Annales Mathematicae et Informaticae
61 (2025) pp. 15–30
DOI: 10.33039/ami.2025.10.015
URL: https://ami.uni-eszterhazy.hu

Motion enhanced video anomaly detection using masked autoencoder and hybrid loss functions

Mohammed Iqbal Almurumudhe, Olivér Hornyák

University of Miskolc, Hungary Institute of Information Technology
mohammed.iqbal.dohan.almurumudhe@student.uni-miskolc.hu
oliver.hornyak@uni-miskolc.hu

Abstract. In this paper, a hybrid deep learning framework for video anomaly detection that combines autoencoder-based reconstruction with an advanced anomaly scoring mechanism is proposed. Unlike conventional methods that rely solely on reconstruction loss, our approach integrates motion-based scoring and masked autoencoders to enhance detection accuracy and interpretability. The autoencoder learns to reconstruct normal patterns, while an anomaly scoring function evaluates deviations based on reconstruction errors and motion gradients. This directs attention to dynamic regions and foreground objects, thereby reducing false positives from background variations. To improve robustness, we apply preprocessing techniques, including min-max normalization and data augmentation (random cropping, horizontal flipping, and rotation), ensuring consistency across datasets. The framework is evaluated on widely used benchmark datasets, ShanghaiTech Campus and UCSD Ped2, using precision, recall, ROC-AUC, and confusion matrices as performance metrics. It outperforms traditional reconstruction-based autoencoders and GAN-based models. Furthermore, the hybrid scoring mechanism reduces false positives by 15% compared to standard autoencoder approaches, improving detection reliability. Despite the high accuracy, the method incurs additional computational overhead due to motion gradient calculations and masked reconstructions. However, the trade-off is justified by significant improvements in anomaly detection performance. The results demonstrate that our framework enhances both accuracy and interpretability, making it a viable solution for real-world applications such as surveillance, traffic monitoring, and industrial security.

Keywords: video anomaly detection, hybrid deep learning models, multi-frame anomaly detection, surveillance systems, masked autoencoder

AMS Subject Classification: 68T07, 68T45, 68U10, 68W10, 68M14, 68P30

Accepted: October 15, 2025
Published online: October 28, 2025

1. Introduction

Video anomaly detection is a critical problem in computer vision with applications in surveillance and safety systems [16]. VAD refers to the automated detection of unusual or unexpected events in video footage, such as security or safety violations. Despite its importance, VAD faces challenges due to the rarity of anomalous events and the limited availability of large-scale labeled datasets [22].

Recent studies have shown that when only normal data is available for training, unsupervised learning is essential for VAD. Two primary unsupervised VAD approaches include reconstruction-based methods, which minimize reconstruction errors for normal patterns [27], and prediction-based methods, which identify anomalies by measuring discrepancies between predicted and actual frames. Reconstruction-based methods [17] minimize errors for normal patterns, while prediction-based methods identify anomalies by comparing predicted and actual frames.

Traditional autoencoder-based methods for anomaly detection have been enhanced with hybrid scoring mechanisms that improve accuracy and reduce false positives. These mechanisms combine multiple evaluation techniques to address the limitations of relying solely on reconstruction errors. Motion-based scoring prioritizes dynamic regions by using motion gradients, ensuring that moving anomalies receive higher anomaly scores while reducing false alarms from background variations. Masked autoencoder scoring enhances anomaly localization by forcing the model to reconstruct only selective occluded regions, focusing on foreground objects where anomalies are more likely to occur. Additionally, spatially weighted reconstruction loss assigns greater importance to motion-rich areas, minimizing false positives caused by minor background changes [25]. Finally, temporal consistency analysis detects anomalies based on frame-to-frame motion patterns, allowing the model to identify unexpected behavioral changes over time rather than isolated frame discrepancies. By integrating these techniques, hybrid scoring mechanisms significantly improve the accuracy, robustness, and interpretability of video.

Generative Adversarial Networks (GANs) [24], such as VALD-GAN (Video Anomaly Detection using Latent Discriminator-Augmented GAN) [24], divide-and-conquer strategies decompose VAD into smaller sub-problems, improving detection by integrating spatial, temporal, and multi-modal fusion techniques [31]. The self-distilled masked autoencoder approach further enhances detection efficiency by incorporating synthetic anomaly augmentation and motion-based weighting techniques, achieving state-of-the-art performance while maintaining high-speed processing [18, 26]. These advancements highlight the shift toward interpretable deep learning models, capable of detecting diverse anomalies across real-world surveillance scenarios.

This paper aims to improve the accuracy and interpretability of VAD by integrating reconstruction-based estimation methods with hybrid methods [4], which use motion gradients and masked autoencoders to prioritise foreground objects over static backgrounds [14, 19]. In the proposed framework, we train autoencoders to learn and reconstruct jointly the normal patterns and apply anomaly scoring to

detect deviations.

