Multimedia Tools and Applications (2024) 83:78517–78534
https://doi.org/10.1007/s11042-024-18604-2

Anomaly detection in video surveillance: a supervised inception encoder approach

Rangachary Kommanduri[1] · **Mrinmoy Ghorai**[1]

Received: 23 October 2023 / Revised: 12 January 2024 / Accepted: 11 February 2024 /
Published online: 26 February 2024

Abstract
Unsupervised video anomaly detection approaches often demand complex models and substantial computational resources for effective performance. In contrast, we introduce a supervised and end-to-end trainable deep learning approach that leverages both performance and computational efficiency by harnessing frame-level annotated data. The framework begins with the utilization of an Inception encoder network in the initial stage to learn feature representations. Notably, the Inception network's proficiency in capturing intricate and high-level features in frames seamlessly extends to the analysis of video data. By using these extracted features, the model excels in identifying deviations from learned patterns, making it highly adept at detecting anomalies in video sequences. The subsequent stage involves a sequence of fully connected layers followed by a classifier that is responsible for classifying input frames as either normal or anomalous based on the extracted features. To thoroughly validate this methodology, extensive experiments are carried out on widely used benchmark datasets. These evaluations involved comprehensive comparisons with contemporary approaches in the field. The experimental findings consistently validate the efficacy and efficiency of the proposed approach, underscoring its outstanding accuracy in identifying anomalies. Additionally, the approach operates with significantly reduced computational overhead, rendering it an appealing solution for real-world applications that demand timely and precise anomaly detection.

Keywords Video surveillance · Inception encoder · Supervised learning · Anomaly detection

1 Introduction

In recent times, urban areas have witnessed a proliferation of surveillance camera installations aimed at capturing a wide array of real-time occurrences. The vast reservoir of video

✉ Rangachary Kommanduri
rangachary.k@iiits.in

Mrinmoy Ghorai
mrinmoy.ghorai@iiits.in

[1] Computer Science and Engineering, Indian Institute of Information Technology, Sri City 517646, Andhra Pradesh, India

data amassed from these installations has underscored the need for technology that not only identifies objects and their activities but also excels at detecting rare and peculiar anomalies or suspicious behavior within extensive volumes of data. The rapid and accurate identification of such anomalous events or objects holds the potential to enhance security measures, optimize operational processes, and facilitate informed decision-making for businesses. Video Anomaly Detection (VAD) has garnered increased attention within the realm of computer vision. Its significance lies in its applications, particularly in the domain of surveillance, where it plays a pivotal role in automatically identifying anomalies, thereby enhancing safety and security in various settings such as airports, traffic management, shopping complexes, and educational institutions.

The challenge in video anomaly detection arises from two intertwined factors: (i) the need to consider the contextual information when identifying anomalies [1], and (ii) the scarcity of abnormal training data. To illustrate these factors, let's consider a simple comparison: running on a regular street, which is considered normal, versus running inside a bank or any government organization, which is considered abnormal. This reliance on context leads to an expansive array of potential anomaly scenarios. However, collecting ample training data for various anomaly types presents a formidable challenge. Gathering video examples for certain types of anomalies, especially those involving potentially harmful or unethical situations, is impractical. As a result, the dependence on contextual information complicates the acquisition of adequate abnormal training data, making it a formidable task.

Video anomaly detection can be approached through various architectural paradigms, including reconstruction-based [2–5], prediction-based [6–8], and classification-based approaches [9–11]. In reconstruction-based methods, the primary objective is to reconstruct the input data, typically frames or sequences, using an autoencoder or a similar architecture. An autoencoder comprises an encoder network that compresses the input data into a lower-dimensional representation (latent space) and a decoder network that attempts to reconstruct the original input from this latent representation. During training, the model learns to minimize reconstruction errors. Anomalies are detected when the reconstruction error exceeds a predefined threshold. Reconstruction-based methods assume that anomalies deviate significantly from normal data and will result in higher reconstruction errors. However, these approaches typically suffer from a drawback wherein the models they rely on, such as autoencoders or Generative adversarial networks (GANs), necessitate retraining when introduced to new sets of normal training videos [12].

Models centered on predictions concentrate on anticipating forthcoming frames within a video sequence by relying on the information from the preceding frames. These architectures use recurrent neural networks (RNNs), GANs, or hybrid architectures to make predictions. Throughout the training process, the model acquires the capability to precisely predict the subsequent frame or event. When an anomaly occurs, it disrupts the regular temporal progression, causing prediction errors to increase. These methods detect anomalies by monitoring these prediction errors, often using thresholds or other anomaly-scoring mechanisms.

Classification-based approaches employ supervised or semi-supervised learning to classify each frame or sequence as normal or anomalous. In supervised settings, labeled data with annotations indicating anomalies are used for training. The model learns to discriminate between normal and anomalous data, typically using CNNs, RNNs, or more advanced architectures. Classification-based methods are versatile and can handle complex scenarios with well-defined anomaly labels. In this study, we present a supervised method that relies on the Inception architecture to classify frames as either normal or anomalous. Notably, our approach emphasizes a straightforward architectural design while maintaining efficiency in computational resource usage. The main contributions are summarised as follows:

a) To address the challenge of video anomaly detection within a deep neural network, an end-to-end trainable feature extractor based on the Inception encoder is introduced. This feature extractor effectively captures visual features that can be accurately classified by the classifier.
b) Through extensive experiments conducted on three datasets, the proposed framework demonstrates competitive performance when compared to state-of-the-art techniques.
c) We performed diverse ablation experiments, encompassing cross-dataset evaluations and assessments on datasets with noise and perturbed patches. This thorough analysis illustrates the capacity of our approach to generalize and remain robust, offering substantial confirmation of the method's effectiveness.

The paper's organization is presented in the following manner: In Section 2, an examination of prior research on video anomaly detection is presented, followed by Section 3, which delves into the specifics of the proposed model. The experimental configuration, outcomes, and insights about the suggested architecture across diverse datasets are elaborated upon in Section 4. Finally, Section 5 encapsulates the key conclusions derived from the study's outcomes.

2 Related work

In the literature, three distinct approaches have been identified for addressing the challenges in the video anomaly detection task. The first prevalent approach, as observed in works such as [2, 3, 5, 13–16], treats anomaly detection as an unsupervised method. These methods do not rely on any labeled data and are designed to detect anomalies solely based on the characteristics of the data itself. For instance, Hasan et al [2] leveraged spatial convolutional autoencoder to learn hand-crafted features for detecting anomalies. Chong et al [3] introduced a spatio-temporal convolutional autoencoder coupled with ConvLSTM to directly model input data for anomaly detection. To tackle sparse coding challenges, Mondal et al [13] employed mean optical flow as contextual information to capture global anomalies. Gandapur et al [14] introduced an end-to-end deep learning model combining Bi-GRU and CNN to detect and prevent criminal activities. Their approach extracts spatial, temporal, and local motion features, employs a focused bag approach, and utilizes a ranked-based loss for precise classification. In [15], Amin et al integrated quantum computing into CNN architectures for video anomaly detection. They introduced two unique architectures, the deep CNN and Javeria quantum CNN, to effectively address challenges such as occlusion, sparse anomalous events, and camera movements. Additionally, Park et al [16] introduced a lightweight autoencoder model using patch transformations to improve the learning of normal features by generating irregular patch cuboids within normal frame cuboids. To enhance the feature extraction and reusability authors in [5] introduced a novel ResNet-based architecture with long-short skip connections to extract spatial information and optical flow network for motion information. Despite the advantages, these unsupervised frameworks require complex architectures to extract spatiotemporal architectures which in turn consumes more computational resources and thus increases train and inference times. To balance both performance and computational complexity, researchers have explored both semi-supervised and supervised approaches.

