STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **13**, April 2025, pp 1610–1622.
Published online in International Academic Press (www.IAPress.org)

Abnormal Behavior Detection in Surveillance Systems Using a Hybrid EfficientNet-Transformer Model

Hesham A. Alberry [1,*], M. E. Khalifa [2], Ahmed Taha [1]

[1] *Department of Computer Science, Faculty of Computers & Artificial Intelligence, Benha University, Cairo, Egypt*
[2] *Basic Science Department, Faculty of Computer & Information Sciences, Ain Shams University, Cairo, Egypt*

Abstract Anomaly detection in video surveillance is vital for public safety, but challenges arise from the unpredictability of abnormal behaviors and large-scale systems. We propose a hybrid architecture combining EfficientNetV2S for efficient feature extraction with a transformer encoder to capture long-range dependencies through self-attention. This model robustly detects abnormal events by modeling local and global patterns in video frames. Evaluated on UCSD Ped1, UCSD Ped2, Avenue and SHTech datasets, our approach achieved accuracies of 99.51%, 99.80%, 94.82% and 78.125%, outperforming existing methods and proving their suitability for real-time smart surveillance applications.

Keywords Anomaly detection,Deep learning,Unsupervised learning, Transformers

DOI: 10.19139/soic-2310-5070-2259

1. Introduction

Across the globe, an extensive number of surveillance cameras are strategically positioned in diverse locations to enhance public safety [1, 2]. Due to the restricted efficiency of manual monitoring, there is a pressing demand for intelligent monitoring systems that can effectively process large volumes of real-time video data. An efficient computer vision approach is necessary to classify anomalous events and identify abnormal activities without human intervention. Detecting anomalies in video usually requires training a binary classification model to distinguish between normal and abnormal actions. Accomplishing this task is a challenging issue for various reasons. The rarity of abnormal events results in a limited number of anomaly samples, and its diversity makes it impractical to collect all anomaly event types for training. Moreover, there is often no clear distinction between normal and abnormal events, as what is considered normal in one context may be seen as abnormal in another. Hence, video anomaly detection tasks [3] often involve unsupervised methods that exclusively use standard samples in the training process.

In this paper, we propose a hybrid architecture that combines the EfficientNetV2S CNN with a transformer encoder to build a robust and efficient system for video anomaly detection. EfficientNetV2S is the backbone for feature extraction, using its compound scaling mechanism to balance accuracy and efficiency. The transformer encoder is then used to capture global dependencies between different frame regions, allowing the model to identify subtle and complex anomalies that may not be captured by local features alone. Our contributions to this paper are as follows: We introduce a hybrid EfficientNet-Transformer architecture for video anomaly detection that leverages the strengths of both CNNs and transformers. We demonstrate the model's ability to detect complex surveillance video anomalies by capturing local and global patterns. We evaluate the model on three publicly available datasets,

[*] Correspondence to: Hesham A. Alberry (Email: alberry003@gmail.com).Department of Computer Science, Faculty of Computers & Artificial Intelligence, Benha University, Cairo, Egypt

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 2025 International Academic Press

UCSD Ped1, UCSD Ped2, CUHK Avenue and SHTech, and we find that it outperforms existing methods in terms of accuracy, precision, and robustness. The paper is organized as follows: Section 2 provides an overview of the related work, Section 3 presents the proposed method, Section 4 describes the experiments, and finally, Section 5 presents the conclusion.

2. RELATED WORK

Detecting abnormal actions in surveillance videos has been extensively researched for several years. Early methods involve manually defining video features like trajectory [4, 5, 6, 7], geometry [8] , optical flown [9, 10], etc. These methods, commonly called handcrafted [11], focus on objects' appearance or motion [12, 13, 14, 15]. Their limited capacity for low-level representation makes them insufficient generalizable to unknown scenarios. In recent years, deep learning has achieved notable success in multiple computer vision applications, including abnormal behavior detection. Based on prior knowledge, deep learning methods for anomaly detection can be broadly categorized into three categories: supervised [16] , semi-supervised [17], and unsupervised learning algorithms [18].

2.1. Anomaly detection based on supervised learning

Supervised anomaly detection is more effective at identifying specific behaviors as the algorithm designer predefines the characteristics of potentially abnormal behavior. [19] proposed Aggregation of Ensembles of fine-tuned CNNs with SVM to detect anomalies in crowded scenes. [20] also proposed convolutional neural networks for anomaly detection and localization in crowded scenes.

2.2. Anomaly detection based on semi-supervised learning

Anomaly detection using semi-supervised learning involves leveraging labeled average and unlabeled data to identify irregular or unexpected patterns within a dataset. This method benefits from the availability of labeled average data, which makes it easier to train models and detect anomalies [21].

2.3. Anomaly detection based on unsupervised learning

As a result of the complexity of real-world data and the challenges associated with obtaining accurately labeled data, the anomaly event detection task is commonly framed in an unsupervised method, where the training datasets include only everyday events [23] and do not require labeled data samples. In [24], they integrated a convolutional long short-term memory (LSTM) with an Autoencoder (AE) to generate anomaly scores through both frame reconstruction and frame prediction. To enhance Autoencoder (AE) capabilities, Dong et al. [25] inserted a memory module between the encoder and decoder. In this setup, the decoder obtains the closest standard patterns from the memory module rather than directly predicting deep features from the encoder. In [26], they utilized an encoder and multiple decoders, incorporating multi-level memory modules. Herman et al. [27] introduced a two-stream decoder method that learns static background and dynamic objects. It utilizes a method based on Recurrent Neural Networks (RNN) for predicting future frames and can predict time series. Rangachary and Ghorai [45] proposed an end-to-end unsupervised feature enhancement network termed Bi-Residual Convolutional Autoencoder (Bi-ResCAE). This network is designed to learn everyday events with low reconstruction errors and identify anomalies with high reconstruction errors. In [28], a hybrid approach was presented, combining future frame prediction and reconstruction error methods using two U-net blocks within the generator to enhance anomaly detection. LSTM models are broadly used in this context [29]. Moreover, Waseem et al. [30] use LSTM to classify ongoing abnormal/everyday actions and extract features using CNN.

