Video Anomaly Detection using Residual Autoencoder: A Lightweight Framework

Mohamed H. Habeb
May A. Salama
Lamiaa A. Elrefaei

Follow this and additional works at: https://mej.researchcommons.org/home
Part of the Architecture Commons, and the Engineering Commons
Video Anomaly Detection Using Residual Autoencoder: A Lightweight Framework

Mohamed H. Habeb*, May A. Salama, Lamiaa A. Elrefaei

Electrical Engineering Department, Faculty of Engineering at Shoubra, Benha University, Cairo, Egypt

Abstract

This paper proposes an efficient lightweight deep spatial residual autoencoder (SRAE) model to detect anomalous events in video surveillance systems. A lightweight network is essential in real-time situations where time is critical. Moreover, it could be deployed on low-resource devices like embedded systems or mobile devices. This makes it a handy option for real-world situations where there may be a shortage of resources. The proposed network comprises a 3-layer residual encoder-decoder architecture adopted to acquire the salient spatial characteristics representative of normal events in videos. Then, the reconstruction loss is used to find abnormalities, where normal frames are recreated well with a low reconstruction loss, and abnormal ones are found as the opposite. The model’s efficiency is tested by two benchmark datasets, the University of California, San Diego (UCSD) Pedestrian 2 (Ped 2) and the CUHK Avenue, achieving AUC ≈ 95% and 81% for the two datasets, respectively. Hence, its performance proves to be comparable with the state-of-the-art models.

Keywords: Anomaly, Residual autoencoder, Security, Video surveillance

1. Introduction

In today’s rapidly evolving digital age, surveillance systems are essential for maintaining public safety and security. The need for efficient anomaly detection systems is highlighted by the billions of dollars lost annually because of criminal activity, vehicular crashes, and other unanticipated events. In the United States, there were 23 million damaged vehicles, 4.5 million injured, and 36,500 deaths from motor vehicle accidents in 2019. In addition to lost productivity, emergency services, legal, and court costs, as well as traffic, property damage, insurance administration, and workplace losses, the economic costs of these crashes came to $340 billion, while the societal harm was nearly $1.4 trillion (Blincoe et al., 2023). Therefore, anomaly identification in surveillance video is not only a computer vision challenge but also a social requirement because it can be used to predict threats and reduce hazards. In computer vision tasks, detecting anomalies in surveillance systems with high reliability and performance is an essential matter due to its wide range of applications, including the identification of criminal activities, traffic accidents, and suspicious behaviors. In addition, the interest in automated approaches for video surveillance has grown because of the developments in computer vision, and there is a pressing need for real-time applications with high performance and accuracy. Nevertheless, it can be hard to recognize an anomalous activity among all the other usual events. Since normal events occur more frequently than abnormal ones, which have rare and unusual characteristics, collecting and classifying all forms of abnormal occurrences is a challenging task. Moreover, abnormal events are also complicated in structure due to many factors, including chaotic or moving objects, messy backgrounds, etc (Liu et al., 2022). Another obstacle is the ambiguous nature of abnormal activities, besides the rare-annotated data. For instance, a specific
activity may be considered abnormal in one situation but normal in another. An event is regarded as regular when a pedestrian uses a crosswalk to cross the street. However, if there is no crosswalk, the same behavior is perceived as abnormal (Le and Kim, 2023).

Deep learning breakthroughs recently introduced a new era of possibilities in computer vision. Deep networks have recently proven their worth in several computer vision applications, including object detection (Szegedy et al., 2013; Zhu et al., 2015), object recognition (Wu and Chen, 2015), and action recognition (Simonyan and Zisserman, 2014). Hence, researchers have begun to experiment with training deep networks before training a detection model, such as a ‘one-class support vector machine’ (SVM). On the other hand, these deep features are not developed or optimized to tackle the issue of anomaly identification, thus they are not optimal.

Deep unsupervised detection techniques have seen extensive usage in the detection of video anomalous occurrences (Baydargil et al., 2021; Deepak et al., 2021; Feng et al., 2021; Gong et al., 2019). Unsupervised visual anomaly detection methods aim to produce models that can mimic our ability to identify objects as humans do. Additionally, the majority of real-world application cases favor the unsupervised scenario, as there is no useful supervision information for the problem of identifying abnormalities in videos. The majority of these approaches collect features from a deep neural network before training a detection model, such as a ‘one-class support vector machine’ (SVM). On the other hand, these deep features are not developed or optimized to tackle the issue of anomaly identification, thus they are not optimal.

Deep unsupervised detection techniques have seen extensive usage in the detection of video anomalous occurrences (Baydargil et al., 2021; Deepak et al., 2021; Feng et al., 2021; Gong et al., 2019). Unsupervised visual anomaly detection methods aim to produce models that can mimic our ability to identify objects as humans do. Additionally, the majority of real-world application cases favor the unsupervised scenario, as there is no useful supervision information for the problem of identifying abnormalities in videos. The majority of these approaches collect features from a deep neural network before training a detection model, such as a ‘one-class support vector machine’ (SVM). On the other hand, these deep features are not developed or optimized to tackle the issue of anomaly identification, thus they are not optimal.

Deep unsupervised detection techniques have seen extensive usage in the detection of video anomalous occurrences (Baydargil et al., 2021; Deepak et al., 2021; Feng et al., 2021; Gong et al., 2019). Unsupervised visual anomaly detection methods aim to produce models that can mimic our ability to identify objects as humans do. Additionally, the majority of real-world application cases favor the unsupervised scenario, as there is no useful supervision information for the problem of identifying abnormalities in videos. The majority of these approaches collect features from a deep neural network before training a detection model, such as a ‘one-class support vector machine’ (SVM). On the other hand, these deep features are not developed or optimized to tackle the issue of anomaly identification, thus they are not optimal.

The fundamental task of the unsupervised anomaly detection technique, such as that described in (Fan et al., 2020; Ionescu et al., 2019), is the use of only normal samples for training the model, after which the test dataset’s samples that deviate from the normal samples can be recognized. Furthermore, most video abnormal event detection algorithms are unable to accomplish online observation. Another main challenge that accompanies video anomaly detection techniques is that the models have several stacked layers, pretrained models, and/or multiple streams; as a result, the structure is more complicated. Thus, the training of such complex and heavyweight models is exposed to vanishing, exploding, and degradation problems, and finding anomalies takes an excessive amount of time. Therefore, this study aims to enable the model to learn the spatial characteristics using some relatively lightweight deep architecture with promising and acceptable high detection performance and accuracy.