Secondly, in semi-supervised learning, the training dataset may have incomplete or noisy labels. Instead of precise frame-level annotations, we might have higher-level annotations, video-level annotations, or labels indicating the presence of anomalies in a video without

specifying their exact locations. Authors in [17] used an unsupervised auto-encoder network in conjunction with the weakly supervised regression model to extract representative features from normal video clips, enhancing the discriminative characteristics between normal and abnormal events. Multiple instance learning (MIL) is frequently utilized in weakly supervised methodologies. For instance, [18] pioneered an approach based on MIL, considering a video as a ""bag" comprising numerous "snippets" treated as individual instances. They applied a ranking loss within the MIL framework to enhance the differentiation between the highest-scoring instances in positive and negative bags. Building upon this framework, Zhu et al [19] further improved it by incorporating motion-aware features. Moreover, Zhong and colleagues [20] tackled the problem of Voice Activity Detection (VAD) by treating it as a challenge that involves clean data labeled with noise. They employed a graph convolution classifier to capture temporal contexts. Weakly supervised methods may struggle to provide precise anomaly localization within frames or clips and handle diverse and complex anomalies due to weak annotations, potentially leading to false positives or missed detections.

The third category of methods belongs to supervised learning, where models are trained on datasets encompassing both normal and anomalous video sequences, with each frame or segment labeled as either "normal" or "anomalous." Despite achieving remarkable accuracy in anomaly detection, these supervised approaches often face a significant challenge in acquiring a substantial amount of labeled training data, a process that can be both costly and time-intensive [21]. Nevertheless, supervised methods offer distinct advantages, delivering precise predictions and high accuracy. Furthermore, they demonstrate efficiency by demanding lower computational resources and less time compared to unsupervised methods. Particularly in scenarios with a well-defined and consistent anomaly type, the deployment of these supervised models in real-time applications emerges as a practical and effective solution.

In their work [22], Zhou et al proposed a novel approach for abnormal behavior detection in crowded scenes, integrating anomaly detection with spatial-temporal CNN. Additionally, Hinami et al [23] introduced a unified framework that combines generic and environment-specific knowledge to address joint abnormal event detection and recounting. The authors of [24] presented a feed-forward neural network leveraging features from Local Binary Patterns (LBP) and employing K-means labeling for abnormality detection. LBP features are extracted and utilized in the neural network, relying on characteristics labeled using K-means clustering. Ma et al [25] proposed a partially supervised method for video abnormal event detection and localization, utilizing only normal samples for training. This approach employs variational autoencoders (VAE) to constrain hidden layer representations of normal samples to a Gaussian distribution, enabling the identification of abnormal samples with lower probabilities within this distribution. In their work [10, 26], the authors introduced a supervised approach to improving the efficiency of detecting abnormal events/objects using transfer learning approach. They employed pre-trained models like VGG16 and MobileNetV2 as feature extractors, integrating a fine-tuned network to classify the input video frames as normal and anomalous.

3 Our method

Problem statement- The concept of video anomaly detection can be formally defined as follows: A video sequence denoted as υ, is given. From this sequence, we extract a consecutive

series of n frames, represented as $\upsilon = (i_1, i_2, i_3, ..., i_n)$. Each frame i_t is assigned a binary label, b, indicating whether it is considered normal or anomalous. In our context, a label of 0 ($b = 0$) signifies that the frame i_t is normal, whereas frames labeled with '1' indicate anomalies. Our primary objective is to determine whether the test frame i_t should be classified as an anomaly or not.

Figure 1 provides a visual representation of our proposed methodology, which comprises three integral phases. The first phase involves feature extraction using an Inception encoder. In the second phase, the features extracted from the initial convolutional layers are transformed and processed through a sequence of four dense layers. This processing typically involves dimensionality reduction and the extraction of higher-level features. Every dense layer utilizes a combination of weights and biases applied to input features, incorporating the ReLU activation function to introduce non-linearities. The progression through the series of dense layers allows the network to learn complex patterns and relationships within the feature maps. The transformations applied by each dense layer enable the network to capture and represent more abstract and informative features. In the final phase, the framework calculates a probability distribution and subsequently classifies the input frame as either normal or anomalous based on thresholding. Given the constraints of limited annotated data, as well as reduced training and inference times, this approach presents an efficient solution for video anomaly detection.

3.1 Feature extraction and anomaly detection

In the domain of video anomaly detection, feature extraction assumes a pivotal role, serving as a fundamental step in distilling pertinent information from video frames, thereby enabling subsequent analysis and anomaly detection. Various researchers have proposed diverse feature extraction techniques, spanning from training models from scratch, as illustrated by [11, 22], to harnessing pre-trained models, as demonstrated by [26, 27]. Furthermore, a subset of researchers has explored the integration of inception layers or feature pyramid networks (FPN) with pre-trained architectures [28, 29].

The foundation of our model relies on the robust capabilities of the Inception encoder, known for its proficiency in capturing intricate image patterns crucial for anomaly identification. The incorporation of two Inception blocks throughout the model optimally balances

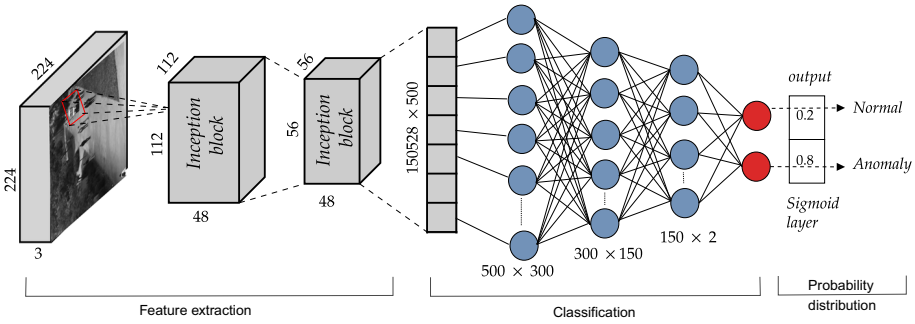

Fig. 1 The proposed methodology is visually depicted through three primary stages. The first stage is dedicated to extracting intrinsic features, followed by the second stage, which focuses on learning intricate feature relationships. The final stage involves the output layer, where the network generates predictions