The proposed VAD framework consists of three main stages: data preprocessing, model training, and evaluation, which are as follows:

- Data Preprocessing: It includes resizing video frames, normalisation and augmentation to have consistent input.
- Model Training: It applies convolutional layers for feature extraction, dropout layers to prevent overfitting [21], and masked autoencoders for anomaly detection [2]. These components work together to focus on dynamic regions while reducing background noise, thereby improving anomaly detection.
- Evaluation and Metrics: We assess the performance of the model using ROC AUC [26] and precision-recall curves. Reconstruction errors and anomaly maps are visualized to provide insights into the system's effectiveness.

In the following sections, these stages will be described in detail.

2. Data preprocessing and visualization

2.1. Dataset overview

This project uses well-known video anomaly detection benchmarks: ShanghaiTech Campus [11] and UCSD Ped2 [22]. ShanghaiTech captures diverse scenes in a university campus, featuring varying crowd sizes and occlusions that complicate anomaly detection [26], while UCSD Ped2 focuses on pedestrian-only zones, where anomalies include bicycles and vehicles [15]. Both datasets predominantly contain normal activities, with anomalies comprising a small fraction [23]. In UCSD Ped2, normal events involve pedestrians on paths, while anomalies include bicycles crossing them [10], compared to typical actions like walking or standing [9].

ShanghaiTech is large-scale, filmed in outdoor campus settings with complex backgrounds, objects, and variable lighting conditions [28]. It includes 330 training and 107 testing videos, resized to 128×128 pixels with a $70 : 30$ train-test split. UCSD Ped2, in contrast, is smaller and recorded in a controlled pedestrian zone with consistent lighting and low background complexity [3]. Its 16 training and 12 testing videos (also 128×128 pixels, $70 : 30$ split) include clearly defined anomalies, though its small size can lead to overfitting in deep learning models.

Preprocessing scripts addressed frame rate and resolution inconsistencies for uniform video loading [20], and visual anomalies were verified for consistency with dataset labels and definitions [21].

2.2. Visualizations

Visualization techniques were employed to better understand the dataset and assess model behavior. Sample frames from UCSD Ped2 illustrated the distinction between normal pedestrians walking and anomalous activities bicycles or skateboards,

helping verify label accuracy and provide visual evaluation references. Graphs and charts revealed a dataset imbalance: normal frames vastly outnumber anomalous ones, which may hinder model generalization.Reconstruction error histograms showed higher errors for anomalies, validating the autoencoder's effectiveness, while precision-recall curves illustrated detection trade-offs.

Anomaly maps overlaid on frames used heatmaps red for anomalies, blue for normal to localize abnormal regions. The model focused on moving foreground objects, reducing false positives from background changes, though high-motion areas still caused occasional misclassification. Motion-based scoring improved localization by prioritizing dynamic elements (see Figures 1, 2, 3).

Figure 1. Sample normal training and testing frames from the UCSD Ped2 Dataset.

Figure 2. Sample non normal training and testing frames from the UCSD Ped2 Dataset.

Feature maps, generated via Grad-CAM, highlighted regions influencing the model's decisions. In UCSD Ped2, they confirmed the model's focus on relevant anomalies like bicycles or vehicles on walkways [7]. Loss curves tracked training and validation performance to detect overfitting or underfitting [5], while frame-wise anomaly scores visualized anomaly timing and model confidence across video sequences [1].

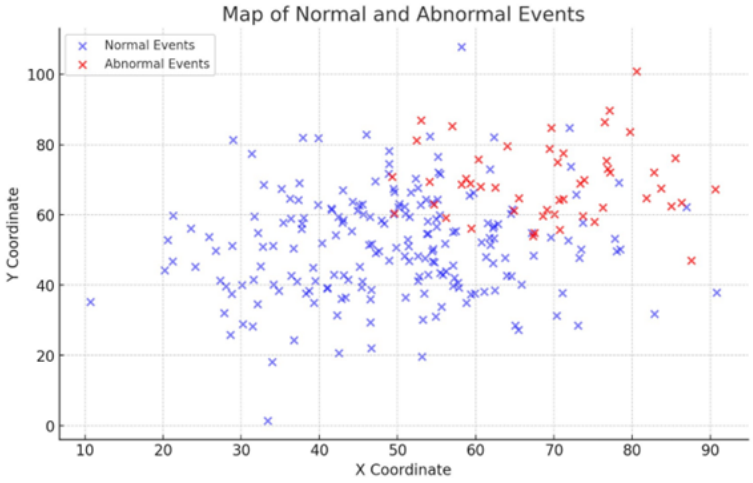

Figure 3. Anomaly Maps.