The following are the main contributions of this study:

1. A lightweight end-to-end unsupervised system was proposed for anomaly identification in videos using a deep residual autoencoder. The proposed model encodes the spatial frame characteristics using a 3-layer encoder and reconstructs the input frame using a 3-layer decoder with doubled activation maps and a skip connection, which enhances anomaly detection accuracy with high performance.

2. A comprehensive evaluation of the proposed method using publicly accessible video anomaly detection datasets was conducted. The experiment reveals that the proposed technique achieves outstanding outcomes in comparison with state-of-the-art approaches.

3. Introducing a lightweight model to construct a real-time system where processing time is reduced as well as processing cost.

The rest of this paper is arranged as follows: A summary of the related work is provided in Section 2. In Section 3, background information about autoencoders and residual autoencoders is described. The details of the proposed model are given in Section 4. In Section 5, dataset description, detailed experiment results, and discussions are introduced. Finally, in Section 6, the conclusion and future perspectives are addressed.

2. Literature review

To detect anomalies in videos, recent studies have employed deep learning-based models for spatial and temporal patterns. The prediction-based (Section 2.1) and the reconstruction-based (Section 2.2) techniques are the two most common deep categories that are utilized for video anomaly detection. In this section, several works will be reviewed from this perspective.

2.1. Prediction-based approach

The fundamental principle of these methods is that normal activities could be predicted but
anomalous ones could not. Prediction-based algorithms predict whether a future frame will be normal or an anomaly based on a small number of prior frames. They typically exploit both the spatial and temporal characteristics of the input video, as it consists of multiple adjacent frames that contain motion characteristics. The majority of available reconstruction-based approaches address the anomaly detection issue by reducing the reconstruction losses in the training data; however, this does not guarantee higher reconstruction losses for anomalous data (Li et al., 2019).

One of the most popular networks is the ‘Generative Adversarial Network’ (GAN), which is comprised of a generator and a discriminator and may be utilized to produce the next frame for video anomaly identification. Based on GAN concepts, the authors (Zhou et al., 2019b) applied the U-Net (Ronneberger et al., 2015) as a generator to guess the future frame while a patch discriminator was utilized to differentiate between the snapshots that were produced via the generator. Moreover, they used the ‘peak signal to noise ratio’ (PSNR) for evaluating image prediction quality instead of the common mean square error (MSE). The authors (Zhou et al., 2019b) aimed to address the issue of the foreground-background imbalance in anomaly identification, which arises solely from the reconstructed loss optimization that leads to overfitting of the background rather than the foreground objects of interest, an attention-driven loss method was proposed. Similarly, to address the same issue but in numerous circumstances, the authors (Yang et al., 2020) proposed an approach that is based on a segmentation map with patch-level loss. They incorporated a map of segmentation into the PSNR by weighing the foreground and background differently. In addition, the patch-level loss was included in the proposed prediction technique to enhance the foreground item’s overall quality. Another U-net structure-like architecture called Multi-scale Multi-path Network Architecture (MsMp-net) was proposed by (Taghinezhad and Yazdi, 2023). The fundamental concept behind this network design is to memorize regular semantic information while simultaneously gathering and including temporal information. The MsMp-net consists of a time-distributed two dimensional encoder for appearance feature extraction, a multiscale memorizer module to store and retrieve semantic data at various levels, a multi-path predictor for preserving motion features, and a 2D decoder for generating predicted frames. Another work (Li et al., 2019) that was based on the GAN framework used a modified spatiotemporal U-Net to replace the generating network for predicting future frames and detecting anomalous events. Convolutional LSTM (ConvLSTM) layers were added to the conventional U-Net for extracting temporal motion information. Since anomalous frames are initially only noticeable at a video scene’s edges, a new regularity scoring algorithm was introduced that uses PSNR in both the present and subsequent frames to determine whether an abnormality has occurred. Hence, in contrast to frames found amid anomalous activity, such frames’ regularity score is high. Using an encoder-decoder reconstruction network and generalization capabilities based on reconstruction error maps and noise maps, the authors (Chen et al., 2021) introduced a GAN-like architecture named ‘Noise-modulated-GAN’, which is modulated by adversarial learning. Furthermore, reconstruction error maps were used to train a CNN-based discrimination network to identify out-of-the-ordinary data and separate it from normal samples. As a result, the reconstruction error maps will deviate from the expected normal distribution if the input data are anomaly samples.

For online anomaly detection, (Doshi and Yilmaz, 2021) combined the lightweight object detector called You Only Look Once (YOLOv3) (Redmon et al., 2016), which is utilized to capture significant location and spatial traits, and the GAN-based future frame predictor, in addition to the nonparametric statistical approach that employs the GAN-based anomaly identification. In (Le and Kim, 2023) a future frame prediction method was proposed where spatial and temporal information is adopted as separate branches inside a unified network. The model was unsupervised and trained with the use of a residual autoencoder structure that incorporates a multiple-phase channel attention decoder and a deep CNN encoder. While channel attention modules are responsible for extracting contextual dependencies, temporal shift methods are employed to make use of the temporal characteristic. Another work that aimed to address the continuous learning issue was proposed by (Doshi and Yilmaz, 2022). The authors employed a deep object detector approach to get pattern embeddings for input video frames. Then, a memory module was utilized to store a collection of nominal feature vectors using k-nearest-neighbors. Moreover, this procedure is trained in numerous sessions to promote continuous learning. A model for learning the difference between two frames, known as the residual frame, was introduced in (Yu et al., 2023) by stacking several ConvLSTM layers with a self-attention mechanism and a concatenation
operation. The next frame in the input video sequence is then predicted using the model. Therefore, if the consecutive frames utilized as input include anomalous objects, the resultant frame would not be reconstructed appropriately. A 3D convolutional multi-branch fusion network called ‘Branch-Fusion Net’ was combined with a Channel Spatial Attention Module (CSAM) and introduced by (Zhang and Lu, 2023) as a local features extractor to find anomalous behaviors in surveillance videos. The authors developed the Branch-Fusion Net to overcome high parameter and poor generalization problems during training and to perceive the input feature maps from several perspectives. Then, the CSAM was adopted to concentrate on spatial feature areas and essential channels to eliminate unnecessary features and increase crucial characteristics. Furthermore, a Bi-Directional Gated Recurrent Unit (Bi-GRU) was used to obtain the global features from videos. A weakly supervised work was proposed by (Ma et al., 2021) that suggested a neural network architecture to extract the essential elements for determining whether a video has abnormalities. The detection process includes assigning video labels after treating a video as an integral input. The authors introduced 3D convolutions to extract spatial and temporal data; then their dependencies are further represented using an LSTM network. Finally, the final score of the video is obtained.