simplicity with efficacy in video anomaly detection. These blocks play a pivotal role in analyzing spatial relationships and feature combinations within each frame. Each Inception block reduces the feature dimension, designed with two levels, and entails a specific composition as depicted in Fig. 2. The initial level of each block consists of two convolutional layers utilizing 1×1 and 5×5 filters, in addition to a 3×3 filter Max Pooling layer. The second level introduces a more comprehensive exploration of features, incorporating three convolutional layers with filter sizes of 1×1, 3×3, and 5×5, each contributing to the model's understanding of the data. As the frame advances within the model, these Inception blocks work collaboratively to systematically decrease the dimensionality of the feature maps. This reduction process is fundamental in distilling the most pertinent information from the frame, thereby preparing it for the subsequent classification phase. The Inception network can be represented as a function $f_\theta(i_t)$, where θ denotes the network's parameters. These features are then flattened using "Flatten" layer which can be formulated as:

$$z_0 = Flatten(f_\theta(i_t)) \tag{1}$$

‘ The flattened feature vectors undergo meticulous processing through a series of dense layers with trainable weights w and bias b. This transformation is encapsulated by a function $h_{w,b}(z)$. Let $z_i = h_{w_i b_i}(z_{i-1})$ represent the application of a dense layer with weights w_i and bias b_i to input z_{i-1} with activation function $\sigma(x)$, where i goes from 1 to 4. Then, the composition of these equations can be represented as:

$$z_4 = h_{w_4 b_4}(z_3) = \sigma(w_4^T \cdot z_3 + b_4) \tag{2}$$

In this single equation, we apply each dense layer successively to the previous layer's output, starting from z_1 and moving through z_2, z_3, and finally z_4. The loss function $\mathcal{L}$ is calculated as the binary cross entropy loss between the predicted probabilities and the actual ground truth labels.

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \cdot \log(z_4) + (1 - y_i) \cdot \log(1 - z_4)] \tag{3}$$

The decision process can be formulated using a threshold as follows:

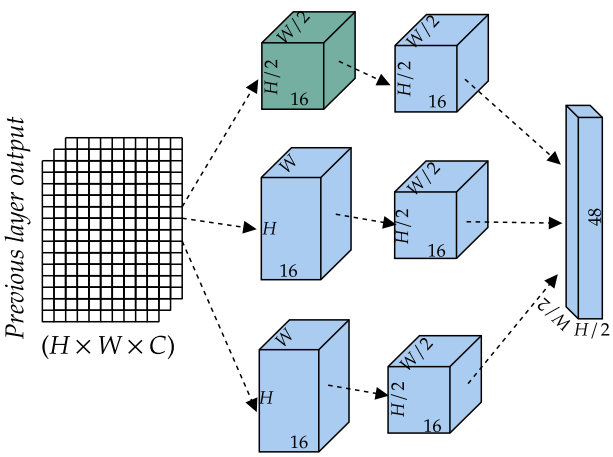

Fig. 2 Illustration of the Inception block used for the feature extraction

Let P_{anomaly} represent the probability assigned to the "anomaly" class by the sigmoid layer, and P_{normal} denote the probability assigned to the "normal" class. The model computes the likelihood of the frame i_t belonging to each class using the sigmoid function:

$$P_{\text{anomaly}} = \sigma(w_{\text{anomaly}}^T \cdot z_4 + b_{\text{anomaly}}) \tag{4}$$

$$P_{\text{normal}} = 1 - P_{\text{anomaly}} \tag{5}$$

To make a classification decision, a predefined threshold Γ is employed. If P_{anomaly} exceeds this threshold, the frame is categorized as an anomaly; otherwise, it is confidently labeled as normal:

$$\text{Frame Classification} = \begin{cases} \text{"Anomaly"} & \text{if } P_{\text{anomaly}} \geq \Gamma \\ \text{"Normal"} & \text{if } P_{\text{anomaly}} < \Gamma \end{cases} \tag{6}$$

This decision process utilizes the probabilities assigned by the sigmoid function to classify the frame as either "Anomaly or "Normal" based on the predefined threshold Γ.

4 Experiments and results

4.1 Datasets and evaluation metric

The UCSD Pedestrian dataset is split into two segments, namely Ped1 and Ped2, each offering unique viewing perspectives. Ped1 encompasses 34 training videos comprising 6,800 frames, and 36 testing videos comprising 7,200 frames. These videos exhibit a frame resolution of 238 × 158 pixels. Conversely, Ped2 comprises 16 training videos with 2,550 frames and 12 testing videos with 2,010 frames. Ped2's frame resolution is set at 360 × 240 pixels. In both datasets, normal scenarios feature pedestrians walking on designated pathways. However, anomalies encompass a variety of instances, including bicycles, carts, skateboards, pedestrians traversing grassy areas, and other vehicles.

The CUHK Avenue dataset comprised 16 training videos, each containing 15,328 frames at a resolution of 640 × 360 pixels. Additionally, it includes 21 test sequences, each comprising 15,324 frames, maintaining the same resolution. This dataset predominantly portrays individuals entering and exiting a building. It poses a significant challenge due to the presence of diverse anomalies, such as individuals tossing bags and papers, children engaged in activities like skipping, jumping, and running, as well as instances of bags placed on the grass.

Frames are initially extracted from the raw videos and resized uniformly to dimensions of 224×224 pixels. To ensure consistent intensity scales across frames, the pixel values of each frame are normalized to the [0, 1] range. Following this, frames from these datasets are categorized into two classes: anomalies (class 1) and non-anomalies (class 0). To facilitate effective model training and assess its performance, we divide the frames into two distinct sets that do not overlap: the training set and the testing set, maintaining an 80:20 ratio. These procedures ensure appropriate preprocessing and organization of the data for subsequent analysis and experimentation. Table 1 provides additional details on these datasets.

In evaluating the effectiveness of the detection algorithm, we utilize two commonly used quantitative metrics: the Equal Error Rate (EER) and the Area under the ROC Curve (AUC). Additionally, for assessing classification tasks, standard metrics such as precision, accuracy, F1 Score, and recall are utilized.

Table 1 Dataset Statistics used for Video Anomaly Detection

Dataset	#Videos	#Anomalous Incidents	Instance Sampling		Data Partitioning	
			Class 0	Class 1	Train	Test
Ped1	70	40	4156	4044	6560	1640
Ped2	28	12	1712	1648	2688	672
Avenue	37	47	4540	4454	7195	1799

4.2 Model parameters and running time

The Keras and TensorFlow frameworks in Python were utilized to implement the proposed methodology. All training and testing processes are conducted on a computer equipped with an NVIDIA GeForce GTX 1080 GPU. The average training time for the model over 20 epochs is approximately 600 seconds. During the training phase, the Binary Cross-Entropy (BCE) loss function is employed, and the SGD optimizer is utilized with a learning rate set to 0.0001. A batch size of 8 is used, and the ReLU activation function is applied during training.