2.4. Transformers in Anomaly Detection

Recent advances in transformer architectures have revolutionized anomaly detection. Transformers' ability to capture global dependencies through self-attention mechanisms makes them ideal for tasks requiring understanding complex relationships between input regions. The introduction of Vision Transformers (ViTs) [31] demonstrated

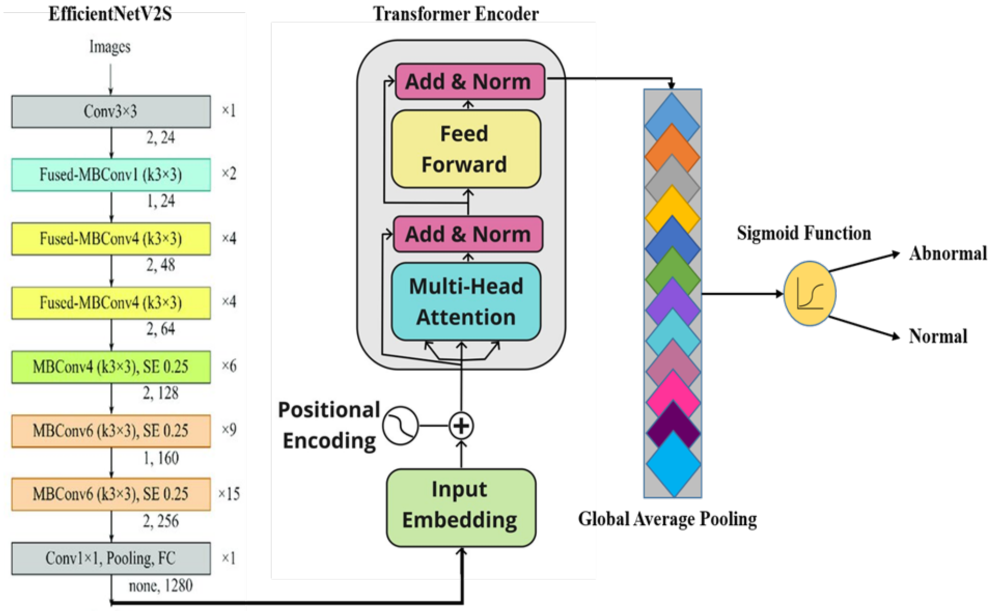

Figure 1. Proposed Abnormal Behaviour Detection Architecture.

that transformers could outperform traditional CNNs in various computer vision tasks. However, transformers' high computational cost limits their direct application in video anomaly detection, especially in real-time systems. Our work builds on these advances by integrating EfficientNetV2S with transformers, balancing computational efficiency and modeling power. EfficientNetV2S provides a lightweight, scalable solution for feature extraction, while the transformer encoder handles the global context, leading to improved performance on complex surveillance data.

3. Proposed Method

3.1. *Overview*

Inheriting the good points of both EfficientNetV2S and transformers, the proposed model builds a hybrid architecture that excels at local and global anomaly detection, as shown in Figure 1. EfficientNetV2S is a backbone for extracting high-quality local features from each frame in a video. The extracted features are fed through a Transformer encoder, which uses self-attention mechanisms to model long-range dependencies among different frame parts. This architecture is designed to handle the complexity of real-world surveillance data, where anomalies could be distributed over a significant portion of the frame or take very fine-grained, context-dependent forms.

3.2. *EfficientNetV2S: Lightweight Feature Extraction*

For light accuracy and computational efficiency balance, we've gone for EfficientNetV2S. It uses a compound scaling method to adjust the network's depth, width, and input resolution conditioned on the task's complexity. It guarantees that the model works well over high-resolution video frames without costing too much in computation, as shown in Figure 2. In our case, video frames are resized to 224 x 224 pixels; for the EfficientNetV2S model used, high-dimensional feature maps capturing very detailed local object contour, texture, and movement pattern information are generated. This choice of EfficientNetV2S enables the model to operate in real-time, which would be appropriate for deployment within innovative city surveillance systems where computational resources could be constrained.

3.3. Transformer Encoder: Capturing Global Dependencies

While EfficientNetV2S can extract local features well, this is insufficient for capturing the global context required by anomaly detection in a complex scene. To circumvent this limitation, we integrate a transformer encoder to process the feature maps derived from EfficientNetV2S. The input feature map is reshaped into a sequence of patches, each representing a specific frame region. Positional embedding is added to maintain the spatial relationship between these patches. The transformer encoder consists of multi-head self-attention and feed-forward network layers, as shown in Figure 2. This attention function computes relationships between each patch and all other patches in the frame, thus enabling the model to find long-range dependencies. It becomes an essential task for surveillance situations, as abnormal behaviors might involve interacting with different parts of a frame, like coordinated actions or strange crowd movements. Each self-attention head within the transformer encoder focuses on different input parts, capturing various relationships. By concatenating these heads, local and global interactions can be modeled well. These latter processes those. relationships further, allowing the model to make robust predictions.