2.2. Reconstruction-based approach

In this category, the trained model is responsible for rebuilding the input video frame. The autoencoder architecture is one of the most widely used models, and it consists of two components: an encoder and a decoder. The encoder condenses the frame into a representation of features with fewer dimensions, and the decoder retrieves the output frame from these reduced representations in such a way that it is as similar as possible to the original input frame. The reconstruction error is then utilized to differentiate between the anomaly and normal occurrences since the anomaly event has a larger reconstruction error than the normal event.

Several techniques (Nawaratne et al., 2019; Wei et al., 2019) learned the normal events using an encoder-decoder framework that included both stacked ConvLSTM layers to acquire the temporal characteristics and stacked CNN layers to get the spatial features. Recently, continuous learning has been used to solve the problem of forgetting that occurs during the training of deep neural networks to identify video anomalies, such as in (Nawaratne et al., 2019) which utilized a professional investigator as a kind of continual learning.

Frequently, a two-stream model (Li and Chang, 2019) was employed to gather both spatial and temporal data. The standard structure of such a model is comprised of an autoencoder and a discriminator. For better abnormality detection accuracy, the two-stream anomaly scores are fused. Similarly, to capture the spatial and temporal aspects of regular occurrences in the video, (Li et al., 2020) also developed a two-stream network. Each network stream had two spatiotemporal autoencoders that utilized 3D video cuboids as input. Several patches were segmented at the same position in successive frames to create the 3D video cuboids.

To capture the motion and appearance traits of regular activities from video surveillance frames and identify anomalies, a residual spatiotemporal autoencoder called (R-STAE) is comprised of 3D convolution, deconvolution, and ConvLSTM layers developed by (Deepak et al., 2021). Another deep autoencoder-based framework was introduced by (Abati et al., 2019), in which a density estimation probabilistic model was developed to estimate the density in the extracted feature space of a deep autoencoder using an autoregressive mechanism. The goal is to optimize the maximum likelihood in combination with normal sample reconstructions to serve as a regularizer for anomaly identification. On the other hand, an attention-based ConvLSTM network and a convolutional autoencoder (Conv. AE) were presented in (Wang and Yang, 2022), resulting in the Convolutional Recurrent Autoencoder (CRAE). The convolutional encoder is used to extract spatial abnormalities, whereas the ConvLSTM network is capable of capturing the temporal pattern’s irregularity. The features of the current frame are retrieved from each of the ConvLSTM layers’ hidden units using the attention technique. A convolutional decoder is then employed to rebuild test-video frames, which seemed to have greater reconstruction error and are thus considered anomalies.

Two methods for acquiring regularities for anomaly identification were implemented using the AE architecture and presented in (Hasan et al., 2016). Both methods use classifiers to make use of Spatiotemporal local features of regular activities in video sequences; the first is a fully connected AE trained using traditional hand-crafted appearance-motion local characteristics of normal behavior. The second method used a fully convolutional feed-forward AE to exploit the spatiotemporal local characteristics of typical activities in video sequences. Similarly, the work of (Luo et al., 2017)
introduced an autoencoder-based model; the spatial features were extracted using CNN, and the temporal characteristics were extracted using LSTM. To find anomalies in RGB movies, (Duman and Erdem, 2019) suggested an unsupervised anomaly detection model that uses Conv. AE and Conv. LSTM. This framework allowed for the acquisition of the pattern of foreground moving objects, including speeds and motion directions. Next, the Conv. AE is utilized to extract the spatial structure of each dense optical flow map volume and the Conv. LSTM network is employed to learn the temporal patterns of encoded optical flow maps of regular activities. Inspired by Deep Neural Networks (DNN) and utilizing sparse coding, (Luo et al., 2019) created a Temporally-coherent Sparse Coding (TSC) approach for video anomaly identification, which introduces a temporal characterization strategy based on appearance features. In addition, a separate stacked Recurrent Neural Network -Autoencoder (sRNN-AE) was designed to encode temporal data in addition to spatial ones. The TSC framework is better suited for anomaly identification since it maintains frames’ similarities between normal and abnormal activities.

To extract features and incorporate clustering to discover abnormal human behavior based on skeletal characteristics, the authors (Liu et al., 2022) presented a self-attention augmenting graph convolution. Accordingly, the skeletal data are robust against varying camera positions, angles, and lighting conditions. So, to gather both local and global information on the joints, a spatial—temporal self-attention augmented graph convolutional autoencoder (SAA-STGCAE) was created by combining an improved spatial graph convolution operator with a modified transformer self-attention operator. Additionally, a spatial self-attention module is utilized to examine the inter-frame interactions of human body components. In (Khaire and Kumar, 2022), a multimodal semi-supervised deep learning framework for video anomaly detection in the Bank-ATM situation was revealed that is also based on a reconstruction error base. The framework employed a MobileNet (Howard et al., 2017) pre-trained model as a feature extractor, followed by a bidirectional LSTM encoder-decoder module, to learn the usual pattern sequences from the training video clips. Afterward, the input feature segment is compared with the rebuilt output feature segment to determine the reconstruction error and define anomalies. A Dual-Stream Memory Network (DSM-Net) was developed by (Wang and Chen, 2023) to provide more context for the anomaly detection system. The proposed model can investigate the historical information between video clips to model regular event patterns by employing moving average encoders to capture the historical data of the events and optical flow to identify the temporal correlations of the behavioral patterns in the frames. The reconstruction loss of both the spatial and temporal branches is fused to give the final anomaly score. Zhang et al. (Zhang et al., 2023) introduced an AE structure-like attention model based on a Dynamic Prototype Unit (DPU) (Lv et al., 2021) and deep separable convolution operation for video anomaly detection. The model utilized the DPU to constitute the prototype pool from the hidden layer’s coding diagram. Then, the attention module is employed to extract spatial features from the video sequence and to decrease the impact of video background information. Finally, a deep separable convolution (Depth-wise (DW)) and a Pointwise convolution (PW) are adopted in the decoder part to reduce the model’s parameters and improve accuracy. The authors used feature reconstruction and frame prediction errors to obtain the anomaly score to identify normal and abnormal frames in the video sequence. Another reconstruction-based approach named dual-generator generative adversarial network (DGGAN) was developed by (Qi et al., 2023). The approach adopted a GAN-like structure that includes two generators: a noise AE to generate pseudo-anomaly frames and a reconstruction AE to learn the normal distribution of the normal frames using the training real frames and the pseudo-anomaly generated frames and then reconstructing them to detect abnormalities in the videos. To make better use of the salient data features, a second-order channel attention module is employed by the authors to learn feature interdependencies. Then, in the end, constraints are imposed to increase the variance between the regular frame and the rebuilt pseudo-abnormal frame and reduce the variance between the regular frame and the reconstructed regular frame.