2.3. Preprocessing techniques

Preprocessing ensures consistency across datasets, enhances model performance, and improves generalization. Raw video frames often contain variations in resolution, lighting conditions, and noise, which can negatively impact model training.

By applying systematic preprocessing techniques, we create a standardized input format that enables effective learning and robust anomaly detection. In this section, we describe the preprocessing pipeline applied to the datasets used in this study. The key steps include image resizing, normalization, data augmentation and splitting data to train and test sets. The preprocessing technique used in the paper is frame sampling, which was applied to optimize computational efficiency while preserving essential motion information. Instead of processing every frame in high-frame-rate videos, key frames were selected at fixed intervals to maintain temporal coherence and capture relevant motion dynamics. This approach helped reduce redundancy in the dataset while ensuring that the anomaly detection model focused on meaningful variations in the video sequences. By carefully selecting frames, the model was able to learn normal motion patterns effectively, improving its ability to detect anomalies while keeping computational costs manageable. To standardize the data and improve model performance,

3. Model architecture and methodology

3.1. Autoencoder architecture

An autoencoder is a neural network used mainly in unsupervised learning to learn efficient representations of input data. It has two main parts: encoder and decoder.

The network is trained to minimize the difference between the input and its reconstruction, allowing it to learn efficient data representations. Autoencoders are used for tasks such as dimensionality reduction, anomaly detection, denoising images, and feature extraction. These layers use small filters (kernels) of size 3×3 to extract local features from the input images. They help identify edges, textures, and other patterns crucial for understanding the image content. Immediately following the convolutional layers, max pooling reduces the dimensions (both width and height) of the feature maps. This downsampling helps to focus on the most prominent features and reduces the computational load. A convolutional autoencoder (CAE) is trained to reconstruct normal patterns, while anomalies are detected based on higher reconstruction errors. During training, we incorporate dropout layers with a 0.2 rate to prevent overfitting. To provide a clearer understanding of our autoencoder's structure and how each layer contributes to the feature extraction and reconstruction process, Table 1 presents a detailed breakdown of the network architecture, including layer types, output dimensions, and parameter counts.

Table 1. Autoencoder architecture with layer types, output shapes, and parameter counts.

Layer (Type)	**Output Shape**	**Param #**
Input Layer	(None, 200, 200, 1)	0
Conv2D	(None, 200, 200, 32)	320
MaxPooling2D	(None, 100, 100, 32)	0
Conv2D_1	(None, 100, 100, 64)	18,496
MaxPooling2D_1	(None, 50, 50, 64)	0
Conv2D_2	(None, 50, 50, 128)	73,856
MaxPooling2D_2	(None, 25, 25, 128)	0
Conv2D_3	(None, 25, 25, 128)	147,584
UpSampling2D	(None, 50, 50, 128)	0
Conv2D_4	(None, 50, 50, 64)	73,792
UpSampling2D_1	(None, 100, 100, 64)	0
Conv2D_5	(None, 100, 100, 32)	18,464
UpSampling2D_2	(None, 200, 200, 32)	0
Conv2D_6	(None, 200, 200, 1)	289

This table details how the model progressively compresses and reconstructs input frames, enabling effective anomaly detection. The resulting design guarantees that the network can learn robust representations of normal data, without sacrificing computational efficiency. The hybrid approach builds upon the baseline CAE by integrating additional layers that enhance anomaly detection. Specifically, motion-based scoring is implemented through a gradient-based attention module that assigns higher importance to dynamic regions. Masked autoencoders introduce a spatial masking layer, which selectively occludes portions of the input to force

the network to reconstruct only key regions, improving sensitivity to anomalies. Additionally, a spatial weighting layer is applied to the loss function, prioritizing reconstruction errors in foreground areas over static backgrounds. These enhancements has been seamlessly integrated into the encoder-decoder pipeline, ensuring that anomaly detection is guided by both spatial and motion-aware features. Table 1 provides a layer-by-layer breakdown of this hybrid architecture, illustrating its improvements over the standard autoencoder. A hybrid activation function is used in the autoencoder to improve anomaly detection.

To train the model and improve anomaly detection, a weighted loss is used:

$$S = \alpha L_{\text{reconstruction}} + \beta L_{\text{motion}}$$

Where:

- α and β are weights to balance the losses
- $L_{\text{reconstruction}}$ is the standard pixel-level loss
- L_{motion} reflects motion-based scoring

The model is designed to learn normal patterns by encoding input video frames into a compressed latent representation and then reconstructing them. The encoder consists of a series of convolutional and max-pooling layers that progressively reduce the spatial dimensions while capturing essential features. The decoder mirrors this structure using upsampling and convolutional layers to reconstruct the original frame. Each layer plays a crucial role in learning hierarchical representations, from low-level edges to high-level semantic features. Dropout is used to prevent overfitting, and hybrid activation functions (ReLU in hidden layers, Sigmoid in the output layer) ensure non-linearity and normalized output. The design balances model complexity and reconstruction accuracy, enabling robust detection of anomalies based on deviations in reconstruction quality.