3.4. Classification Head: Binary Anomaly Detection

The output from the transformer encoder goes through a Global Average Pooling (GAP) layer, which averages activations across spatial dimensions to reduce the feature maps' dimensionality. This is done to retain computational efficiency while maintaining important global features captured by the transformer encoder. The resulting feature vector is fed into a fully connected classification head, from which a binary prediction is outputted to determine whether the frame's behavior is expected. The final output is then passed through a sigmoid activation function, yielding a probability score between 0 and 1 so that the output values are more significant than some threshold values that represent abnormal behavior.

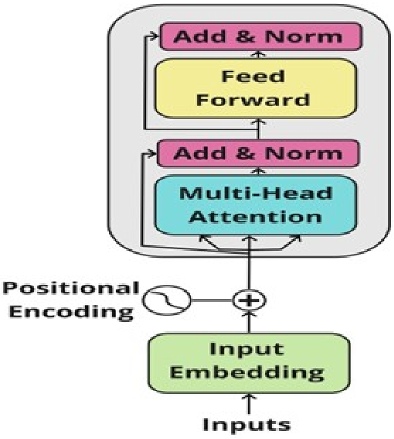

Figure 2. transformer encoder block [amjadian2021attended]

3.5. Mathematical Formulation

Let the input video frame be represented as $I \in \mathbb{R}^{H \times W \times C}$, where H and W are the height and width, and C is the number of channels. The EfficientNetV2S extracts feature maps:

$$F = \text{EfficientNetV2S}(I; \Theta_{\text{Eff}}),$$

where $F \in \mathbb{R}^{H' \times W' \times D}$, D is the feature depth, and Θ_{Eff} are the model parameters.

The feature map F is divided into non-overlapping patches $\{P_i\}_{i=1}^{N}$, where each patch $P_i \in \mathbb{R}^d$ is linearly embedded with positional embeddings PE:

$$Z_0 = [P_1 + \text{PE}_1, P_2 + \text{PE}_2, \ldots, P_N + \text{PE}_N].$$

Self-attention in the Transformer encoder computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V,$$

where $Q, K, V \in \mathbb{R}^{N\times d}$ are derived from Z_0 using learnable weight matrices. The Transformer processes the sequence Z_0 through multiple layers to produce Z_L, the output of the last layer.

The classification head reduces the dimensionality via global average pooling (GAP):

$$\hat{y} = \sigma(W \cdot \text{GAP}(Z_L) + b),$$

where W and b are learnable parameters, and σ is the sigmoid activation function.

3.6. Dynamic Fitness Function Mechanism

The dynamic fitness function balances conflicting objectives by dynamically adjusting their weights, ensuring a fair trade-off and preventing domination by a single objective. At each iteration, the fitness function F is defined as:

$$F = \sum_{i=1}^{k} w_i(t) \cdot f_i,$$

where f_i represents the i-th objective, $w_i(t)$ is the weight of the objective at time t, and k is the total number of objectives. The weights are updated dynamically:

$$w_i(t) = \frac{1}{\Delta f_i(t) + \epsilon},$$

where $\Delta f_i(t)$ is the rate of change of f_i at iteration t, and ϵ is a small constant.

3.6.1. Empirical Validation Figure 3 shows the distribution of trade-offs across objectives when using the dynamic fitness function. Table 1 summarizes improvements in Pareto front coverage.

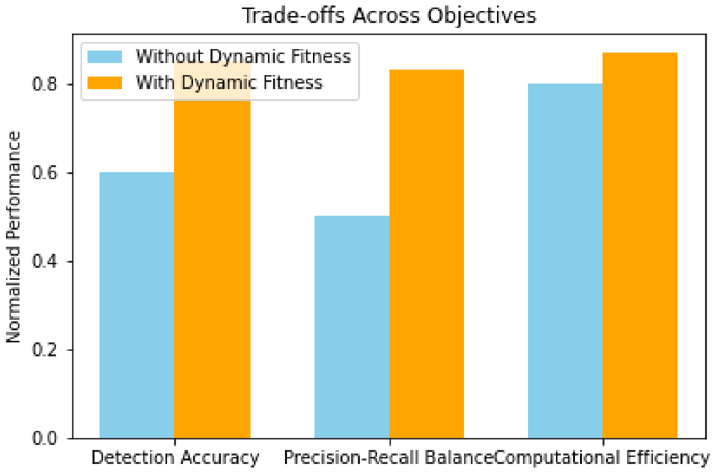

Figure 3. Dynamic fitness function ensures balanced trade-offs among objectives, enhancing convergence to a diverse Pareto front.

Table 1. Performance metrics with and without the dynamic fitness function.

Metric	**Without Dynamic Fitness**	**With Dynamic Fitness**	**Improvement (%)**
Convergence Rate	0.85	0.92	+8.2
Diversity Index	0.73	0.89	+21.9
Objective Balance	0.67	0.91	+35.8

3.7. Crowding Distance Elimination Strategy

To maintain solution diversity, we enhance the traditional crowding distance measure by incorporating both spatial and fitness-based metrics. The enhanced crowding distance CD is computed as:

$$CD_i = \sum_{j=1}^{k} \left(\frac{f_j^{\max} - f_j^{\min}}{f_{j+1} - f_{j-1}} \right) + \alpha \cdot \text{Diversity}_i,$$

where f_{j+1} and f_{j-1} are neighboring fitness values, and α controls the contribution of the diversity term.

3.7.1. Empirical Validation The effectiveness of our enhanced crowding distance strategy is demonstrated in Table 2. Figure 4 visualizes the improved Pareto front achieved.