While numerous studies have explored video anomaly detection using deep networks, many have predominantly focused on detection accuracy rather than performance, ignoring the need to balance efficiency and complexity. Although these approaches are efficient in some scenarios, they often have limitations in real situations with reliable accuracy and performance due to their complex architectures with numerous stacked stages and streams, as shown in Table 1, that negatively affect the detection process with reliable performance for real-time applications. In contrast, the introduced work introduces a lightweight spatial residual model specifically designed to tackle the complexity of the models with only a 3-layer encoder-decoder
architecture with a doubled unit number in the decoder layers that enables an efficient and fast anomaly detection process with acceptable accuracy to be a seed of real-time applications, especially for limited capability hardware. Not only does the proposed method offer better processing times, but it also paves the way for robust real-time anomaly detection systems. By integrating deep unsupervised techniques to be able to deal with rare, unseen, and unknown anomaly situations, the proposed model has been able to achieve comparable detection efficiency AUC = 95 % and 81 % for the University of California, San Diego (UCSD)-Ped 2 and the CUHK-Avenue datasets, respectively with only 4.8M parameters and 1.6G floating-point operations (FLOPs) which proofs its performance regarding state-of-the-art approaches that incorporate more complex architectures.

As a result of the investigation into the current alternatives for video anomaly detection, the reconstruction-based approach was adopted to detect video anomalies, wherein a larger reconstruction error between the input and reconstructed frame is indicative of a likely anomaly frame and a smaller reconstruction error indicates a normal frame. In all investigated review studies, the networks used were, to the best of our knowledge, heavyweight. There is a potential drawback to these networks since they are built on the foundation of pretrained models, several stacked layers, and/or various streams. Heavyweight models, or sophisticated models, need a lot of processing time and resources, which works against efficient anomaly identification. As a result of different studies investigation, the proposed work suggests a model that is both lightweight and fast, thanks to its simplistic design and shallow architecture. Furthermore, a model was proposed that is not only lightweight but also highly accurate in detecting abnormalities across a variety of datasets and application areas. Hence, applying this to real-time detection might be both a contribution and an additional benefit.

3. Autoencoder for anomaly detection: a background

In this section, a brief illustration of the usefulness and drawbacks of the autoencoder for anomaly detection (Section 3.1) as a backbone model and also the advantages of using the residual autoencoder (Section 3.2), adopted in the proposed work, over the simple autoencoder for anomaly detection.

3.1. Autoencoder-based anomaly detection

Autoencoder (AE) is a (one/more)-hidden-layer(s) neural network that comprises two primary parts, an encoder, and a decoder, is trained on unlabeled data and is completely unsupervised. A deep AE converts the input data to its hidden low-dimensional representation utilizing a nonlinear mapping function. The objective is to train the AE to
reconstitute the input patterns at the network’s output. When trained entirely on normal data samples, AEs are unable to reproduce anomalous data samples. Hence, anomalies or abnormal actions deviate from the trained model, resulting in poor reconstruction (high reconstruction error). Anomaly identification techniques based on deep AEs are generally categorized into model architectures using AEs and hybrid models. Model architectures that use AEs as a basic component make use of the reconstruction loss as an anomaly score between both the input and reconstructed output frames to determine the presence of anomalies. On the other hand, hybrid models primarily employ an AE as a feature extractor in which the decoder part of the AE is dropped after the training process and then the learned feature representations from the AE hidden layer are obtained and fed into another anomaly detection technique(s) (e.g. Clustering, Generative models, etc.). AEs suffer from some drawbacks that include:

1. A tendency to have vanishing gradients.
2. Memorization and lower-quality reconstruction of the input data, especially for images.

3.2. Residual autoencoder-based anomaly detection

Residual networks are simpler to tune, and a significant increase in network depth can improve their precision. A fundamental residual block in a residual model architecture, as shown in Fig. 1a (He et al., 2016), has an identity skip connection. This aids in the transmission of information from preceding layers and facilitates the flow of gradients during backpropagation, hence preventing the problem of disappearing gradients. Residual

---

**Fig. 1. The residual blocks and their activation functions.**
networks are used in this work to circumvent the vanishing gradients issue that exhausts the deep architectures. In the proposed model experiment, the residual connection is slightly different, as shown in Fig. 1b, from the regular one used in literature, where the used residual connection is comprised of element-wise addition of the first layer of the encoder to the second layer in the decoder, and after that, Leaky-Rectified Linear Unit (Leaky-ReLU) activation function is applied, and then the batch normalization. Fortunately, this modification improves the training efficiency and gives promising results in anomaly detection. Since ReLU’s activation values, as shown in Fig. 1c, have zero lower bounds, Leaky-ReLU, shown in Fig. 1d, is used as the residual block’s activation function to allow slightly negative values to not be suppressed. From our point of view and through the literature works’ investigation, Leaky-ReLU has a good effect on mitigating the vanishing and exploding gradient problems during the training of the proposed model. Further, a regularization strategy is used to improve training effectiveness, which is called Batch Normalization (BN). The used residual block’s equation, given x as input, is provided by:

\[ Y(x) = F(x) + x \]  

where x is the input frame, F is the residual function learned by the block’s stacked layers, and Y is the output of the residual block.

Like AEs, models that include residual AEs as fundamental components utilize the reconstruction difference between the reconstructed output and the input to identify anomalies.

4. Proposed method

To identify abnormalities in videos, robust deep learning networks that successfully acquire the distinctions between abnormal and regular actions are seen as being of utmost relevance. Moreover, deep models can learn more generic features that address intra-class variances found in a variety of normal events. In this work, we introduce an unsupervised learning model for identifying anomalous occurrences in the video.

4.1. Network architecture

A spatial residual autoencoder model was developed to obtain appearance features from input video frames that can discriminate between normal and anomalous events. Fig. 2 depicts the spatial residual autoencoder architecture, which forms the backbone of the proposed model and comprises an encoder and a decoder, each having three layers. The encoder consists of three two dimensional convolution layers with 32, 64, and 128 units, respectively. While the decoder comprises three 2D deconvolutional layers with 64, 32, and 16 units, respectively. Since anomaly detection is mainly concerned with ‘low-level contours’ and ‘edge features’, three Conv-layers are employed in the encoder for extracting the spatial patterns. In contrast, the deconvolutional layers in the decoder

![Fig. 2. The overall proposed model architecture for video anomaly detection.](image-url)
are responsible for reconstructing video frames and densifying sparse inputs through repeated filter operations. It is worth mentioning that the doubled unit number in the decoder layers, in comparison with the corresponding encoder layers, enhanced the frame’s reconstruction process and enabled the model not to need more layers to detect anomalies, which in turn reduced the model’s complexity.