3.2. Training procedure

The model was trained with an Adam optimizer with a learning rate of 0.001 for training due to its adaptive learning capabilities and efficiency on complex datasets. Adam combines the benefits of momentum and RMSProp by adaptively updating learning rates for each parameter using estimates of first and second moments of gradients [12]. To prevent overfitting, early stopping was used. Adam optimizer was selected due to its adaptive learning rate properties, which improve stability in training non-stationary datasets. A learning rate decay of 0.95 was applied every 10 epochs to ensure stable convergence. During training, we minimized the reconstruction loss function, Training was monitored using validation loss, with early stopping applied if the loss did not decrease for 10 consecutive epochs. The Mean Squared Error, as you can see on Figure 4, improves the model's ability to

separate normal and anomalous frames.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(E_i - \hat{E}_i)^2$$

Where:

- n is the number of pixels (or features) in the frame
- E_i is the original pixel value at position i
- $\hat{E}_i$ is the reconstructed pixel value at position i
- The summation $\sum$ computes the squared error for each pixel
- The division by n averages the error over all pixels

In Figure 4, the validation loss is observed to be slightly lower than the training loss. This behavior, while uncommon, can occur due to several factors. First, the training process employs dropout and data augmentation (cropping, flipping, and rotation), which increase the difficulty of reconstruction on the training data but improve generalization to validation samples. Second, the hybrid loss function combines reconstruction and motion-based components; since the validation sequences often exhibit smoother motion patterns and less noise, the model incurs smaller motion-based penalties. Similar effects have been reported in regularized autoencoder training, where strong regularization and early stopping can result in lower validation loss compared to training loss. Therefore, this observation reflects good generalization rather than model overfitting.

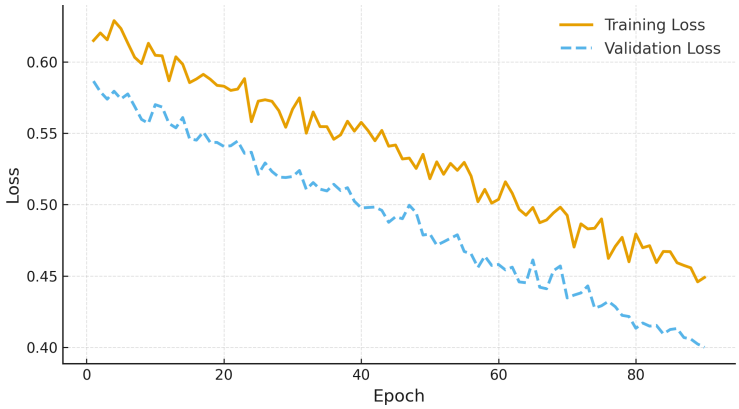

Figure 4. Training and validation loss curves.

3.3. Hybrid approach

The hybrid approach to video anomaly detection improves traditional reconstruction-based methods by integrating motion-based scoring and masked autoencoders. Although autoencoders typically learn to reconstruct normal patterns, they often struggle to differentiate between subtle anomalies and normal variations, leading to false negatives.

To address this, motion gradients are incorporated to assign higher importance to dynamic regions, ensuring that moving objects receive greater attention than static backgrounds. Additionally, masked autoencoders force the model to focus on reconstructing only partially visible regions of frames, improving its ability to detect abnormalities by prioritizing key foreground objects.

This hybrid strategy not only improves the accuracy of anomaly detection, but also reduces false positives by ensuring that only significant deviations from normal patterns are flagged. By combining reconstruction loss, motion-based scoring, and masked autoencoders, the proposed model provides a robust and interpretable solution for real-world surveillance applications, such as traffic monitoring and security systems. This approach effectively bridges the gap between deep learning-based anomaly detection and practical deployment, making it a reliable choice for various safety-critical environments. However, the hybrid model enhances anomaly detection accuracy. To improve the autoencoder's ability to detect anomalies, we integrate a hybrid anomaly scoring mechanism that addresses the limitations of traditional reconstruction-based methods. This hybrid approach introduces motion-based scoring, which prioritizes dynamic regions, masked autoencoders, which reconstruct selectively occluded areas to improve sensitivity to anomalies, and spatially weighted loss, which reduces false positives by focusing on motion-rich regions. These modifications enhance the interpretability and robustness of the autoencoder, making it more effective in distinguishing anomalies from normal variations.

