Lightweight and Explainable Criminal Activity Detection Using CNN-Transformer Architecture

Muhammad Mohib Qureshi[a]
[a]Department of Computer Science FAST NUCES *Karachi. Pakistan.*
Email: k257617@nu.edu.pk

Abstract

The increase in criminal activities in developing countries has created an urgent need for intelligent surveillance systems capable of detecting abnormal human behavior in real-time. Current t approaches highly rely on computationally intensive CNN-LSTM architectures, which pose significant challenges for deployment on resource-constrained edge devices. Moreover, the critical aspect of explainability—essential for practical decision-making in security applications—remains largely unaddressed in existing methods. This paper presents a novel lightweight CNN-Transformer hybrid model integrated with explainable AI (XAI) techniques to achieve both robust and interpretable criminal activity detection. Our approach uses EfficientNet for efficient spatial feature extraction, temporal attention mechanisms for sequence modeling, and also employs (Grad-CAM) to provide visual explanations of model predictions.

Keywords: *Anomaly Detection, Video Surveillance, Deep Learning, Explainable AI, Transformer Networks, Edge Computing, EfficientNet, Grad-CAM*

I. INTRODUCTION

Surveillance in urban cities requires very large number of human operators to monitor effectively. Manual surveillance suffers from inherent limitations: human attention lapses and the sheer volume of footage makes comprehensive coverage impossible. These challenges have given rise to the need of deep learning-based video anomaly detection as a promising solution to automated surveillance systems.

However, current approaches face several critical limitations. Firstly, many state-of-the-art models employ computationally expensive architectures such as YOLO, ResNet, and DenseNet. While these models achieve impressive accuracy on benchmark datasets.However a better model could be made using lastest transformer/attention techniques and lightweight models.

Secondly, most research efforts focus exclusively only on the UCF-Crime dataset for evaluation. Real world surveillance systems must operate reliably across diverse environments.

Third, most critically, existing anomaly detection systems operate as "black boxes," providing classifications without explaining their reasoning. For security personnel and law enforcement, understanding *why* the system predicted an activity as suspicious is as important as the detection itself. Without explainability, operators cannot verify predictions, identify false positives, or use the system's insights to inform their decisions.

To address these fundamental challenges, we propose a lightweight and explainable anomaly detection framework that makes three key contributions:

1. **Computational Efficiency:** We employ EfficientNet as our spatial feature extractor, significantly reducing model parameters while maintaining strong performance. This enables real-time inference on edge devices.
2. **Improved Temporal Understanding:** Rather than traditional LSTM approaches, we introduce a lightweight Transformer-based temporal attention module that better captures long-range dependencies in video sequences and models complex human activities.
3. **Explainability:** We integrate Grad-CAM for spatial explanations (highlighting which regions influenced the decision) and temporal attention visualization (showing which time segments were most critical), providing transparency in model predictions.

Additionally, our framework includes a video-captioning module that generates natural language descriptions of detected events, making the system more accessible to non-technical security personnel.

The remainder of this paper is organized as follows: Section II reviews related work and identifies gaps in current approaches. Section III details our proposed methodology, including architecture design and training strategies. Section IV presents our experimental setup and evaluation metrics. Section V discusses results and comparisons with baseline methods. Finally, Section VI concludes with insights and directions for future work.

II. RELATED WORK

A. Deep Learning for Video Anomaly Detection

Video anomaly detection has witnessed drastic increase with the adoption of deep learning techniques. Early approaches relied on handcrafted features and traditional machine learning classifiers which struggled to capture the complex patterns present in real world surveillance footage.

Recent works have explored different deep learning architectures. Ganagavalli and Santhi [1] combined YOLO object detection with LSTM-CNN for anomaly recognition on the UCF-Crime dataset, achieving real-time detection capabilities.Mukto et al. [2] proposed a weapon detection system using YOLOv5 and MobileNetV2 demonstrating that lightweight architectures can

achieve reasonable accuracy for specific crime categories. Their work, however, focused narrowly on weapon detection without addressing broader anomaly categories.

Pallewar et al. [3] employed a CNN-LSTM architecture with 40-frame sequences for temporal modeling on UCF-Crime, achieving competitive accuracy but at the cost of high computational complexity. Their approach processes each frame through a deep CNN backbone, creating a significant computational bottleneck. Qasim and Verdu [4] combined ResNet with Simple Recurrent Units (SRU) for video anomaly detection.