In this study, residual networks were employed to get around the vanishing gradient problem that frequently affects deep networks. The residual connection experiment differs from the standard residual connection utilized in the literature, which consists of adding the first layer of the encoder and the second layer of the decoder element by element before applying Leaky-ReLU and batch normalization. Fortunately, this adjustment boosts training effectiveness and provides encouraging anomaly detection outcomes. Moreover, incorporating a residual connection into the system aids in obtaining minimal reconstruction loss compared with systems without residual connections. Leaky-ReLU is utilized as the residual block’s activation function because ReLU’s activation values have a zero-lower bound, which permits significantly less suppression of negative values. According to the observations and the results of the proposed experiment, Leaky-ReLU effectively reduces the issues of vanishing and exploding gradients during the training of the suggested model.

During the training step, normal frames, which represent only normal activities, are used to train the model, and the loss function is optimized by adjusting the learnable parameters. These parameters are updated in a SGD manner by computing the partial derivatives of the loss function.

4.2. Feature learning

The process of learning new features is a crucial component of the model training stage. During the encoding step, the model can learn both the crucial background details as well as the appearance characteristics of the monitored object that are included within the video frames. The encoder utilizes the input video frames and maps them into a latent representation, as shown in Fig. 3, which represents common data patterns and functions as a compressed version of the input video frame. In this study, only normal video frames were utilized to get the model trained with a latent space of size 256. Therefore, the obtained latent characteristics represent only the normal. The decoder uses the generated ‘bottleneck’ characteristics from this latent space to reconstruct the original video frames and achieve optimal accuracy when comparing the original and video-reconstructed frames. Hence, the goal is to successfully reconstruct the regular video frames in a meaningful way by reducing the reconstruction loss throughout the learning phase.

The step-by-step description of the operations in this model is as follows:

(1) The Encoder Part:
   (a) Input: A grayscale frame with shape (192, 192, 1).
   (b) Conv2D_0: Apply a convolutional layer with 64 filters, a kernel size of (3, 3), and a stride of 2. This will reduce the spatial dimensions of the input image by half (96, 96) while increasing the number of channels to 64.
   (c) Conv2D_1: Apply a convolutional layer with 32 filters, a kernel size of (3, 3), and a stride of 2. This will again reduce the spatial dimensions by half (48, 48) while decreasing the number of channels to 32.
   (d) Conv2D_2: Apply a convolutional layer with 16 filters, a kernel size of (3, 3), and a stride of 2. This will further reduce the spatial dimensions by half (24, 24) while decreasing the number of channels to 16.
   (e) Flatten: Flatten the feature maps into a one-dimensional vector of length 9216.
   (f) Dense_0: Apply a fully connected layer with 256 units and a linear activation.

![Fig. 3. The flowchart diagram of the proposed model.](https://example.com/fig3.png)
function to obtain the latent features representation of the input frame.

(2) The Decoder Part:

(a) Dense_1: Apply another fully connected layer with 9216 units and a linear activation function to start to reconstruct the features of the encoded frame.

(b) Reshape: Reshape the output of the Dense_1 layer back into the shape (24, 24, 16), which is the same shape as the output of the Conv2D_2 layer.

(c) Conv2DTranspose_0: Apply a transposed convolutional layer (sometimes called ‘deconvolution’ or ‘upsampling’) with 32 filters, a kernel size of (3, 3), and a stride of 2. This will double the spatial dimensions (48, 48) while increasing the number of channels to 32.

(d) Conv2DTranspose_1: Apply another transposed convolutional layer with 64 filters, a kernel size of (3, 3), and a stride of 2. This will double the spatial dimensions again (96, 96) while increasing the number of channels to 64.

(e) Add: Element-wise addition of the output of Conv2D_0 (step 2) and the output of Conv2DTranspose_1 (step 10). This is a form of skip connection, which helps the model learn more effectively by providing an additional path for gradient flow during backpropagation and allows the model to overcome the degradation problem that occurs during training.

(f) Leaky_ReLU: Apply a Leaky ReLU activation function with a small negative slope for input values less than zero. This introduces nonlinearity into the model.

(g) Batch_Normalization: Normalize the output of the LeakyReLU layer to improve training stability and convergence.

(h) Conv2DTranspose_2: Apply a transposed convolutional layer with 128 filters, a kernel size of (3, 3), and a stride of 2. This will double the spatial dimensions (192, 192) while increasing the number of channels to 128.

(i) Conv2D_3: Apply a convolutional layer with 1 filter, a kernel size of (3, 3), and a stride of 1. This will maintain the spatial dimensions (192, 192) while reducing the number of channels to 1, matching the input frame’s shape.

Assume the layer Conv2D_0 (r) with input and output $x_r$ and $y_r$, respectively and the layer Conv2DTranspose_1 (R) with input and output $x_R$ and $y_R$, respectively, then the residual (skip) connection contribution to the output of the layer R can be represented as follows:

$$y_R = y_r + F(x_r, W_R)$$  \hspace{1cm} (2)

since Conv2DTranspose_1 is deeper than Conv2D_0, the above equation can be rewritten as:

$$y_R = y_r + F(\ldots f(x_r, W_r) \ldots)$$  \hspace{1cm} (3)

Where $f$ and $F$ are activation functions.

4.3. Reconstruction error

Normal frames are utilized to train the spatial residual model, and normal events are learned while reconstructing the given input. The training is conducted on the reconstruction loss (Nayak et al., 2021):

$$L_{red} = \| V_f - \hat{V}_f \|_2$$  \hspace{1cm} (4)

where $V_f$ is the inputted frame and $\hat{V}_f$ is the reconstructed frame.