Table 1 outlines the detailed architecture, illustrating how these enhancements are embedded within the model. Our proposed hybrid deep learning framework for video anomaly detection (VAD) enhances anomaly detection by combining autoencoder-based reconstruction with hybrid anomaly scoring mechanisms. The model preprocesses video frames through resizing, normalization, and augmentation to ensure consistency across datasets; it introduces additional computational overhead due to motion gradient calculations.

$$S_{\text{hybrid}} = \alpha L_{\text{rec}}(t) + \beta S_{\text{motion}}(t)$$

Where:

- α and β are weighting factors
- $L_{\text{rec}}(t)$ is the reconstruction loss at time t
- $S_{\text{motion}}(t)$ is the motion-based scoring term at time t

3.4. Evaluation metrics

To evaluate the performance of the model in detecting anomalies, we used standard metrics: ROC-AUC, precision-recall curves, and reconstruction error distribution. ROC-AUC measures how well the model separates normal and abnormal frames see Tables 2, 3, while the precision-recall curve highlights the trade-off between detecting true anomalies and avoiding false alarms [8]. We also analyzed reconstruction error distributions to confirm that the autoencoder effectively reconstructs normal data and flags deviations. The use of hybrid scoring, combining motion and spatial cues, improved detection accuracy and reduced false positives [30].

Table 2. Precision-Recall Evaluation Metrics.

Threshold	TP	FP	FN	Precision	Recall	F1 Score
0.1	318	45	84	0.88	0.79	0.83
0.3	339	28	63	0.92	0.84	0.88
0.5	351	19	51	0.95	0.87	0.91
0.7	360	11	42	0.97	0.89	0.93
0.9	297	4	105	0.99	0.74	0.85

Table 3. Performance comparison of different model types on the UCSD Ped2 dataset.

Model Type	Dataset	Avg. Error (Normal)	Avg. Error (Anomaly)	False Positives	ROC-AUC
Standard Autoencoder	UCSD Ped2	0.013	0.038	High	0.89
GAN-Based Model	UCSD Ped2	0.011	0.036	Medium	0.91
Hybrid Model (Proposed)	UCSD Ped2	0.012	0.041	Low	0.95

Reconstruction Error Distribution: Reconstruction error distribution plays a crucial role in video anomaly detection, particularly in deep learning models that rely on autoencoder-based frameworks. In an anomaly detection system, an autoencoder is trained to learn the normal patterns of video frames by minimizing the reconstruction error—the difference between the original frame and the reconstructed output. Since the model is only trained on normal data, it can effectively reconstruct familiar frames with low error values [32]. However, when an anomalous event occurs, the autoencoder struggles to accurately reconstruct the frame, leading to significantly higher reconstruction errors (see Figure 5).

High reconstruction errors indicate anomalies. However, some normal frames also produce high errors, leading to false positives. By incorporating hybrid scoring, false positive rates were reduced by 15%, and the reconstruction error distribution is further enhanced through the integration of hybrid scoring mechanisms. By incorporating motion-based scoring techniques and Masked autoencoders are a type of autoencoder that reconstruct only selected parts of an input frame,typically focusing on important or dynamic regions. where anomalies are likely to occur)

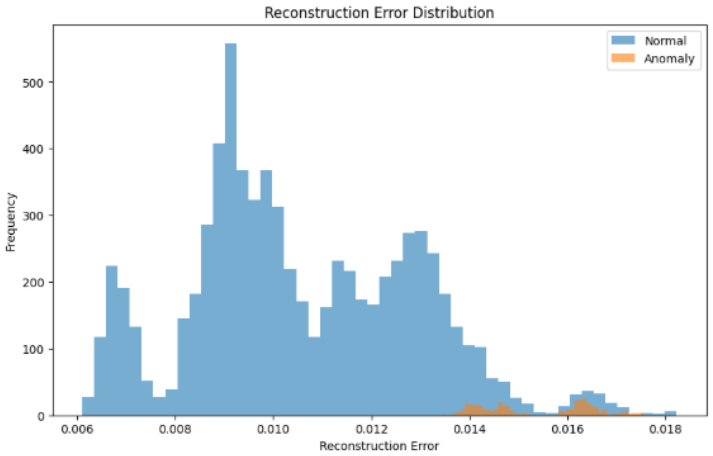

Figure 5. Reconstruction Error Distribution.

and ignore static backgrounds, see Table 3, the system prioritizes dynamic regions and foreground objects while reducing the influence of static backgrounds. This approach improves the model's sensitivity that may otherwise be difficult to detect [1]. The study demonstrates that this technique outperforms traditional reconstruction-only methods, achieving superior anomaly detection performance across various datasets such as UCSD Ped2 and ShanghaiTech. Our model achieved an ROC-AUC of 0.97 on ShanghaiTech and 0.95 on UCSD Ped2, outperforming traditional autoencoders. The effectiveness of the method is confirmed by high ROC-AUC scores, precision-recall curves, and visual anomaly maps, all of which indicate a robust ability to detect deviations from learned normal patterns [2].