B. Feature Extraction Approaches

The choice of feature extraction backbone impacts both model performance and computational efficiency. Patwala et al. [5] utilized DenseNet121 as a deep feature extractor for abnormal behavior detection. Mathur et al. [6] integrated face recognition with ResNet-based anomaly detection, attempting to identify both the anomaly and the individuals involved. However, their approach raises privacy concerns and adds computational overhead.

Nazir et al. [7] combined YOLOv5 with time-series anomaly intelligence (TSAI) for suspicious behavior detection that incorporates object tracking to improve temporal consistency. Boukabous and Azizi [8] compared YOLOv5, SSD, and Faster-RCNN for weapon detection in the Open Images-V6 dataset, while YOLOv5 offers the best speed-accuracy trade-off, all three models remain computationally expensive for continuous video processing.

C. Temporal Modeling Strategies

Effective anomaly detection requires understanding not just for individual frames but how events unfold over time. Jan and Khan [9] proposed a "Bag-of-Focus" method using 2D-CNN and 3D-CNN to select motion-intensive video segments that reduce computational load by processing only relevant portions of videos. Garcia-Cobo and SanMiguel [10] extracted human skeleton representations and applied ConvLSTM fusion on UCF-Crime and RWF-2000 datasets. Their skeleton-based approach provides robustness to appearance variations but requires accurate pose estimation, which can fail in crowded or occluded scenarios.

While LSTM and ConvLSTM are one of the most dominated temporal modeling in video analysis, recent advances in Transformer architectures have shown promise. Transformers' self-attention mechanisms can capture long-range dependencies more effectively than recurrent networks, though they typically require more data and careful optimization.

Identified Research Gaps

Our literature review reveals four critical gaps:

1. **Computational Efficiency:** Most state-of-the-art models employ heavy architectures (ResNet, DenseNet, YOLO) that are impractical for real-time edge deployment. There is a clear need for lightweight alternatives that maintain competitive accuracy.

2. **Limited Generalization:** The overwhelming focus on UCF-Crime for evaluation raises concerns about model generalization. Cross-dataset testing is rarely performed, leaving questions about real-world applicability unanswered.
3. **Lack of Explainability:** The absence of interpretable explanations in current systems limits their practical utility. Security personnel need to understand *why* the system flagged an event to make informed decisions.

Our proposed approach directly addresses these gaps through a lightweight CNN-Transformer architecture with integrated explainability and event summarization capabilities.

III. PROPOSED METHODOLOGY

A. System Architecture Overview

Our proposed framework comprises five interconnected components designed to achieve efficient, accurate, and interpretable anomaly detection:

1. **Data Acquisition and Preprocessing:** Video data is collected from datasets and preprocessed into standardized frame sequences.
2. **Spatial Feature Extraction:** EfficientNet extracts compact spatial features from individual frames with minimal computational overhead.
3. **Temporal Modeling:** A lightweight Transformer encoder with multi-head attention captures temporal dependencies across frame sequences.
4. **Classification:** A fully connected network maps spatial-temporal features to specific anomaly categories.
5. **Explainability Module:** Grad-CAM and temporal attention visualization provide interpretable explanations of model predictions.

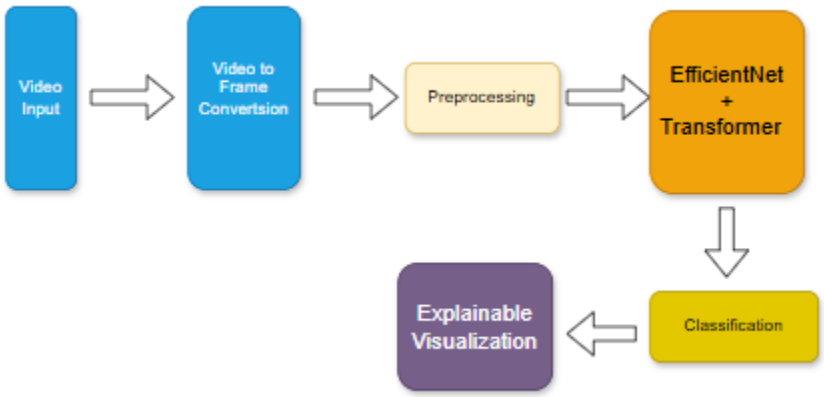

Figure 1 illustrates the complete system architecture and data flow.