4.3 Results and analysis

We conducted a series of experiments on three datasets, and the results are depicted in Table 2. The outcomes of our novel approach are highlighted in bold to underscore their significance. These highlighted metrics underscore the remarkable capability of the framework in differentiating between regular and anomalous frames across various datasets. To maintain uniformity in our assessments, we utilize the same network design and training parameters across all datasets.

In Fig. 3, a confusion matrix for binary classification of anomalies and normal data is presented. This matrix serves the purpose of providing a comprehensive visual representation of a classification model's performance. From the confusion matrix, a comprehensive array of performance metrics, such as Accuracy, Precision, Recall, and F1 score are computed

Table 2 Evaluation of AUC and EER scores in comparison with state-of-the-art studies

Methods ↓	AUC (%)			EER (%)		
	UCSD Ped1	UCSD Ped2	CUHK Avenue	UCSD Ped1	UCSD Ped2	CUHK Avenue
AMDN [30]	92.1	90.8	-	16	17	-
Spatial-temporal CNN [22]	-	-	-	24.0	24.4	-
AL [31]	89.7	90.1	-	17	18	-
Ionescu et al [9]	-	97.6	90.4	-	-	-
Anomaly-Net [32]	83.5	94.9	86.1	25.2	10.3	22
SIGnet [33]	86	96.2	86.8	-	-	-
Tian et al [34]	-	98.6	-	-	-	
Ruchika et al [11]	98.5	97.9	95.1	11	9	11.5
Mohammad et al [35]	94.2	95.1				
VAD-TL [26]	98.6	99.4	95.5	4.2	2.6	8.9
Ours	**98.9**	**99.6**	**98.1**	**3.42**	**1.2**	**1.1**

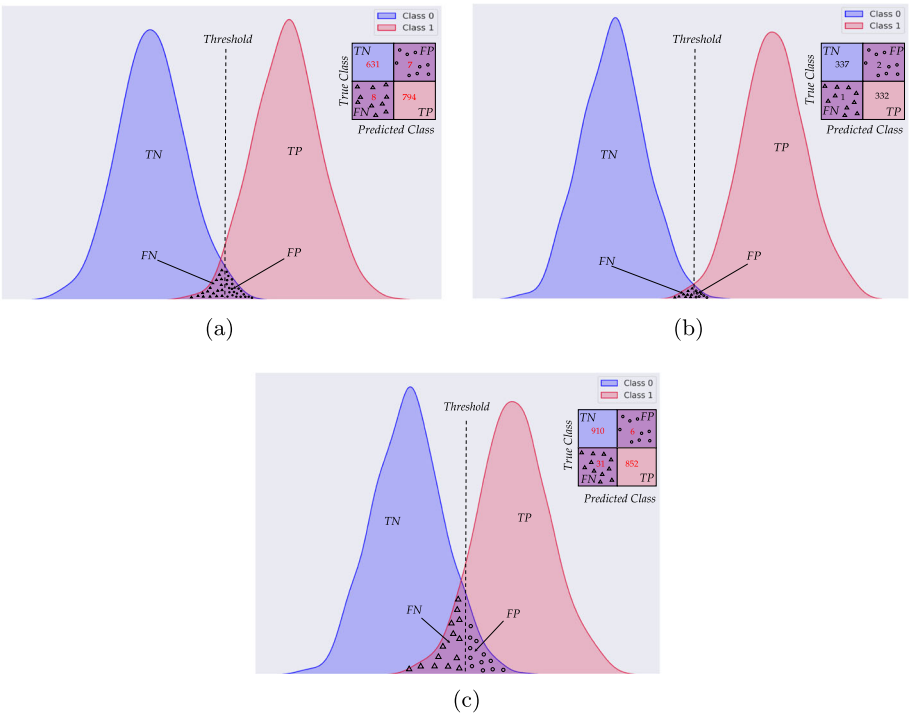

Fig. 3 Confusion matrices depicting model performance across (a) Ped1, (b) Ped2, and (c) Avenue datasets

and thoughtfully displayed in Fig. 4. These metrics offer a nuanced understanding of the framework's performance across various dimensions.

4.3.1 Analyzing weight distributions

The weight histograms offer valuable insights into the learned parameters of our network. To delve into the weight distributions of our CNN layers, we concentrate our analysis on

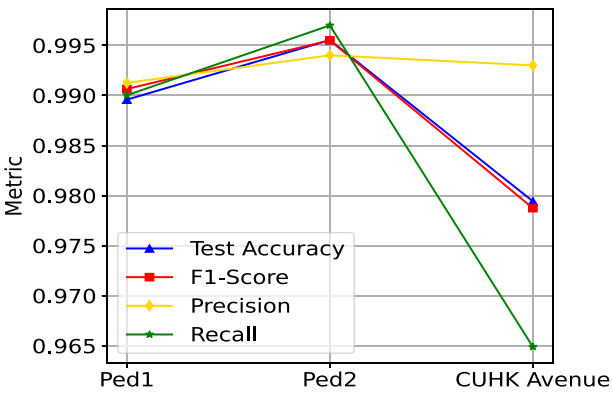

Fig. 4 Quantitative performance metrics for different datasets

two pivotal layers: the final convolution layer in block2 and the dense-2 layer. Figure 5 visually represents these weight distributions. A remarkable observation from these weight distributions is their Gaussian-like characteristics. Furthermore, the majority of weight values cluster around zero. These findings illuminate the stability of the learning process, showcasing the model's ability to steer clear of extreme weight values. This stability plays a crucial role in averting numerical instability and convergence challenges during training.