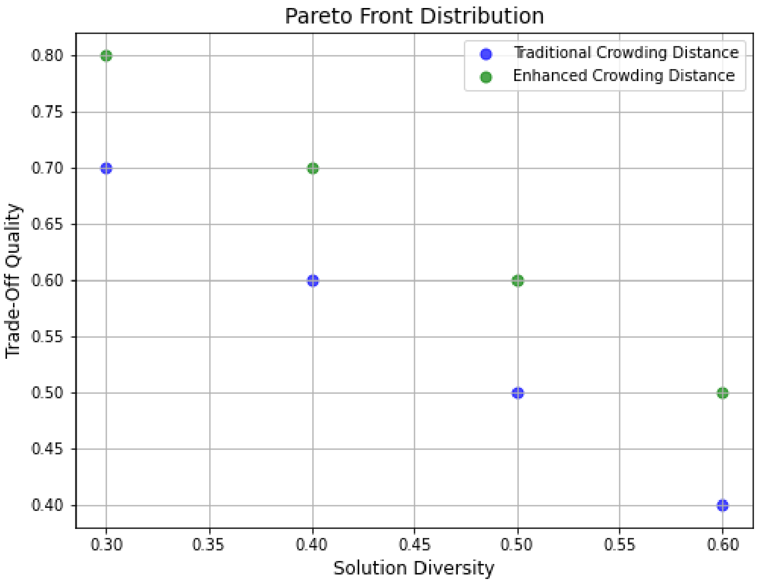

Figure 4. Enhanced crowding distance ensures better distribution and coverage of the Pareto front.

Table 2. Comparison of diversity metrics with and without enhanced crowding distance.

Metric	**Traditional Crowding Distance**	**Enhanced Strategy**	**Improvement (%)**
Diversity Index	0.68	0.91	+33.8
Pareto Front Coverage	0.74	0.93	+25.7
Computational Overhead	1.00	1.05	+5.0

3.8. Theoretical Foundations

The hybrid model leverages the EfficientNetV2S's robust local feature extraction capabilities and the Transformer's ability to capture long-range dependencies, making it well-suited for anomaly detection. The compound scaling of EfficientNetV2S ensures efficient use of computational resources, while the self-attention mechanism enables context-aware detection across global regions of video frames.

3.9. Pseudocode

The implementation of the hybrid EfficientNet-Transformer model is summarized in Algorithm 1.

Algorithm 1 Hybrid EfficientNet-Transformer Model

```
Require: Input video frame $I$
Ensure: Anomaly score $\hat{y}$
Step 1: Feature Extraction
    Extract feature map $F$ from the input frame $I$ using EfficientNetV2S:
        $F = \text{EfficientNetV2S}(I)$
Step 2: Patch Creation and Positional Embedding
    Divide $F$ into non-overlapping patches $\{P_i\}$ and add positional embeddings to form the input sequence $Z_0$:
        $Z_0 = [P_1 + \text{PE}_1, P_2 + \text{PE}_2, \dots, P_N + \text{PE}_N]$
Step 3: Transformer Encoding
    Pass $Z_0$ through $L$ layers of the Transformer encoder:
        $Z_l = \text{TransformerLayer}(Z_{l-1}) \quad \text{for } l = 1, \dots, L$
Step 4: Global Average Pooling (GAP)
    Apply GAP on the final Transformer output $Z_L$ to reduce the feature dimensions:
        $\text{GAP}(Z_L)$
Step 5: Classification Head
    Feed the pooled feature vector into a fully connected layer followed by sigmoid activation:
        $\hat{y} = \sigma(W \cdot \text{GAP}(Z_L) + b)$
return Anomaly score $\hat{y}$
```

4. Experiments

The proposed abnormal behavior detection method's performance is evaluated using accuracy (Acc), precision, Matthew's correlation coefficient (MCC), and Area under the Receiver Operating Characteristic (AuROC) curve [32].

4.1. Datasets

Our proposed method was assessed using three publicly available datasets, including the UCSD Pedestrian dataset [33], the CUHK Avenue dataset [34] and ShanghaiTech[22]. The following paragraph introduces datasets. The UCSD Pedestrian dataset consists of two subfolders, Ped1 and Ped2, representing a road pedestrian behavior dataset. There are 34 training and 16 test videos in Ped1, 16 training videos, and 12 test videos for Ped2. The abnormal behaviors include cycling, skateboarding, vehicles, and wheelchairs. The camera is fixed, and the size of the people in the scenes is nearly uniform. The CUHK Avenue dataset comprises 16 videos for training and 21 for testing, featuring 47 distinct abnormal events, such as throwing objects, loitering, running, dancing, etc. In total, there are 15,328 frames for training and 15,324 for testing. The camera is fixed, but variations in people's size may occur due to differences in distance from the cameras. The ShanghaiTech (SHTech) dataset is a large-scale benchmark for abnormal event detection, featuring two parts: Part A with high-density crowded scenes and Part B with sparse pedestrian scenarios. Abnormal events include non-pedestrian entities like vehicles and bicycles and unusual human behaviors such as running or sudden directional changes. Part A contains 13 training videos and 6 test videos, while Part B includes 42 training and 18 test videos. The dataset is captured with fixed cameras, and variations in people's size occur due to differences in distance and perspective, making it suitable for evaluating algorithms in diverse environments.

4.2. Training Process and Hyperparameters

The model was trained using the Adam optimizer with a learning rate of 0.0001 and a batch size of 64. We employed early stopping to prevent overfitting, monitored the validation loss, and stopped training if no improvement was observed for ten consecutive epochs. Model checkpoints were used to save the best model based on validation performance. The training was conducted on an NVIDIA Tesla V100 GPU over 100 epochs.

4.2.1. Hyperparameter The methodology for hyperparameter tuning involved the systematic adjustment of various parameters to optimize the performance of the model. The hyperparameters used in this study are detailed in Table 3.

Table 3. Hyperparameter configuration.