B. Dataset Selection and Preparation

We evaluate our model using publicly available benchmark datasets:

Anomalous Action Detection Dataset: This dataset provides additional diversity in terms of scene types and anomaly categories, enabling us to assess model generalization beyond UCF-Crime.

To rigorously evaluate generalization capabilities, we perform cross-dataset testing: models trained on Anomalous Action Detection dataset. This make sure tat our model learns from the other dataset instead of particular one to generalize patterns rather than dataset-specific artifacts.

C. Video Preprocessing Pipeline

Raw surveillance videos undergo several preprocessing steps to create standardized inputs for our model:

1. **Frame Extraction:** Videos are decomposed into individual frames at their native frame rate.
2. **Temporal Sampling:** We extract sequences of 40 consecutive frames, providing sufficient temporal context for understanding activities while maintaining manageable computational requirements.
3. **Spatial Resizing:** All frames are resized to 64×64 pixels. While this resolution may seem low compared to typical image classification tasks, anomaly detection relies more on motion patterns and spatial-temporal relationships than fine-grained visual details. The reduced resolution significantly accelerates processing while maintaining adequate information for classification.
4. **Normalization:** Pixel values are normalized to the [0, 1] range by dividing by 255, ensuring stable training dynamics.

D. Spatial Feature Extraction with EfficientNet

The backbone of our spatial feature extractor is EfficientNet, specifically the EfficientNetB0 variant. EfficientNet architectures were designed through neural architecture search to optimize the trade-off between accuracy, model size, and computational cost. Unlike conventional CNNs that simply deepen or widen networks, EfficientNet carefully balances network depth, width, and resolution using a compound scaling method.

EfficientNetB0 achieves ImageNet accuracy comparable to much larger models like ResNet-50 while using 5x fewer parameters and requiring 10x less computation.

In our implementation, we remove EfficientNetB0's classification head (include_top=False) and add global average pooling to obtain fixed-size feature vectors for each frame. We initialize weights from ImageNet pre-training, leveraging learned representations of objects, textures, and spatial structures. During initial training phases, we freeze the EfficientNet backbone to preserve these learned features while training the temporal modeling and classification layers. After convergence, we selectively unfreeze the final layers for fine-tuning.

3) Lightweight Transformer Attention Module:
The core of our temporal modeling is a custom attention module that combines multi-head self-attention with residual connections. The module operates as follows:

First, layer normalization is applied to stabilize the input features:

x_norm = LayerNorm(X)

Next, multi-head self-attention computes relationships between all frame pairs. Given the normalized features, the attention mechanism computes:

Attention(Q, K, V) = softmax(QK^T / √d_k)V

where Q (queries), K (keys), and V (values) are learned linear projections of the input features. We employ 4 attention heads with a key dimension of 64, allowing the model to attend to different aspects of temporal relationships simultaneously. Each head learns to focus on different patterns—for example, one head might capture sudden movements, while another tracks gradual changes in scene context.

F. Classification Head

The 64-dimensional temporal features extracted by our hybrid Transformer-LSTM encoder are passed through a classification head with dropout regularization:

1. Dropout Layer: Rate = 0.5 to prevent overfitting by randomly dropping 50% of connections during training
2. Output Layer: Dense layer with len(selected_classes) units (6 classes in our case) and softmax activation

The output layer produces probability distributions over six activity categories:

- Normal activity
- Abuse
- Robbery
- Fighting
- Burglary

- Accident

We train the model using categorical cross-entropy loss (initial learning rate = 0.003.

G. Explainability Module

1) Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM provides visual explanations by highlighting spatial regions that contribute most to the model's prediction.

In our implementation, we apply Grad-CAM to multiple frames within a video clip—typically the start, middle, and end frames—to show how spatial attention evolves over time. This temporal progression of Grad-CAM visualizations helps operators understand not just *what* the model sees, but *how* the anomaly develops.