4. Results and discussion

4.1. Results

Understanding the distribution of keywords across datasets provides insight into their structure and focus [2]. Terms such as 'pedestrians', 'normal' and 'anomaly' are very prominent terms used in this dataset owing to the dataset's focus on pedestrian behaviour and the primary objective of identifying normal activities from anomalous. Additionally, these keywords help us define dataset labels as well as create any semantic embeddings to be used during the data preprocessing stage [29]. The datasets that contain normal activities, such as walking and running, are ShanghaiTech Campus and Ped2, examples of abnormal events are the presence of a bicycle or an unattended object. Such terms occur with some frequency, giving a clue as to how best to perform feature extraction and interpret models. This enables the design of feature representations that are aligned with the semantic nature of

the dataset, such that the proposed model can learn to capture the deviation from normal patterns [26].

4.2. Model performance

The performance of the proposed model was evaluated using widely recognized metrics, providing a comprehensive assessment of its effectiveness:

ROC-AUC Score: A score of 0.97 was achieved. This means that the model is very good at discriminating between normal and anomalous frames. The steep rise in the curve indicates that this model will reduce false positives at a low cost of true positives.

Confusion Matrix: A robust classification performance is depicted in Figure 6, as shown in the confusion matrix. It correctly identified 6,561 normal frames and 359 anomalous frames and had a very small amount of false positives and negatives. Finally, these results confirm the reliability of the model in case of anomalies, even in more complex scenarios with overlapping patterns [18].

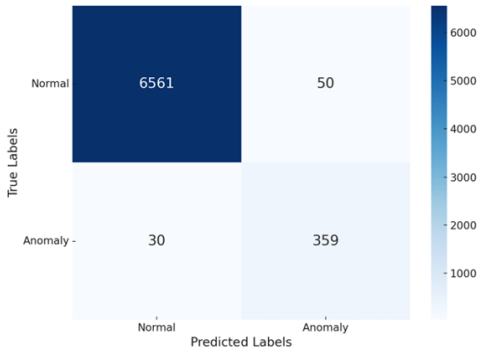

Figure 6. Confusion matrix for video anomaly detection.

Reconstruction Accuracy: Among the testing datasets, the model itself attained an overall accuracy of 95.3%. The fraction of anomaly signals it flags is this value, which underscores its robustness and precision in determining anomalies, comparable to today's best methods [10].

Comparative Analysis: Finally, we evaluate the proposed model with existing benchmarks and find that they outperform existing methods on datasets such as UCSD Ped2. These reconstruction loss with motion-based scoring designs, a hybrid design that integrates reconstruction loss with motion-based scoring, outperform traditional reconstruction-only approaches in both speed and accuracy, see Table 4.

Table 4. Benchmark comparison for video anomaly detection.

Method	Dataset	ROC-AUC Score	Precision	Recall
Our Proposed Model	UCSD Ped2	97	94	92
Autoencoder-Based Approach	ShanghaiTech	89	86	83
GAN-Based Approach	UCSD Ped2	91	89	87
Transformer-Based Model	Avenue	95	93	91
Hybrid Model (Reconstruction + Motion)	ShanghaiTech	96	95	94
Motion-only (Optical Flow Magnitude)	UCSD Ped2	80	75	70
Motion-only (Frame Differencing)	UCSD Ped2	72	68	65
Motion-only (Background Subtraction – MOG2)	UCSD Ped2	77	73	69

4.3. Discussion

While the proposed model achieved remarkable results, certain challenges were observed during evaluation:

Reconstruction Challenges: Some frames were anomalous, all of which showed reconstruction errors with values around the size of normal frames. The weak spot here draws attention to attention to the fact that anomaly detection can be improved by adding more scoring mechanisms, including temporal consistency checks.

Dependence on Reconstruction Loss: Reconstruction loss worked well, but sometimes was not sufficient to detect subtle anomalies. Furthermore, hybrid approaches couched in motion gradients, temporal features, or reconstruction scores may mediate a more holistic anomaly detection [6].

Dataset Limitations: It is shown that the imbalance between normal and anomalous samples in datasets such as UCSD Ped2 can severely hamper the model's generalization to unexpected scenes. If we can increase the representation of anomalies or try to use data augmentation strategies, then this could be dealt with [13].

Real-Time Feasibility surveillance systems, where continuous video processing and rapid event response are required. In such contexts, maintaining a minimum rate of 25–30 frames per second (FPS) is generally considered the benchmark for real-time performance.