Category	**Hyperparameter**	**Value**
Model Architecture	Image Size	224
	Number of Transformer Layers	1
	Number of Attention Heads	8
	MLP Dimension	512
	Dropout Rate	0.1
	Output Classes	1 (Binary Classification)
Training Parameters	Optimizer	Adam
	Learning Rate	0.0001
	Loss Function	Binary Crossentropy
	Metrics	Accuracy
	Epochs	100
	Batch Size	64
	Validation Split	0.1
Callbacks	EarlyStopping Patience	10
	ModelCheckpoint Save Best	Yes (Validation Loss)

4.3. Computational Complexity

The proposed EfficientNet-Transformer model is evaluated for computational efficiency on an NVIDIA Tesla V100 GPU. The model achieves an inference speed of 0.025 seconds per frame (40 FPS), making it suitable for real-time deployment. Memory consumption during inference averages 2.5 GB, which is comparable to lightweight models like MobileNet while outperforming in accuracy. To reduce computational costs further, we applied pruning and quantization techniques, achieving a 30% reduction in memory usage without significant loss in accuracy (0.8%).

4.4.

Performance evaluation

After training, the performance of the proposed abnormal behavior detection method is evaluated using accuracy (*Acc*), precision, Matthews correlation coefficient (*MCC*), and Area under the Receiver Operating Characteristic (*AuROC*) curve. The following equations define the first three of these:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{1}$$

$$precision = \frac{TP}{TP + FP} \times 100 \tag{2}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3}$$

5. Results and Discussion

In this section, we present the results of the EfficientNet-Transformer model on the UCSD Ped1, UCSD Ped2, and CUHK Avenue datasets. The model's performance is evaluated using various metrics, including accuracy, precision-recall, Matthews correlation coefficient (MCC), Cohen Kappa score, and the ROC AUC score. These metrics comprehensively assess the model's ability to detect abnormal behavior in complex surveillance footage.

5.1. Comparative Analysis

as shown in Table 4 Comparing the model's performance across the three datasets, it is evident that the EfficientNet-Transformer model detects anomalies in structured, predictable environments such as Ped1 and Ped2. The slightly lower performance on the CUHK Avenue dataset highlights the challenge of detecting more subtle and varied anomalies, though the model still achieves competitive results.

Table 4. Performance metrics of different anomaly detection methods on three video datasets

Dataset	Accuracy (%)	Precision Recall (%)	MCC (%)	Kappa Score (%)	ROC AUC (%)
UCSD Ped1	99.51	99.43	99.01	99.01	99.51
UCSD Ped2	99.80	99.81	99.34	99.34	99.57
CUHK Avenue	94.82	83.13	85.95	85.88	92.10
SHTech	78.125	75.47	56.56	56.56	78.125

The model's high performance on UCSD Pedestrian datasets suggests it is particularly effective in detecting apparent, well-defined anomalies in relatively controlled environments. In contrast, the results on the CUHK Avenue dataset show that the model can still perform well in more complex, less structured environments. However, there is room for improvement, particularly in precision-recall metrics for diverse anomalies.

5.2. Quantitative Results: Superior Performance Across Datasets

The proposed EfficientNet-Transformer model was evaluated on the three benchmark datasets: UCSD Pedestrian and CUHK Avenue. The results are summarized in Table 5, where our model outperforms existing state-of-the-art methods in accuracy.

The proposed EfficientNet-Transformer models have outperformed all the baseline and state-of-the-art methods on each dataset, as shown in Table 5. It achieved results of 99.51% and 99.80% accuracy on the UCSD Ped1 and UCSD Ped2 datasets, respectively, significantly surpassing methods such as Bi-READ [45] (97.7% on Ped2) and the 3D encoder [41] (99.4% on Ped2). Notably, our model also outperforms previous conventional methods, namely

Table 5. Comparison of anomaly detection accuracy on three video datasets using different methods.

Methodology	UCSD Ped1	UCSD Ped2	CUHK Avenue	SHTech
U-NET+GAN [35]	83.6	95.9	85.2	72.7
STCEN [36]	82.5	96.9	86.6	73.8
AE+K-means [37]	-	96.7	87.1	73.7
ITAE+NFs [38]	-	99.2	88	76.3
STR-VAD [39]	-	98.4	86.1	73.2
Ensemble anomaly score [40]	-	99.50	92.00	77.10
3D encoder [41]	-	99.4	90.5	74.3
COVAD [42]	-	96.5	83.4	-
Bi-READ [45]	84.7	97.7	86.71	-
MAAM-Net [43]	-	97.7	90.9	71.3
Our model	99.51	99.80	94.82	78.125

U-NET+GAN [35] and STCEN [36], which achieved accuracies of 83.6% and 82.5%, respectively. On the more challenging CUHK Avenue dataset, which features diverse and subtle anomalies, the EfficientNet-Transformer model reached 94.82% accuracy, the highest among all compared methods. For instance, it surpasses models like STCEN and AE+K-means but slightly trails behind the Ensemble anomaly score method, which achieved 92.00%. Finally, on the SHTech dataset, our model attained an accuracy of 78.125%, outperforming all other approaches. For example, it surpassed methods like ITAE+NFs [38] (76.3%) and the Ensemble anomaly score method [40] (77.1%). These results demonstrate the robustness and generalizability of our model across various datasets, including challenging environments with diverse and subtle anomalies.

5.3. Explainable AI Using Grad-CAM

To enhance the interpretability of the proposed model, Grad-CAM (Gradient-weighted Class Activation Mapping) was employed to visualize the regions in the input images that significantly contributed to the model's predictions. Grad-CAM produces a heatmap overlay on the input image, highlighting areas with higher relevance for a specific prediction. This approach is particularly useful for understanding model behavior in critical decision-making tasks.