The mean squared error (MSE) is adopted to determine how far off the reconstructed frame is from the original one by calculating the average squared difference between them. This is because the values of MSE will be larger for anomalous frames and less for regular ones (as the model is trained for normality only). The objective is to reconstruct the normal video parts meaningfully. To do this, the reconstructive loss during the training process must be reduced by using stable structures. For each video frame ($f$), The reconstruction error of the pixel spatial value ($I$) at location $(x,y)$ is computed as follows:

$$P(x,y,f) = \| I(x,y,f) - M(I(x,y,f)) \|_2$$  \hspace{1cm} (5)

Where $M$ refers to the proposed spatial residual autoencoder model. Then, the reconstruction error was computed for each frame by adding together all the pixel-based reconstruction errors.

$$R(f) = \sum_{(x,y)} P(x,y,f)$$  \hspace{1cm} (6)

Then, the frame’s normalized anomaly score $SA$ is then computed by utilizing the following formulas:

$$S_A(f) = \frac{R(f) - \min R(f)}{\max R(f) - \min R(f)}$$  \hspace{1cm} (7)

and the frame’s normalized normality score is:
Both the anomaly score and normality score are in the [0–1] range. A cutoff point or threshold is then experimentally determined, which means that a frame is considered abnormal during testing if its anomaly score is over a threshold number and vice versa. In the performed experiment, the Precision-Recall Curve and ROC Curve were utilized to investigate the accuracy of the proposed model. As a result, for video frame anomaly detection, the grid search could be used to tune the threshold and find the optimal value that maximizes the F1-score metric, which is the harmonic mean of precision and recall, as shown in the following section.

5. Experimental results

5.1. Datasets

The datasets utilized in this study are intended for unsupervised modeling; as a result, the data utilized for training includes only normal video frames, whereas the data used for testing will also include video frames that are anomalous. The most used benchmark datasets for video anomaly identification are the UCSD Pedestrian 2 (Mahadevan et al., 2010) and CUHK Avenue (Lu et al., 2013) datasets because they provide a wide range of difficult scenarios and abnormalities that are typical of actual surveillance systems. While the CUHK dataset has anomalies like people loitering, running, jumping, and crossing in unexpected places, the UCSD dataset contains abnormalities like bikers, skateboarders, tiny carts, and pedestrians crossing a walkway or in its surrounding grass. Additionally, these datasets are also labeled, which makes it possible to assess how well the anomaly detection algorithms are working. The fact that so many studies have utilized these datasets makes it easier to compare and evaluate the effectiveness of the introduced model. The following are the two datasets that are used to evaluate the proposed model in detail:

1. The UCSD-Pedestrian 2 (Ped 2) dataset: is among the most utilized datasets for unsupervised video abnormality identification. This dataset was created in 2010 by researchers at the UCSD to capture regular pedestrian movement as well as occasional abnormalities in a controlled environment. Along this dataset, pedestrians crossed in front of the camera, and normal situations typically involve individuals strolling in walkways, whereas abnormalities are characterized by the appearance of unusual things (such as carts, bicycles, skateboards, etc.), unusual motion patterns (such as skating on a board) and walking on grass. Ped2 comprises 240 * 360 pixels resolution videos, with 12 videos that include 2010 frames for testing and 16 videos with 2550 frames for training. The videos, captured from a single scene at different times throughout the day, encompass variations in lighting and shadow effects, adding layers of complexity to the anomaly detection process.

2. The CUHK Avenue dataset: This dataset, which was created in 2013 by the Chinese University of Hong Kong (CUHK), offers a wider range of anomalies in an open campus environment. The clips are taken from a single view of the Chinese University of Hong Kong’s campus avenue. The dataset has 21 test videos and 16 training videos, totaling 15 328 frames for training and 15 324 frames for testing. All frames have a resolution of 360 by 640 pixels. 47 unusual activities were recorded, including object throwing, loitering, and running through the gate. Each training and testing clip in the CUHK Avenue dataset has a length of no more than 2 min.

A preprocessing step was performed before inputting the frames into the model, in which the input video frames were scaled to 112 by 112 pixels. Then, the input data is standardized by utilizing the centering and standard deviation methods to be ready for the model’s training and testing.

5.2. Evaluation metrics

Anomaly detection is a one-class estimation problem; samples with high probabilities (low reconstruction errors) are given the value 0, whereas anomalies with low probabilities (high reconstruction errors) are given the value 1. When compared with the normal class, abnormalities are statistically an uncommon class that only sometimes exists. The predicted frequency of anomalies’ appearance is the most typical way to describe this pattern (Kiran et al., 2018).

The classification of data points as normal or anomalous is a core objective shared by video abnormality identification techniques and other computer vision fields, despite some fundamental differences between them. Video anomaly detection is thus a binary classification problem that may be
handled (Nayak et al., 2021). Additionally, any of the fundamental components of performance assessment, such as true negative (TN), true positive (TP), false negative (FN), and false positive (FP), can be used to describe each binary classifier choice. These four performance metrics can be presented in a tabular format called a confusion matrix or matching matrix, as shown in Fig. 4, for both unsupervised and supervised learning, respectively. This error matrix can then be used to derive other advanced performance metrics, including Precision, Recall or True Positive Rate (TPR), and F1-Score, as shown in equations 9–11.

\[
\text{Recall} = \frac{TP}{TP + FN} \\
\text{Precision} = \frac{TP}{TP + FP} \\
\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]

The receiver operating characteristic (ROC) curve and Precision-recall (PR) curve are also two metrics that could be used for performance assessment of anomaly detection techniques (Nayak et al., 2021). Detection performance at different thresholds could be analyzed by constructing a plot of sensitivity (TPR) (in the Y-axis) vs. false positive rate (FPR) (in the X-axis). This plot is known as the ROC curve. The anomaly detection accuracy of a particular test set may be measured using the Area Under-ROC curve (AU-ROC) metric. Within the bounds of zero and one, the value of the AU-ROC ought to be maximized. The PR curve represents the relationship between recall (along the X-axis) and precision (along the Y-axis). Instead of using the area under receiver operating characteristics (AU-ROC), the area under precision-recall (AU-PR) curve should be used when trying to identify anomalies. This is due to the difficulty in detecting anomalies when there is an uneven distribution between normal and anomaly data, in which TNs are disproportionately big relative to TPs. In addition, the PR curve is more concentrated on positive or anomalous predictions. The highest obtainable value for the AU-PR is somewhere between zero and one. On the other hand, the number of model parameters and the floating-point operations (Flops), which are the theoretical measures of computational complexity that allow agnostic comparisons between models because they are independent of hardware and frameworks, were utilized to measure the proposed model’s efficiency.

5.3. Model configuration/implementation

To accomplish this research, a desktop computer equipped with a Core i7-11700F processor, 16 GB of RAM, and an NVIDIA GeForce RTX 3060 GPU with 12 GB of memory is utilized for investigations. On the TensorFlow-Keras deep learning framework, the spatial residual autoencoder architecture was developed, and the suggested system’s code was created on Ubuntu 20.04 using the Python Jupyter Notebook. The Adam optimizer (Kingma and Ba, 2014) is adopted to optimize the autoencoder, which is a simple and reliable method. In the 2D convolutional layers, ReLU activation was utilized because ReLU offers non-linearity and successfully learns the hidden patterns. The model was trained in batches, with a batch size of 16. The AU-ROC score is utilized to assess the suggested approach’s efficiency on the selected datasets. Table 2 includes the complete configuration of the proposed technique.