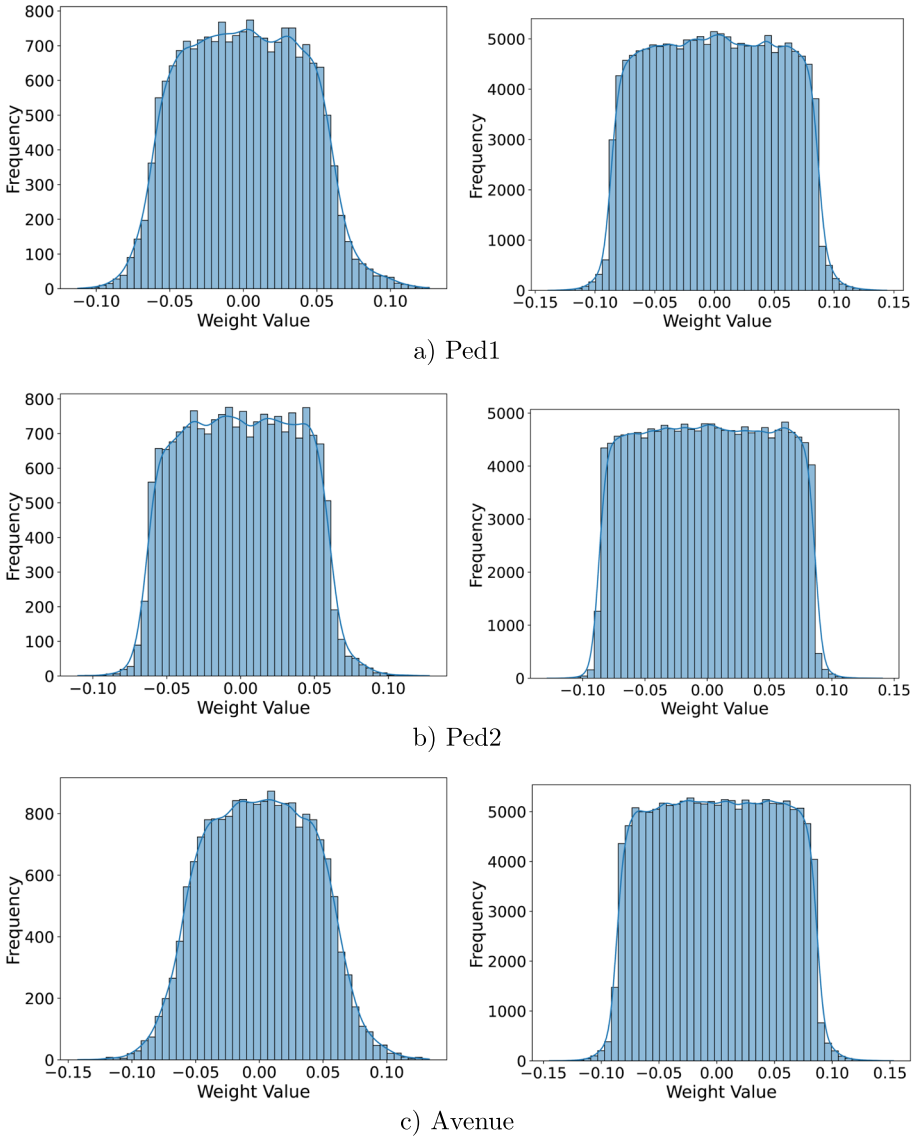

Fig. 5 Weights distributions in Inception block1_convolution layer2 (column1) and Dense layer (column2) for three datasets

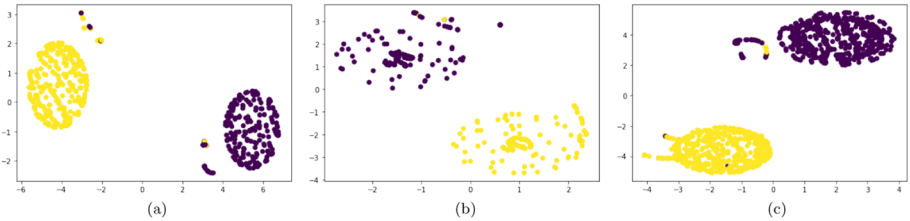

Fig. 6 2-D t-SNE plot of features obtained from the framework for the Ped1, Ped2, and Avenue datasets, respectively

4.3.2 Visualization of feature embeddings

The detection accuracy of a model relies heavily on how features are grouped or organized within the target domain. We propose that our modules adapt the model and improve the feature representation by making the grouping of features more distinguishable or separable. To validate this hypothesis and evaluate the quality of the feature representation achieved by our modules, we utilize t-SNE (t-distributed Stochastic Neighbor Embedding) to visualize the extracted features from the model. Fig. 6 illustrates the t-SNE plot for the result obtained for three datasets. By examining this plot, we can gain insights into how well our model captures and differentiates the underlying features in the target domain.

4.4 Ablation study

4.4.1 Evaluating the generalization ability of the model

I. Cross-dataset experiment In evaluating the model's capacity to generalize and endure domain shifts, we conduct cross-dataset testing across three diverse datasets. To visually represent the predicted class gradients, heatmaps were employed. However, it is crucial to note that heatmaps generated directly from the last convolutional layer highlight activated features without specific class information. In contrast, heatmaps generated using the Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm [36] add a class-specific context by considering the gradients associated with a target class. This approach provides a more refined visualization, aiding in the understanding of not just what the model is looking at but also what it deems important for a particular prediction. Thus, we employed the Grad-CAM technique to generate heatmaps, as illustrated in Fig. 7.

Analyzing the Grad-CAM heatmaps presented in Fig. 7, several key inferences can be drawn. The visualizations in subfigures 7b and 7f highlight the regions within the input images that significantly contribute to the model's predictions when heatmaps are generated using the conventional setup. Moreover, in cross-dataset scenarios, such as Ped1 $\rightarrow$ Ped2, Grad-CAM heatmaps allow us to discern how well the model adapts its attention to relevant features despite domain shifts. However, challenges arise when transferring knowledge between datasets with distinct characteristics. For instance, as depicted in subfigures 7c, 7d, 7g, and 7h the model encounters difficulties in adapting to the unique anomalies and scenes presented in the target dataset. This nuanced analysis, facilitated by Grad-CAM, enhances our understanding of the model's performance in different domains and provides valuable insights for future improvements.

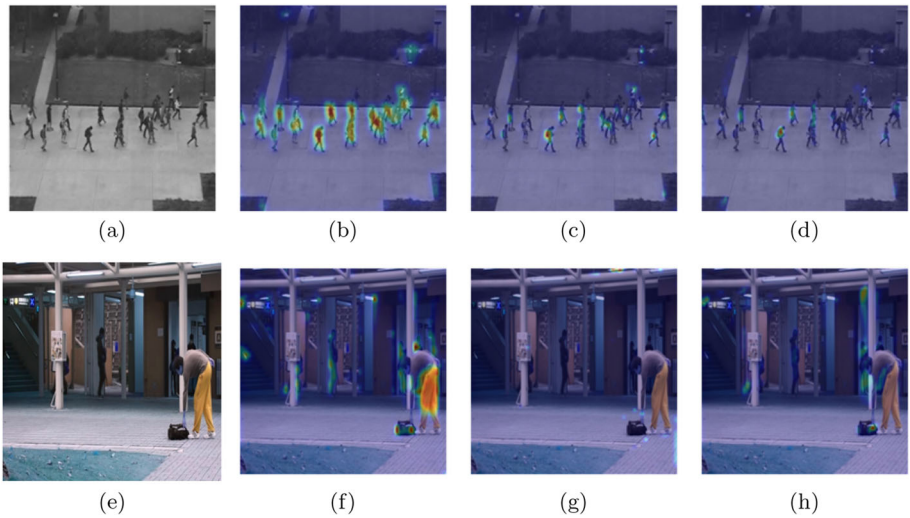

Fig. 7 Comparison of heatmaps generated using different setups on Ped2 and Avenue datasets. (a) and (e) depicts the original frames of the Ped2 and Avenue datasets. (b) and (f) show the GRAD-CAM heatmaps generated by the conventional setup. (c) and (g) display the heatmaps resulting from cross-testing between Avenue and Ped2, and Ped2 and Avenue, respectively. (d) and (h) illustrate the heatmaps produced by cross testing between Ped1 and Ped2, and Ped1 and Avenue, respectively

To further emphasize the limitations of cross-dataset generalization, the AUC values exhibit significant differences as illustrated in Table 3. The Ped2 → Avenue experiment yields an AUC of 61.42, considerably lower than the AUC of 73.77 observed in the Ped1 → Ped2 scenario. These results underscore the difficulty of applying learned anomaly detection patterns from Ped2 to the Avenue dataset. Similarly, the Avenue → Ped2 and Ped1 → Avenue experiments show AUC values of 59 and 58, respectively, reinforcing the model's challenge in generalizing across these diverse datasets.