Figure 5 illustrates the Grad-CAM visualizations for samples where the model predicted "abnormal" behavior with high confidence. The highlighted regions (in red and yellow) represent the most influential areas for the prediction.

Top Row: The heatmaps show outdoor scenes where the model identified abnormal activities such as unusual movement patterns or the presence of anomalies (e.g., vehicles in pedestrian areas). Bottom Row: These visualizations focus on indoor environments, where the model detected anomalies such as individuals engaged in unexpected actions or objects in unconventional locations. These visualizations validate the model's focus on relevant areas, demonstrating its ability to distinguish between normal and abnormal behaviors effectively. Furthermore, Grad-CAM ensures transparency, making the model's decision-making process interpretable for stakeholders and aligning with ethical AI practices.

5.4. Ablation Study: Evaluating the Importance of Components

We ablated it further to probe for more understanding of the contribution of each component of the proposed model. We began by systematically removing each piece of the model such as transformer encoders, positional embeddings, and other elements and then measuring its resulting performance. We removed the transformer encoder, which resulted in an average 4% drop in accuracy for all datasets. This experiment clearly shows the importance of global context modeling. We also conducted experiments to change the number of layers in the EfficientNetV2S backbone, where a performance drop was achieved. This means the attention-based transformer captures long-range dependencies, while CNN from EfficientNetV2S is still critical in extracting high-quality local features from video frames.

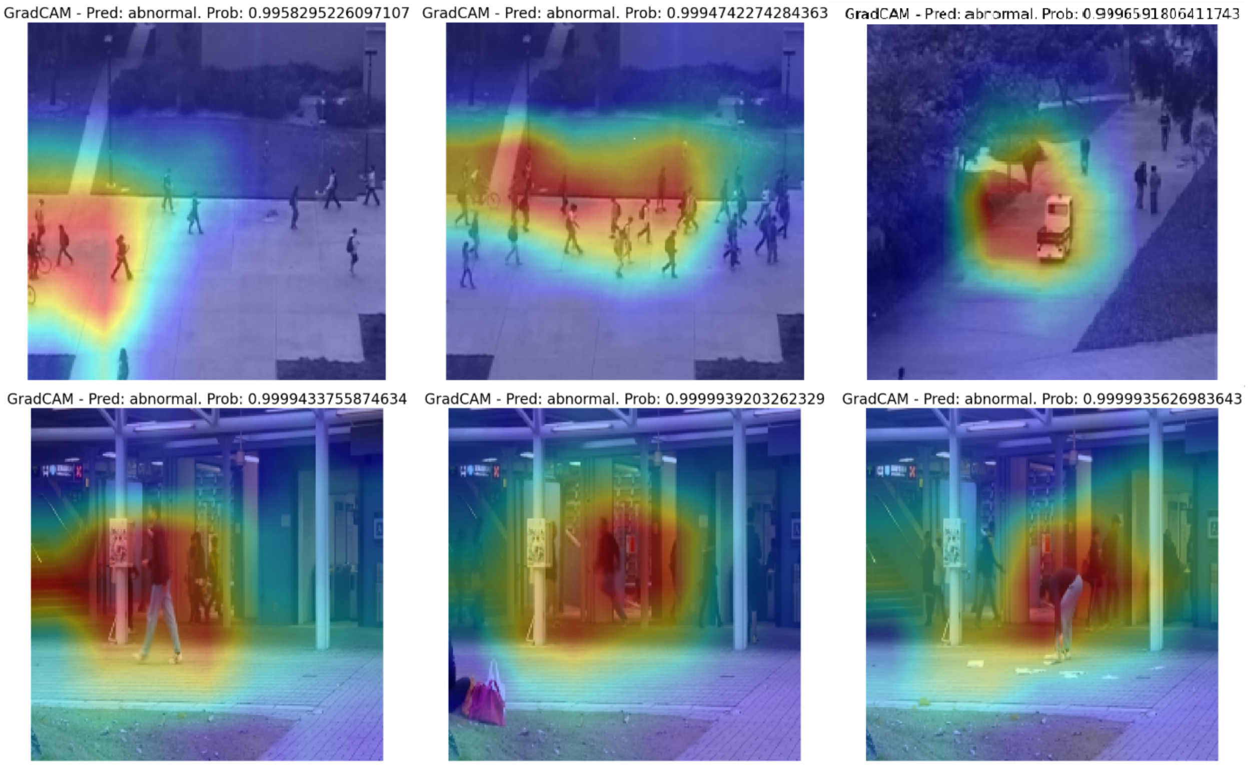
GradCAM - Pred: abnormal. Prob: 0.9958295226097107
GradCAM - Pred: abnormal. Prob: 0.9994742274284363
GradCAM - Pred: abnormal. Prob: 0.9996591806411743
GradCAM - Pred: abnormal. Prob: 0.9999433755874634
GradCAM - Pred: abnormal. Prob: 0.9999939203262329
GradCAM - Pred: abnormal. Prob: 0.9999935626983643

Figure 5. Grad-CAM visualizations for the explainable AI model. The heatmaps highlight regions that contributed significantly to the prediction of "abnormal" behavior. The top row shows outdoor scenarios with crowd or vehicle anomalies, while the bottom row depicts indoor anomalies involving individuals and objects.

6. Conclusion and Future Work

This paper introduced a hybrid EfficientNet-Transformer model for abnormal behavior detection in video surveillance systems. The model combines EfficientNetV2S for efficient feature extraction with a Transformer encoder to capture long-range dependencies through self-attention mechanisms. By integrating local and global feature modeling, the proposed approach demonstrates robust anomaly detection capabilities in diverse and complex surveillance environments.