5.4. Experimental evaluation

As mentioned before, two benchmark datasets were utilized for the performance assessment of the proposed methodology: the UCSD Pedestrian dataset and the CUHK Avenue dataset. Fig. 5 displays some examples of these datasets. The training set of each dataset consists solely of normal frames, whereas the test set includes both normal and anomalous frames. In each test video, the ground

<table>
<thead>
<tr>
<th>Parameter</th>
<th>UCSD-Ped 2</th>
<th>Avenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Width</td>
<td>192</td>
<td>192</td>
</tr>
<tr>
<td>Height</td>
<td>192</td>
<td>192</td>
</tr>
<tr>
<td>Optimizer</td>
<td>SGD ‘Adam’</td>
<td>SGD ‘Adam’</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.01</td>
<td>1e-5</td>
</tr>
<tr>
<td>Loss</td>
<td>MSE</td>
<td>MSE</td>
</tr>
<tr>
<td>Epoch</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>Batch size</td>
<td>16</td>
<td>64</td>
</tr>
<tr>
<td>Stride</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
truth annotation comprises a binary flag per frame that indicates whether the frame contains an anomalous occurrence. Thus, label 0 represents the regular frame, whereas label 1 represents the anomalous frame. After that, to determine the anomaly score, the frame that represents the ground truth is compared with the frame that represents the model's prediction. AUC-ROC at the frame level was utilized to assess the effectiveness of the developed system because it is the most frequently used metric in state-of-the-art methods. The AUC-ROC value was determined by calculating the area under the ROC for the anomalous scores at various threshold settings. A higher AUC value shows greater effectiveness in anomaly identification. Additionally, other classification metrics were harnessed, as mentioned in the previous subsection, for more exploration of the introduced system.

Table 3 displays the AUC performance (%) achieved by the proposed model when applied to the UCSD-Ped 2 and CUHK Avenue datasets. Notably, the introduced spatial residual autoencoder model (SRAE) attained the highest performance compared with the spatial model only, reaching 95.41 % for the UCSD Ped2 dataset and 80.9 % for the CUHK Avenue dataset, respectively. This demonstrates that the developed residual model provided more information to encode the input frames. Furthermore, the adopted residual block architecture enabled the proposed model to overcome the degradation problem. This problem causes the accuracy to be saturated when the depth of the network grows. Hence, the residual block allowed the model to be lightweight with a minimum depth level. On the other hand, the spatial model without the residual block achieved better performance for the CUHK Avenue dataset than the residual one. As the CUHK Avenue dataset has too many frames with various unusual events, the residual connection probably causes the model to be less generalized as it reduces the model's capacity during training. Moreover, the CUHK Avenue dataset has many problems (Nguyen and Meunier, 2019), including a few outliers that are present in the training data, pedestrians who appear too far from the camera, people who walk toward and away from the camera, some events that are categorized as normal in the training data but are viewed as anomalous in the test data (for example, people walking on grass), and some challenging environmental conditions such as illumination and camera shaking in some videos that cause the detection performance to be slightly low for both the proposed residual and non-residual models. Hence, a deeper model or using temporal features besides the spatial ones can overcome this issue.

Moreover, the ROC curves for five test videos from the UCSD Ped 2 dataset and the CUHK-Avenue dataset are displayed in Figs. 6 and 7, respectively. The curves show how the proposed

Table 3. Comparison between the spatial model with and without residual block in terms of area under curve (%) for both the University of California, San Diego -Ped 2 and Chinese University of Hong Kong Avenue datasets.

<table>
<thead>
<tr>
<th>SRAE Model</th>
<th>UCSD-Ped 2</th>
<th>CUHK-Avenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Residual Block</td>
<td>86.28</td>
<td>82.19</td>
</tr>
<tr>
<td>With Residual Block</td>
<td>95.41</td>
<td>80.90</td>
</tr>
</tbody>
</table>

Fig. 5. Samples of normal and anomalous activities in the University of California, San Diego Ped 2 and the Avenue datasets.
SRAE model was trained well and how it can identify anomalies with high detection performance. Figs. 8 and 9 illustrate the visualization of two anomaly scores along with video frames for the two datasets; the animated version is available on GitHub (https://github.com/Mohamed-Habeb/Video-Anomaly-Detection-A-Lightweight-Model/tree/main) Anomaly scores (shown as blue lines in the figures) fluctuate quickly between normal and abnormal events, demonstrating that the introduced network can identify rare abnormal occurrences among numerous normal ones in a video. For the test 02 video from the UCSD-Ped 2 dataset, Fig. 8 illustrates the range of normal and abnormal occurrences’ anomaly scores. In the first two shots, people are seen strolling alone, but in the latter pair, a cyclist joins the group. Take note of how the anomaly score rises as soon as the rider enters the frame and how it stays relatively high until the rider leaves the frame. For the CUHK Avenue dataset, the test 15 video in Fig. 9, shows a person moving toward the camera, which is a very usual event in normal situations, but all normal events in this dataset are represented by people moving parallel or away from the camera. Hence, the anomaly event here is represented by a person moving in the wrong direction when he becomes close to the camera. Fig. 9 depicts the resulting change in the anomaly score.

In Table 4, the F1-score, recall, and precision metrics for some test videos from both the UCSD-Ped 2 and the CUHK Avenue datasets are presented. The results reveal promising and outstanding video abnormality detection performance. However, in anomaly detection, these metrics may not accurately reflect the models’ performance due to highly unbalanced data, as anomalous events are extremely rare in comparison to normal ones.

