta, Robert Hudec, Patrik Kamencay, and Peter Sykora. "Human activity classification using the 3DCNN architecture." Applied Sciences 12, no. 2 (2022): 931.

[33] Ur Rehman, Atiq, Samir Brahim Belhaouari, Md Alamgir Kabir, and Adnan Khan. "On the use of deep learning for video classification." Applied Sciences 13, no. 3 (2023): 2007.

[34] Ogawa, Takahiro, Yuma Sasaka, Keisuke Maeda, and Miki Haseyama. "Favorite video classification based on multimodal bidirectional LSTM." IEEE Access 6 (2018): 61401-61409.

[35] Hendi, Sajjad H., Hazeem B. Taher, and Karim Q. Hussein. "Automated video events detection and classification using CNN-GRU model." Wasit Journal of Computer and Mathematics Science 2, no. 4 (2023): 77-86.

[36] Zhuang, Xuqiang, Fang'ai Liu, Jian Hou, Jianhua Hao, and Xiaohong Cai. "Modality attention fusion model with hybrid multi-head self-attention for video understanding." Plos one 17, no. 10 (2022): e0275156.

[37] Chi, Lu, Guiyu Tian, Yadong Mu, and Qi Tian. "Two-stream video classification with cross-modality attention." In Proceedings of the IEEE/CVF international conference on computer vision workshops, 0-0. 2019.

[38] Long, Xiang, Chuang Gan, Gerard De Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. "Attention clusters: Purely attention based local feature integration for video classification." In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, 7834-7843.

[39] Ling, Charles X., Jin Huang, and Harry Zhang. "AUC: a better measure than accuracy in comparing learning algorithms." In Conference of the canadian society for computational studies of intelligence, Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, 329-341.

[40] Borji, Ali. "Enhancing sensor resolution improves CNN accuracy given the same number of parameters or FLOPS." arXiv preprint arXiv:2103.05251 (2021).

[41] Liu, Wen, Weixin Luo, Dongze Lian, and Shenghua Gao. "Future frame prediction for anomaly detection–a new baseline." In Proceedings of the IEEE conference on computer vision and pattern recognition, 6536-6545. 2018.