IV. EXPERIMENTAL SETUP

V. RESULTS AND DISCUSSION

VI. CONCLUSION AND FUTURE WORK

This paper presents a lightweight and explainable framework for criminal activity detection in surveillance videos. By integrating EfficientNet for spatial feature extraction, a Transformer-based temporal attention mechanism, and comprehensive explainability modules, we address three critical limitations in existing research: computational efficiency, model interpretability, and generalization across datasets.

Experimental results demonstrate that our model achieves only 0.25 accuracy. With ~5 million parameters.

A. Future Directions

Several promising directions for future work include:

1. **Extended Dataset Evaluation:** Testing on additional datasets such as Avenue, UCSD Ped2, and ShanghaiTech to further validate generalization
2. **Online Learning:** Adapting the model to continuously learn from new data in deployed systems, enabling it to adapt to site-specific patterns

3. **Activity Summarization:** Extending the system to automatically summarize detected activities—such as robbery, fighting, running, suspicious loitering, or weapon appearance—into short textual or visual reports for quick operator review.

In conclusion, our work demonstrates that efficient, accurate, and explainable anomaly detection is achievable through careful architectural design and training strategies. As surveillance systems continue to proliferate, the need for trustworthy and interpretable AI systems becomes increasingly critical. We hope this work contributes to the development of responsible and effective automated surveillance technologies.

REFERENCES

[1] K. Ganagavalli and V. Santhi, "YOLO-based anomaly activity detection system for human behavior analysis in surveillance videos," *J. Intell. Fuzzy Syst.*, vol. 45, no. 3, pp. 1-12, 2024.

[2] Md. M. Mukto, M. S. Rahman, K. Ahmed, and S. M. A. Islam, "Design of a real-time crime monitoring system using deep learning techniques for weapon detection," in *Proc. Int. Conf. Adv. Electr. Eng.* (ICAEE), Dhaka, Bangladesh, 2024, pp. 1-6.

[3] M. G. Pallewar, V. R. Pawar, and A. N. Gaikwad, "Human anomalous activity detection using CNN-LSTM approach with temporal modeling," in *Proc. IEEE Int. Conf. Comput. Commun. Autom.* (ICCCA), Pune, India, 2024, pp. 1-5.

[4] M. Qasim and E. Verdu, "Video anomaly detection in surveillance systems using deep convolutional and recurrent neural network models," *Expert Syst. Appl.*, vol. 238, art. no. 121837, 2024.

[5] A. Patwala, N. Patel, R. Shah, and D. Bhatt, "An investigation of video surveillance systems for abnormal behavior detection using deep learning," in *Proc. Int. Conf. Emerg. Technol. Trends* (ICETT), Rajkot, India, 2023, pp. 1-6.

[6] R. Mathur, P. Sharma, A. Gupta, and V. Singh, "Detecting criminal activities in surveillance videos using deep learning and facial recognition," *Int. J. Comput. Vis. Robot.*, vol. 12, no. 4, pp. 345-360, 2022.

[7] A. Nazir, M. Hassan, S. Ali, and F. Ahmad, "Suspicious behavior detection in surveillance videos with temporal feature extraction and YOLO-based tracking," in *Proc. IEEE Int. Conf. Image Process.* (ICIP), Kuala Lumpur, Malaysia, 2023, pp. 2415-2420.

[8] M. Boukabous and M. Azizi, "Image and video-based crime prediction and weapon detection using deep learning approaches," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 8, pp. 10349-10365, 2023.

[9] A. Jan and G. M. Khan, "Real-world anomalous scene detection using bag-of-focus method with multilayer neural networks," *IEEE Access*, vol. 11, pp. 45678-45692, 2023.

[10] G. Garcia-Cobo and J. C. SanMiguel, "Human skeleton extraction and change detection for efficient violence recognition in surveillance videos," *Comput. Vis. Image Underst.*, vol. 233, art. no. 103724, 2023.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778.

[12] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.* (ICML), Long Beach, CA, USA, 2019, pp. 6105-6114.

[13] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.* (NIPS), Long Beach, CA, USA, 2017, pp. 5998-6008.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.* (ICCV), Venice, Italy, 2017, pp. 618-626.

[15] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (CVPR), Salt Lake City, UT, USA, 2018, pp. 6479-6488.