II. Robustness to noise

Environmental challenges such as illumination fluctuations, adverse weather conditions, and camera degradation can significantly impact model performance. To enhance our model's resilience in the face of these challenges, we systematically evaluate and strengthen its capabilities through both adversarial testing and adversarial training. To simulate such real-world scenarios, we introduce perturbations to the original frames in the Avenue dataset. Specifically, we manipulate pixel values by introducing Gaussian noise across a range of standard

Table 3 Cross-domain performance when trained on source dataset and tested on target dataset

Training → Test	AUC (%)
Ped1 → Ped2	73.77
Ped2 → Avenue	61.42
Avenue → Ped1	60
Ped2 → Ped1	71.68
Avenue → Ped2	59
Ped1 → Avenue	58

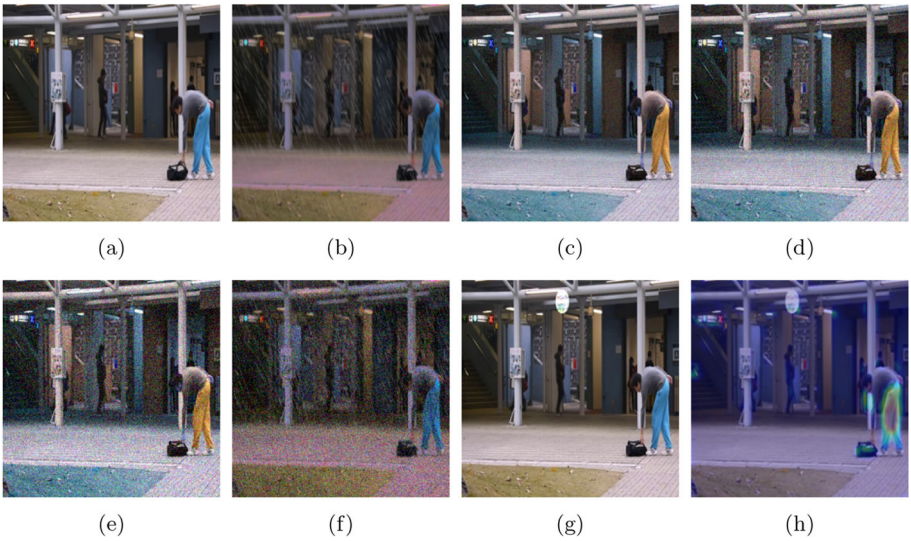

Fig. 8 Illustration of distorted and perturbed frames in the Avenue dataset. (a) Original frame, (b) Intense rain effect, (c), (d), (e) depict Gaussian noise with σ values of 0.3, 0.6, and 0.9 respectively, (f) Combined noise and rain effect, (g) Perturbed patch image, and (h) Adversarial training applied to the perturbed patch image

deviations ($\sigma = 0.3, 0.6,$ and 0.9) as illustrated in subfigures 8c, 8d, and 8e. Additionally, we apply rain effect to the frames using the Automold toolkit[1] as depicted in Fig. 8b. For adversarial testing, we train the model on original frames and assess its performance on distorted frames. The results are tabulated in Table 4. From the table, it can be observed that our method demonstrates robustness when tested with frames distorted by rain and noise individually. As expected, model performance diminishes with increasing noise intensity combined with rain. However, notably, our model consistently outperforms expectations under these challenging conditions. These results affirm the generalizability of our approach to unseen data characterized by noise and diverse weather conditions.

Introducing adversarial patches [37, 38] into video frames adds an extra layer of complexity and validates the model's robustness against adversarial patch attacks. To achieve this, we apply a spherical patch of size 30×30 to Avenue and Ped2 frames at random locations. Subsequent experiments involving adversarial training yield impactful results, as vividly presented in Fig. 9. The figure reveals a notable impact on the model's performance due to the application of adversarial training with the introduced patch. This signifies the model's heightened resilience against perturbations introduced by adversarial attacks, further reinforcing its effectiveness in the realm of anomaly detection for video frames. subfigures 8g and 8h visually depict an example of a perturbed frame and the resulting outcome after applying adversarial training to that image.

4.4.2 Effect of different feature extractors

To assess the effectiveness of our proposed feature extractor, we replaced our Inception encoder with established pre-trained models such as InceptionV3, VGG16, MobileNetV2, and DenseNet201 in the proposed framework. The experimental results, represented as AUC

[1] https://github.com/UjjwalSaxena/Automold--Road-Augmentation-Library

Table 4 Real-time performance on the Avenue dataset for frame distortions

Method	Standard Deviation (σ)			
	0	0.3	0.6	0.9
Baseline	98.11	-	-	-
Only rain	66.5	-	-	-
Only noise	98.11	98.11	97.06	95.2
Noise+rain	98.11	62.5	55.5	51.47

scores, are illustrated in Fig. 10 to provide a nuanced understanding of our novel feature extractor's impact on video anomaly detection. This comparative analysis delves into specific metrics, highlighting the model's strengths and contributions across diverse feature extraction architectures.

4.5 Unveiling training insights

Analyzing the convergence plot of training loss depicted in Fig. 11, we assess the influence of our novel deep learning model on video anomaly detection using the Avenue dataset. This evaluation involves a comparison with two alternative networks referenced as [11] and [26]. While the training schedules for [26] and our method remain identical, [11] utilizes an input shape of 200 $\times$ 200. All methods undergo 20 epochs of training to unveil their respective training tendencies. Throughout the training process, all three networks consistently exhibit a decline in training loss, indicating convergence towards minima. Notably, our inception encoder outperforms the other networks by demonstrating a lower training loss and superior learning efficiency. These specific architectural nuances underscore its proficiency in effectively learning and representing complex data.