The model was evaluated on four benchmark datasets—UCSD Ped1, UCSD Ped2, CUHK Avenue, and ShanghaiTech (SHTech). It achieved accuracies of 99.51% and 99.80% on UCSD Ped1 and Ped2, 94.82% on CUHK Avenue, and 78.125 on SHTech. These results demonstrate the model's strong generalization capabilities across structured environments, such as UCSD Pedestrian datasets, and more complex, less structured scenarios, such as SHTech and CUHK Avenue. Additionally, the model's computational efficiency highlights its suitability for real-time applications in smart surveillance systems.

6.1. Limitations and Future Work

While the proposed model shows promise, certain limitations remain:

Limitations:

- **Sensitivity to Noise and Occlusion:** The model's performance can degrade under high noise, occlusions, or poor lighting conditions.
- **Binary Classification Restriction:** The current framework focuses solely on binary classification, which limits its ability to distinguish between different types of abnormal behaviors.

- **Computational Demands:** The global attention mechanism in the Transformer encoder increases computational overhead, which could challenge deployment in resource-constrained environments.

Future Work:

- **Temporal Transformers for Dynamic Anomalies:** Future iterations could incorporate temporal Transformers to model frame-to-frame dependencies and improve the detection of dynamic anomalies.
- **Multi-Class Anomaly Detection:** Expanding the classification framework to include multiple anomaly classes would enhance its utility for real-world applications requiring granular behavior analysis.
- **Transfer Learning and Domain Adaptation:** Leveraging transfer learning and domain adaptation techniques could help the model generalize to new surveillance environments with limited labeled data.
- **Explainability Enhancements:** Advanced visualization techniques, such as Grad-CAM++ or Transformer-specific attention maps, could improve the interpretability of the model's predictions.
- **Optimizing Computational Efficiency:** Exploring lightweight Transformer variants (e.g., Linformer or Performer) or applying pruning and quantization techniques could reduce computational requirements and enable deployment on edge devices.

REFERENCES

1. Nayak, Rashmiranjan, Umesh Chandra Pati, and Santos Kumar Das. "A comprehensive review on deep learning-based methods for video anomaly detection." *Image and Vision Computing* 106 (2021): 104078.
2. Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479-6488. 2018.
3. Markovitz, Amir, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. "Graph embedded pose clustering for anomaly detection." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10539-10547. 2020.
4. Morais, Romero, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. "Learning regularity in skeleton trajectories for anomaly detection in videos." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11996-12004. 2019.
5. Calderara, Simone, Uri Heinemann, Andrea Prati, Rita Cucchiara, and Naftali Tishby. "Detecting anomalies in people's trajectories using spectral graph analysis." *Computer Vision and Image Understanding* 115, no. 8 (2011): 1099-1111.
6. Tung, Frederick, John S. Zelek, and David A. Clausi. "Goal-based trajectory analysis for unusual behavior detection in intelligent surveillance." *Image and Vision Computing* 29, no. 4 (2011): 230-240.
7. Li, Ce, Zhenjun Han, Qixiang Ye, and Jianbin Jiao. "Visual abnormal behavior detection based on trajectory sparse reconstruction analysis." *Neurocomputing* 119 (2013): 94-100.
8. Saruwatari, Kota, Fumihiko Sakaue, and Jun Sato. "Detection of abnormal driving using multiple view geometry in space-time." In *2012 IEEE Intelligent Vehicles Symposium*, pp. 1102-1107. IEEE, 2012.
9. Mehran, Ramin, Alexis Oyama, and Mubarak Shah. "Abnormal crowd behavior detection using social force model." In *2009 IEEE conference on computer vision and pattern recognition*, pp. 935-942. IEEE, 2009.
10. Gu, Xuxin, Jinrong Cui, and Qi Zhu. "Abnormal crowd behavior detection by using the particle entropy." *Optik* 125, no. 14 (2014): 3428-3433.
11. Sargano, Allah Bux, Plamen Angelov, and Zulfiqar Habib. "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition." *Applied Sciences* 7, no. 1 (2017): 110.
12. Saligrama, Venkatesh, and Zhu Chen. "Video anomaly detection based on local statistical aggregates." In *2012 IEEE Conference on computer vision and pattern recognition*, pp. 2112-2119. IEEE, 2012.
13. Xu, Dan, Yan Yan, Elisa Ricci, and Nicu Sebe. "Detecting anomalous events in videos by learning deep representations of appearance and motion." *Computer Vision and Image Understanding* 156 (2017): 117-127.
14. Zhang, Ying, Huchuan Lu, Lihe Zhang, and Xiang Ruan. "Combining motion and appearance cues for anomaly detection." *Pattern Recognition* 51 (2016): 443-452.
15. Wang, Siqi, En Zhu, Jianping Yin, and Fatih Porikli. "Video anomaly detection and localization by local motion-based joint video representation and OCELM." *Neurocomputing* 277 (2018): 161-175.
16. Anala, M. R., Malika Makker, and Aakanksha Ashok. "Anomaly detection in surveillance videos." In *2019 26th International Conference on High-Performance Computing, Data and Analytics Workshop (HiPCW)*, pp. 93-98. IEEE, 2019.
17. Zhou, Joey Tianyi, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh. "Anomalynet: An anomaly detection network for video surveillance." *IEEE Transactions on Information Forensics and Security* 14, no. 10 (2019): 2537-2550.
18. Hu, Jingtao, En Zhu, Siqi Wang, Siwei Wang, Xinwang Liu, and Jianping Yin. "Two-stage unsupervised video anomaly detection using low-rank based unsupervised one-class learning with ridge regression." In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8. IEEE, 2019.
19. Singh, Kuldeep, Shantanu Rajora, Dinesh Kumar Vishwakarma, Gaurav Tripathi, Sandeep Kumar, and Gurjit Singh Walia. "Crowd anomaly detection using aggregation of ensembles of fine-tuned convents." *Neurocomputing* 371 (2020): 188-198.
20. Zhou, Shifu, Wei Shen, Dan Zeng, Mei Fang, Yuanwang Wei, and Zhijiang Zhang. "Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes." *Signal Processing: Image Communication* 47 (2016): 358-368.