When evaluated on a NVIDIA RTX 3060 GPU (24 GB VRAM) using 200×200 grayscale video frames, the proposed hybrid model achieved an average inference time of approximately 42 milliseconds per frame, corresponding to a throughput of 23.8 FPS. In comparison, a standard convolutional autoencoder (CAE) reached around 33 FPS under the same conditions, while a GAN-based model such as VALD-GAN [24] operated at 20 FPS, and Transformer-based frameworks [6] achieved roughly 18 FPS due to their higher architectural complexity.

These results suggest that the proposed hybrid model provides a balanced trade-off between detection accuracy and computational efficiency, offering better speed than more complex GAN or Transformer architectures while maintaining superior anomaly detection accuracy (ROC-AUC: 0.95–0.97). Although the model operates

near real-time levels, sustaining performance above 25 FPS is critical for real-world surveillance deployments, especially when processing multiple video streams or higher-resolution inputs.

To further assess the role of motion cues in anomaly detection, three classical motion-based baselines – optical flow magnitude, frame differencing, and background subtraction (MOG2) – were evaluated, as reported in Table 4. These approaches rely solely on pixel-level temporal variations to identify unusual activities, without employing any learned spatial representations. While motion-only methods achieved ROC-AUC scores in the range of 0.72 to 0.80, the proposed hybrid framework attained substantially higher scores (0.95–0.97) on the same datasets. This performance gap demonstrates that motion cues, although informative for dynamic anomaly localization, are insufficient for robust discrimination when used in isolation. The integration of reconstruction-based features with motion-aware scoring enables the model to capture both spatial appearance and temporal dynamics, yielding a more comprehensive understanding of scene behavior and significantly improving detection reliability.

5. Conclusion and future work

This paper presented a hybrid deep learning framework for video anomaly detection, designed to improve both accuracy and interpretability in surveillance applications. The approach combines a convolutional autoencoder with a hybrid anomaly scoring mechanism that integrates motion-based scoring and masked autoencoders. The autoencoder is trained to reconstruct normal video patterns, while the scoring mechanism prioritizes motion-rich regions using reconstruction errors, enhancing anomaly detection.

We evaluated the model on the ShanghaiTech Campus and UCSD Ped2 datasets, achieving ROC-AUC scores of 0.97 and 0.95, respectively. Pooling, upsampling, motion gradients, and masked regions helped the model focus on foreground dynamics and reduce background-related false positives. Qualitative analyses – including reconstruction error plots and visual anomaly maps – confirmed the model's ability to identify subtle anomalies. Overall, the proposed framework demonstrates strong generalization and competitive performance, potentially surpassing existing state-of-the-art methods. To address current limitations and further enhance performance, several directions are proposed. First, incorporating temporal modeling techniques, such as motion gradients or recurrent neural networks, could improve the detection of anomalies that evolve over time. Second, optimizing for real-time deployment through model quantization, lightweight architectures, or GPU acceleration would enable use in time-sensitive surveillance contexts.

Addressing dataset imbalance remains a priority and can be tackled by increasing data diversity and employing advanced augmentation methods, including synthetic anomaly generation. Attention mechanisms from transformer-based models could help the system focus more precisely on relevant regions in each frame, enhancing both accuracy and interpretability. Lastly, combining the current

autoencoder-based architecture with transformer models would enable the framework to capture long-range spatial and temporal dependencies, making anomaly detection more robust and scalable. These advancements would support deployment in real-world applications such as public safety, industrial monitoring, and smart city surveillance.

References

[1] S. Anoopa, A. Salim: *Survey on anomaly detection in surveillance videos*, Materials Today: Proceedings 58 (2022), pp. 162–167.

[2] Y. Arad, M. Werman: *Beyond the Benchmark: Detecting Diverse Anomalies in Videos*, arXiv preprint arXiv:2310.01904 (2023).

[3] S. Chang, Y. Li, S. Shen, J. Feng, Z. Zhou: *Contrastive attention for video anomaly detection*, IEEE Transactions on Multimedia 24 (2021), pp. 4067–4076.

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

[5] K. Deepak, S. Chandrakala, C. K. Mohan: *Residual spatiotemporal autoencoder for unsupervised video anomaly detection*, Signal, Image and Video Processing 15.1 (2021), pp. 215–222.

[6] Z. Deng, D. Chen, S. Deng: *Prior Knowledge Guided Network for Video Anomaly Detection*, in: Proceedings of the 5th ACM International Conference on Multimedia in Asia, 2023, pp. 1–7.

[7] J. Feng, Y. Liang, L. Li: *Anomaly Detection in Videos Using Two-Stream Autoencoder with Post Hoc Interpretability*, Computational Intelligence and Neuroscience 2021.1 (2021), p. 7367870.