5 Conclusion

In this paper, we design a novel supervised framework for efficient video anomaly detection, which is end-to-end trainable. We have demonstrated the superiority of this approach

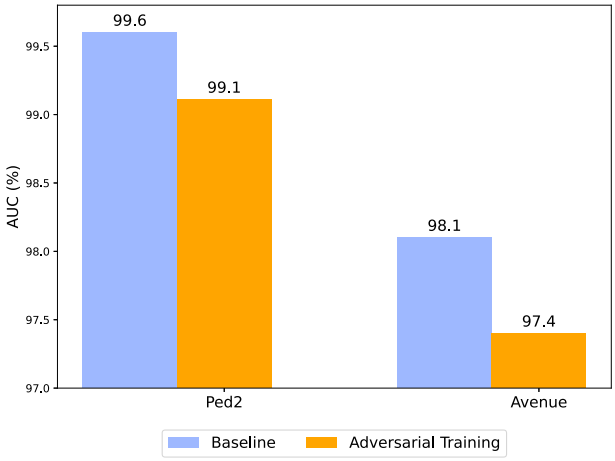

Fig. 9 AUC scores comparison on perturbed Avenue dataset frames with Baseline and adversarial training

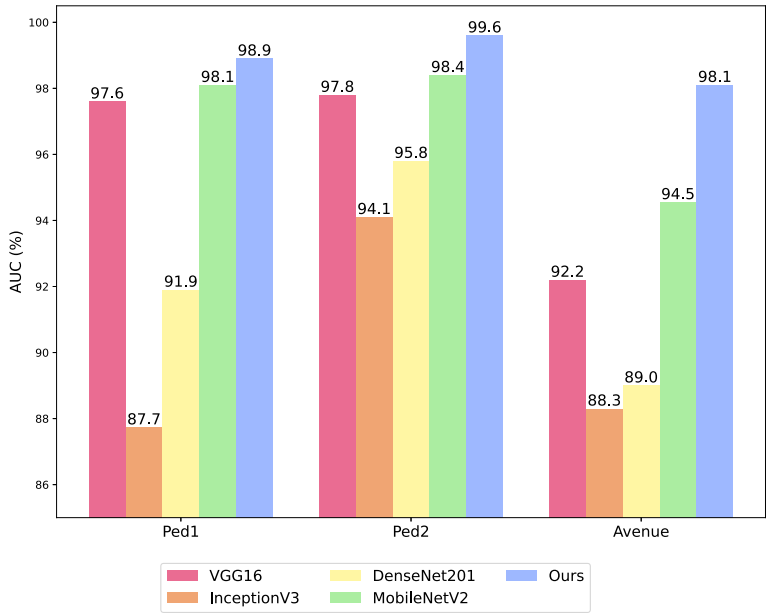

Fig. 10 Comparison of AUC scores across various feature extractors

over traditional unsupervised methods, which often require complex models and significant computational resources. Our method leverages annotated data at the frame level, striking a harmonious balance between performance and computational efficiency. The foundation of our approach lies in the utilization of the Inception encoder network, which excels at learning intricate and high-level features in video frames. This network seamlessly extends its capabilities to video data analysis, enabling the detection of anomalies by identifying deviations from established patterns. Our extensive evaluations demonstrate the compelling

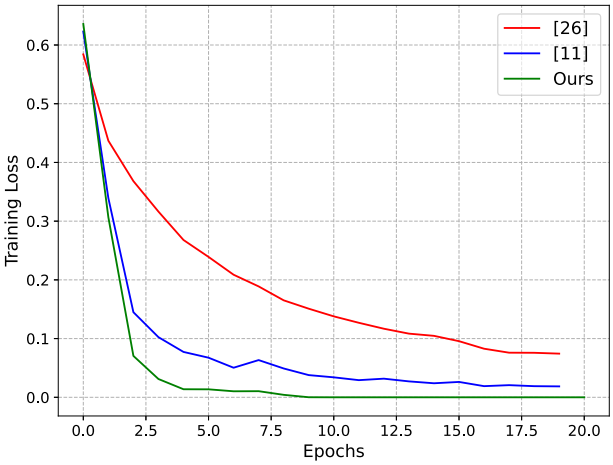

Fig. 11 Training error convergence plots correspond to the [11, 26], and proposed networks

performance of our proposed method, achieving impressive AUC values of 98.9%, 99.6%, and 98.1% on the Ped1, Ped2, and Avenue datasets, respectively. Moreover, we conduct ablation experiments that underscore the effectiveness of our network compared to several pre-trained models. In our pursuit of comprehensive validation, we also subject our model to cross-dataset testing, and adversarial training on these datasets to gauge its generalization capabilities and robustness to domain shifts. It is worth noting that while our approach may encounter challenges when confronted with unforeseen anomalies that lack representation in the training data, and data labeling can incur significant expenses, it excels in scenarios where the complexity of the model and performance are pivotal considerations.

Author Contributions All authors contributed equally to the research and manuscript preparation. Each author played a significant role in the design of the study, data collection, data analysis, and manuscript writing.

Funding No external funding was received for this research.

Availability of data and materials The data used in this study are available upon request.

Code availability The code used for data analysis is available upon request.

Declarations

Consent to participate All participants provided informed consent to participate in the study.

Consent for publication All authors consent to the publication of this manuscript.

Ethics approval This research study was conducted in compliance with all relevant ethical standards and guidelines.

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Sikdar A, Chowdhury AS (2020) An adaptive training-less framework for anomaly detection in crowd scenes. Neurocomputing 415:317–331. https://doi.org/10.1016/j.neucom.2020.07.058
2. Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS (2016) Learning temporal regularity in video sequences. In: 2016 IEEE Conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp 733–742. https://doi.org/10.1109/CVPR.2016.86
3. Chong YS, Tay YH (2017) Abnormal event detection in videos using spatiotemporal autoencoder. In: Cong F, Leung AC, Wei Q (eds) Advances in neural networks - ISNN 2017 - 14th International symposium, ISNN 2017, Sapporo, Hakodate, and Muroran, Hokkaido, Japan, June 21-26, 2017, Proceedings, Part II. Lecture Notes in Computer Science, vol 10262, pp 189–196. https://doi.org/10.1007/978-3-319-59081-3_23
4. Li N, Chang F, Liu C (2021) Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. IEEE Trans Multim 23:203–215. https://doi.org/10.1109/TMM.2020.2984093
5. Kommanduri R, Ghorai M (2023) Bi-read: Bi-residual autoencoder based feature enhancement for video anomaly detection. J Vis Commun Image Represent 95:103860. https://doi.org/10.1016/j.jvcir.2023.103860
6. Akcay S, Abarghouei AA, Breckon TP (2018) Ganomaly: Semi-supervised anomaly detection via adversarial training. In: Jawahar CV, Li H, Mori G, Schindler K (eds) Computer vision - ACCV 2018 - 14th Asian conference on computer vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part III. Lecture Notes in Computer Science, vol 11363, pp 622–637. https://doi.org/10.1007/978-3-030-20893-6_39
7. Chen D, Yue L, Chang X, Xu M, Jia T (2021) NM-GAN: noise-modulated generative adversarial network for video anomaly detection. Pattern Recognit 116:107969. https://doi.org/10.1016/j.patcog.2021.107969