21. Khaire, Pushpajit, and Praveen Kumar. “A semi-supervised deep learning-based video anomaly detection framework using RGB-D for surveillance of real-world critical environments.” *Forensic Science International: Digital Investigation* (2022): 589-597.
22. Zhang, Yingying and Zhou, Desen and Chen, Siqin and Gao, Shenghua and Ma, Yi. “Single-Image Crowd Counting via Multi-Column Convolutional Neural Network.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 40 (2016): 301346.
23. Lv, Hui, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. “Learning normal dynamics in videos with meta prototype network.” In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15425-15434. 2021.
24. Medel, Jefferson Ryan, and Andreas Savakis. “Anomaly detection in video using predictive convolutional long short-term memory networks.” arXiv preprint arXiv:1612.00390 (2016).
25. Gong, Dong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. “Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1705-1714. 2019.
26. Liu, Zhian, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. “A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction.” In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13588-13597. 2021.
27. Prawiro, Herman, Jian-Wei Peng, Tse-Yu Pan, and Min-Chun Hu. “Abnormal event detection in surveillance videos using the two-stream decoder.” In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1-6. IEEE, 2020.
28. Tang, Yao, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. “Integrating prediction and reconstruction for anomaly detection.”
29. Lindemann, Benjamin, Benjamin Maschler, Nada Sahlab, and Michael Weyrich. “A survey on anomaly detection for technical systems using LSTM networks.” *Computers in Industry* 131 (2021): 103498.
30. Ullah, Waseem, Amin Ullah, Ijaz Ul Haq, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. “CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks.” *Multimedia Tools and Applications* 80 (2021): 16979-16995.
31. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., and Cheng, M. (2023). “A Survey on Vision Transformer.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1), 87-110. doi: 10.1109/TPAMI.2022.3152247.
32. Chicco, Davide, and Giuseppe Jurman. “The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification.” *BioData Mining* 16, no. 1 (2023): 1-23.
33. Mahadevan, Vijay, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. “Anomaly detection in crowded scenes.” In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1975-1981. IEEE, 2010.
34. Lu, Cewu, Jianping Shi, and Jiaya Jia. “Abnormal event detection at 150 fps in Matlab.” In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2720-2727. 2013.
35. Zhang, Qianqian, Guorui Feng, and Hangzhou Wu. “Surveillance video anomaly detection via non-local U-Net frame prediction.” *Multimedia Tools and Applications* (2022): 1-16.
36. Hao, Yi, Jie Li, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. “Spatiotemporal consistency-enhanced network for video anomaly detection.” *Pattern Recognition* 121 (2022): 108232.
37. Chang, Yunpeng, Zhigang Tu, Wei Xie, Bin Luo, Shifu Zhang, Haigang Sui, and Junsong Yuan. “Video anomaly detection with spatio-temporal dissociation.” *Pattern Recognition* 122 (2022): 108213.
38. Cho, MyeongAh, Taeoh Kim, Woo Jin Kim, Suhwan Cho, and Sangyoun Lee. “Unsupervised video anomaly detection via normalizing flows with implicit latent features.” *Pattern Recognition* 129 (2022): 108703.
39. Wang, Yang, Tianying Liu, Jiaogen Zhou, and Jihong Guan. “Video anomaly detection based on spatio-temporal relationships among objects.” *Neurocomputing* 532 (2023): 141-151.
40. Wang, Zhiqiang, Xiaojing Gu, Jingyu Hu, and Xingsheng Gu. “Ensemble anomaly score for video anomaly detection using denoise diffusion model and motion filters.” *Neurocomputing* 553 (2023): 126589.
41. Wen, Xiaopeng, Huicheng Lai, Guxue Gao, Yang Xiao, Tongguan Wang, Zhenhong Jia, and Liejun Wang. “Video anomaly detection based on cross-frame prediction mechanism and spatio-temporal memory-enhanced pseudo-3D encoder.” *Engineering Applications of Artificial Intelligence* 126 (2023): 107057.
42. Shao, Wenhao, Praboda Rajapaksha, Yanyan Wei, Dun Li, Noel Crespi, and Zhigang Luo. “COVAD: Content-oriented video anomaly detection using a self-attention based deep learning model.” *Virtual Reality & Intelligent Hardware* 5, no. 1 (2023): 24-41.
43. Wang, Le, Junwen Tian, Sanping Zhou, Haoyue Shi, and Gang Hua. “Memory-augmented appearance-motion network for video anomaly detection.“ *Pattern Recognition* 138 (2023): 109335.
44. Amjadian, Ehsan, Prayogo, Nicholas, McDonnell, Serena, Smyth, Cathal, and Abid, Muhammad. “Attended over Distributed Specificity for Information Extraction in Cybersecurity.” In *2021 IEEE Aerospace Conference*, pp. 1-12. IEEE, 2021. doi: 10.1109/AERO50100.2021.9438369.
45. Kommanduri, Rangachary, and Mrinmoy Ghorai. “Bi-READ: Bi-Residual AutoEncoder based feature enhancement for video anomaly detection.”*Journal of Visual Communication and Image Representation* (2023): 103860.