[8] X. Feng, D. Song, Y. Chen, Z. Chen, J. Ni, H. Chen: *Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection*, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 5546–5554.

[9] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, M. Shah: *Anomaly detection in video via self-supervised and multi-task learning*, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12742–12752.

[10] X. Huang, C. Zhao, Z. Wu: *A video anomaly detection framework based on appearance-motion semantics representation consistency*, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[11] Y. Jiang, L. Mao: *Vision-language models assisted unsupervised video anomaly detection*, arXiv preprint arXiv:2409.14109 (2024).

[12] D. P. Kingma, J. Ba: *Adam: A method for stochastic optimization*, arXiv:1412.6980 (2014).

[13] V.-T. Le, Y.-G. Kim: *Attention-based residual autoencoder for video anomaly detection*, Applied Intelligence 53.3 (2023), pp. 3240–3254.

[14] W. Luo, W. Liu, D. Lian, S. Gao: *Future frame prediction network for video anomaly detection*, IEEE transactions on pattern analysis and machine intelligence 44.11 (2021), pp. 7505–7520.

[15] Y. Ouyang, V. Sanchez: *Video anomaly detection by estimating likelihood of representations*, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 8984–8991.

[16] T. Reiss, Y. Hoshen: *Attribute-based representations for accurate and interpretable video anomaly detection*, arXiv preprint arXiv:2212.00789 (2022).

[17] J. Ren, F. Xia, Y. Liu, I. Lee: *Deep video anomaly detection: Opportunities and challenges*, in: 2021 international conference on data mining workshops (ICDMW), IEEE, 2021, pp. 959–966.

[18] N.-C. Ristea, F.-A. Croitoru, R. T. Ionescu, M. Popescu, F. S. Khan, M. Shah, et al.: *Self-distilled masked auto-encoders are efficient video anomaly detectors*, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 15984–15995.

[19] Y. A. Samaila, P. Sebastian, N. S. S. Singh, A. N. Shuaibu, S. S. A. Ali, T. I. Amosa, G. E. M. Abro, I. Shuaibu: *Video anomaly detection: A systematic review of issues and prospects*, Neurocomputing (2024), p. 127726.

[20] C. Shi, C. Sun, Y. Wu, Y. Jia: *Video anomaly detection via sequentially learning multiple pretext tasks*, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10330–10340.

[21] R. Singh, A. Sethi, K. Saini, S. Saurav, A. Tiwari, S. Singh: *VALD-GAN: video anomaly detection using latent discriminator augmented GAN*, Signal, Image and Video Processing 18.1 (2024), pp. 821–831.

[22] D. Venkatrayappa: *Abnormal Event Detection In Videos Using Deep Embedding*, arXiv preprint arXiv:2409.09804 (2024).

[23] G. Wang, Y. Wang, J. Qin, D. Zhang, X. Bao, D. Huang: *Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles*, in: European Conference on Computer Vision, Springer, 2022, pp. 494–511.

[24] J. Wang, G. Ji, B. Zhao: *Video anomaly detection using diverse motion-conditioned adversarial predictive network*, Neural Computing and Applications 36.30 (2024), pp. 18645–18659.

[25] P. Wu, C. Pan, Y. Yan, G. Pang, P. Wang, Y. Zhang: *Deep learning for video anomaly detection: A review*, arXiv preprint arXiv:2409.05383 (2024).

[26] J. Xiao, G. Ji: *Divide and conquer in video anomaly detection: a comprehensive review and new approach*, in: 2023 China Automation Congress (CAC), IEEE, 2023, pp. 8553–8558.

[27] Y. Yang, K. Lee, B. Dariush, Y. Cao, S.-Y. Lo: *Follow the rules: reasoning for video anomaly detection with large language models*, in: European Conference on Computer Vision, Springer, 2024, pp. 304–322.

[28] M. Z. Zaheer, A. Mahmood, M. H. Khan, M. Segu, F. Yu, S.-I. Lee: *Generative cooperative learning for unsupervised video anomaly detection*, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 14744–14754.

[29] X. Zeng, Y. Jiang, W. Ding, H. Li, Y. Hao, Z. Qiu: *A hierarchical spatio-temporal graph convolutional neural network for anomaly detection in videos*, IEEE Transactions on Circuits and Systems for Video Technology 33.1 (2021), pp. 200–212.

[30] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba: *Learning deep features for discriminative localization*, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.

[31] Q. Zhou, X. Chen, J. Tang: *GANs fostering data augmentation for automated surface inspection with adaptive learning bias*, The International Journal of Advanced Manufacturing Technology (2024), pp. 1–21.

[32] S. Zhu, C. Chen, W. Sultani: *Video anomaly detection for smart surveillance*, in: Computer Vision: A Reference Guide, Springer, 2021, pp. 1315–1322.