8. Luo W, Liu W, Lian D, Gao S (2022) Future frame prediction network for video anomaly detection. IEEE Trans Pattern Anal Mach Intell 44(11):7505–7520. https://doi.org/10.1109/TPAMI.2021.3129349
9. Ionescu RT, Khan FS, Georgescu M, Shao L (2019) Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In: IEEE Conference on computer vision and pattern recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pp 7842–7851. https://doi.org/10.1109/CVPR.2019.00803
10. Bansod SD, Nandedkar AV (2019) Transfer learning for video anomaly detection. J Intell Fuzzy Syst 36(3):1967–1975. https://doi.org/10.3233/JIFS-169908
11. Lalit R, Purwar RK, Verma S, Jain A (2022) Correction to: Crowd abnormality detection in video sequences using supervised convolutional neural network. Multim Tools Appl 81(22):32701. https://doi.org/10.1007/s11042-022-13375-0
12. Ramachandra B, Jones MJ, Vatsavai RR (2022) A survey of single-scene video anomaly detection. IEEE Trans Pattern Anal Mach Intell 44(5):2293–2312. https://doi.org/10.1109/TPAMI.2020.3040591
13. Mondal R, Chanda B (2018) Anomaly detection using context dependent optical flow. In: ICVGIP 2018: 11th Indian conference on computer vision, graphics and image processing, Hyderabad, India, 18-22 December, 2018, pp 22–1228. https://doi.org/10.1145/3293353.3293375
14. Gandapur MQ (2022) E2E-VSDL: end-to-end video surveillance-based deep learning model to detect and prevent criminal activities. Image Vis Comput 123:104467. https://doi.org/10.1016/j.imavis.2022.104467
15. Amin J, Anjum MA, Ibrar K, Sharif M, Kadry S, Crespo RG (2023) Detection of anomaly in surveillance videos using quantum convolutional neural networks. Image Vis Comput 135:104710. https://doi.org/10.1016/j.imavis.2023.104710
16. Park C, Cho M, Lee M, Lee S (2022) Fastano: Fast anomaly detection via spatio-temporal patch transformation. In: IEEE/CVF Winter conference on applications of computer vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022, pp 1908–1918. https://doi.org/10.1109/WACV51458.2022.00197
17. Liu Y, Liu J, Zhao M, Li S, Song L (2022) Collaborative normality learning framework for weakly supervised video anomaly detection. IEEE Trans Circuits Syst II Express Briefs 69(5):2508–2512. https://doi.org/10.1109/TCSII.2022.3161061
18. Sultani W, Chen C, Shah M (2018) Real-world anomaly detection in surveillance videos. In: 2018 IEEE Conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pp 6479–6488. https://doi.org/10.1109/CVPR.2018.00678
19. Zhu Y, Newsam SD (2019) Motion-aware feature for improved video anomaly detection. In: 30th British machine vision conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019, p 270
20. Zhong J, Li N, Kong W, Liu S, Li TH, Li G (2019) Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: IEEE Conference on computer vision and pattern recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pp 1237–1246. https://doi.org/10.1109/CVPR.2019.00133
21. Thakare KV, Raghuwanshi Y, Dogra DP, Choi H, Kim I (2023) Dyannet: A scene dynamicity guided self-trained video anomaly detection network. In: IEEE/CVF Winter conference on applications of computer vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023, pp 5530–5539. https://doi.org/10.1109/WACV56688.2023.00550
22. Zhou S, Shen W, Zeng D, Fang M, Wei Y, Zhang Z (2016) Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes. Signal Process: Image Commun 47:358–368. https://doi.org/10.1016/j.image.2016.06.007
23. Hinami R, Mei T, Satoh S (2017) Joint detection and recounting of abnormal events by learning deep generic knowledge. In: IEEE International conference on computer vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pp 3639–3647. https://doi.org/10.1109/ICCV.2017.391
24. Ruchika P (2019) Abnormality detection using lbp features and k-means labelling based feed-forward neural network in video sequence. Int J Innovative Technol Exploring Eng 8:629–633
25. Ma Q (2021) Abnormal event detection in videos based on deep neural networks. Sci Program 2021:6412608–164126088. https://doi.org/10.1155/2021/6412608
26. Kommanduri R, Ghorai M (2023) A supervised approach for efficient video anomaly detection using transfer learning, pp 209–217. https://doi.org/10.1007/978-3-031-45170-6_22
27. Liu D, Cui Y, Yan L, Mousas C, Yang B, Chen Y (2021) Densernet: Weakly supervised visual localization using multi-scale feature aggregation. Proceedings of the AAAI Conference on Artificial Intelligence 35(7):6101–6109. https://doi.org/10.1609/aaai.v35i7.16760
28. Liu D, Cui Y, Tan W, Chen Y (2021) Sg-net: Spatial granularity network for one-stage video instance segmentation. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR). https://doi.org/10.1109/cvpr46437.2021.00969

29. Liu D, Liang J, Geng T, Loui A, Zhou T (2023) Tripartite feature enhanced pyramid network for dense prediction. IEEE Trans Image Process 32:2678–2692. https://doi.org/10.1109/tip.2023.3272826
30. Xu D, Ricci E, Yan Y, Song J, Sebe N (2015) Learning deep representations of appearance and motion for anomalous event detection. In: Procedings of the British machine vision conference 2015. BMVC 2015. https://doi.org/10.5244/c.29.8
31. He C, Shao J, Sun J (2017) An anomaly-introduced learning method for abnormal event detection. Multimed Tools Appl 77(22):29573–29588. https://doi.org/10.1007/s11042-017-5255-z
32. Zhou JT, Du J, Zhu H, Peng X, Liu Y, Goh RSM (2019) Anomalynet: An anomaly detection network for video surveillance. IEEE Trans Inf Forensics Secur 14(10):2537–2550. https://doi.org/10.1109/TIFS.2019.2900907
33. Fang Z, Liang J, Zhou JT, Xiao Y, Yang F (2022) Anomaly detection with bidirectional consistency in videos. IEEE Trans Neural Netw Learn Syst 33(3):1079–1092. https://doi.org/10.1109/tnnls.2020.3039899
34. Tian Y, Pang G, Chen Y, Singh R, Verjans JW, Carneiro G (2021) Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: 2021 IEEE/CVF International conference on computer vision (ICCV). https://doi.org/10.1109/iccv48922.2021.00493
35. Sabih M, Vishwakarma DK (2022) A novel framework for detection of motion and appearance-based anomaly using ensemble learning and lstms. Expert Syst Appl 192:116394. https://doi.org/10.1016/j.eswa.2021.116394
36. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International conference on computer vision (ICCV). https://doi.org/10.1109/iccv.2017.74
37. Brown TB, Mané D, Roy A, Abadi M, Gilmer J (2017) Adversarial patch. arXiv preprint arXiv:1712.09665. https://doi.org/10.48550/arXiv.1712.09665
38. Cheng Z, Liang J, Choi H, Tao G, Cao Z, Liu D, Zhang X (2022) Physical attack on monocular depth estimation with optimal adversarial patches, pp 514–532 . https://doi.org/10.1007/978-3-031-19839-7_30