Table 5 provides a comparison of the proposed model to current state-of-the-art methods for the...
two datasets utilized in this study. Reconstruction-based techniques, where the model belongs, and prediction-based techniques are both broad categories into which these methodologies fall, and these are the two categories in consideration for comparison. Although the proposed model utilized only the appearance (spatial) features, the model shows an efficient and promising anomaly detection performance in comparison to the other techniques that use both spatiotemporal features, with an AUC of 95.41 % and 80.9 % for the UCSD-Ped 2 and the

![Fig. 9. The test 15 video’s anomaly score from the Chinese University of Hong Kong Avenue dataset. The shaded area shows the abnormality frames for a man moving towards the camera (wrong direction). The blue line represents the anomaly score which increases when an abnormal event appears.](image)

Table 4. F1-score, recall, and precision for some test videos for the two datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Video#</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1-score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCSD Ped 2</td>
<td>Test 02</td>
<td>97.56</td>
<td>93.02</td>
<td>95.23</td>
</tr>
<tr>
<td></td>
<td>Test 03</td>
<td>99.31</td>
<td>99.31</td>
<td>99.31</td>
</tr>
<tr>
<td></td>
<td>Test 04</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>Test 07</td>
<td>95.74</td>
<td>1.0</td>
<td>97.82</td>
</tr>
<tr>
<td></td>
<td>Test 12</td>
<td>80.90</td>
<td>77.42</td>
<td>79.12</td>
</tr>
<tr>
<td>CUHK Avenue</td>
<td>Test 05</td>
<td>90.44</td>
<td>87.85</td>
<td>89.13</td>
</tr>
<tr>
<td></td>
<td>Test 07</td>
<td>77.01</td>
<td>80.52</td>
<td>78.73</td>
</tr>
<tr>
<td></td>
<td>Test 15</td>
<td>97.83</td>
<td>1.0</td>
<td>98.90</td>
</tr>
<tr>
<td></td>
<td>Test 21</td>
<td>84.13</td>
<td>1.0</td>
<td>91.37</td>
</tr>
</tbody>
</table>

Table 5. Frame-level area under curve receiver operating characteristic comparison with recent state-of-the-art techniques for video abnormality identification (ST: SpatioTemporal, SP: Spatial, and TE: Temporal).

<table>
<thead>
<tr>
<th>Method</th>
<th>Reference</th>
<th>Model</th>
<th>Feature Type</th>
<th>UCSD-Ped 2</th>
<th>CUHK Avenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction</td>
<td>Li et al. (2019)</td>
<td>U-Net + ConvLSTM</td>
<td>ST</td>
<td>96.5</td>
<td>84.5</td>
</tr>
<tr>
<td></td>
<td>Lu et al. (2019)</td>
<td>Variational Autoencoder (VAE) + ConvLSTM</td>
<td>ST</td>
<td>96.0</td>
<td>85.7</td>
</tr>
<tr>
<td></td>
<td>Liu et al. (2018)</td>
<td>Least Square GAN (Mao et al., 2017) + U-Net</td>
<td>ST</td>
<td>95.4</td>
<td>85.1</td>
</tr>
<tr>
<td></td>
<td>Doshi and Yilmaz (2021)</td>
<td>Predictive neural network (PredNet)</td>
<td>ST</td>
<td>97.2</td>
<td>86.4</td>
</tr>
<tr>
<td></td>
<td>Zhou et al. (2019b)</td>
<td>U-Net + patch discriminator</td>
<td>ST</td>
<td>96.0</td>
<td>86.0</td>
</tr>
<tr>
<td></td>
<td>Le and Kim (2023)</td>
<td>Conv-encoder and a multi-stage channel</td>
<td>ST</td>
<td>97.4</td>
<td>86.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>attention-based decoder</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Taghinezhad and Yazdi (2023)</td>
<td>U-Net-like architecture + memorizer module</td>
<td>ST</td>
<td>97.6</td>
<td>89.0</td>
</tr>
<tr>
<td></td>
<td>Nawaratne et al. (2019)</td>
<td>Incremental Spatio-Temporal Learner (ISTL)</td>
<td>ST</td>
<td>91.1</td>
<td>76.6</td>
</tr>
<tr>
<td></td>
<td>Hu et al. (2019)</td>
<td>Ensemble Random Projection-based</td>
<td>ST</td>
<td>95.9</td>
<td>84.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reconstruction Loss Neural Network (E-RR-Net)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Li and Chang (2019)</td>
<td>Multivariate Gaussian Fully Convolution</td>
<td>ST</td>
<td>91.6</td>
<td>84.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Adversarial Autoencoder (MGFC-AAE)</td>
<td>ST</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Zhou et al. (2019a)</td>
<td>Sparse Long short-term memory (SLSTM)</td>
<td>ST</td>
<td>94.9</td>
<td>86.1</td>
</tr>
<tr>
<td></td>
<td>Fan et al. (2020)</td>
<td>Gaussian Mixture Fully Convolutional</td>
<td>ST</td>
<td>92.2</td>
<td>83.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Variational Autoencoder (GMFC-VAE)</td>
<td>ST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reconstruction</td>
<td>Deepak et al. (2021)</td>
<td>Residual Autoencoder</td>
<td>ST</td>
<td>83.0</td>
<td>82.0</td>
</tr>
<tr>
<td></td>
<td>Wang and Yang (2022)</td>
<td>Attention-based ConvLSTM network + Conv</td>
<td>ST</td>
<td>95.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Autoencoder</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wang and Chen (2023)</td>
<td>Dual-Stream Memory Network (DSM-Net)</td>
<td>ST</td>
<td>98.3</td>
<td>88.6</td>
</tr>
<tr>
<td></td>
<td>Zhang et al. (2023)</td>
<td>AE + DPU + Attention module</td>
<td>ST</td>
<td>97.9</td>
<td>85.9</td>
</tr>
<tr>
<td></td>
<td>Qi et al. (2023)</td>
<td>Noise AE + AE + Second-order Channel</td>
<td>ST</td>
<td>97.9</td>
<td>86.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Attention</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Feng et al. (2021)</td>
<td>Two-Stream Autoencoder</td>
<td>TE</td>
<td>84.0</td>
<td>77.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ST</td>
<td>84.5</td>
<td>80.3</td>
</tr>
<tr>
<td></td>
<td>The proposed model</td>
<td>Residual Autoencoder</td>
<td>SP</td>
<td>95.41</td>
<td>80.9</td>
</tr>
</tbody>
</table>
CUHK Avenue datasets, respectively. In particular, the introduced technique surpasses most contemporary abnormality determination algorithms in terms of AUC-ROC performance.

It is obvious that for the UCSD-Ped 2, the proposed model outperforms the models proposed by (Fan et al., 2020; Nawaratne et al., 2019; Li and Chang, 2019; Zhou et al., 2019a), in terms of AUC by +4.3 %, +3.8 %, +0.5 %, and +3.2 %, respectively. Although these models used a more sophisticated and complex architecture that exploited the spatio-temporal features in videos rather than the proposed one, which exploits only the spatial features and is a lightweight model with only 3 layers for the encoder and decoder parts, the proposed model is superior to them. For instance, (Nawaratne et al., 2019) used an Incremental Spatio-Temporal Learner, (Li and Chang, 2019) utilized a multivariate Gaussian Fully Convolution Adversarial Autoencoder, (Zhou et al., 2019a) adopted sparse long short-term memory, and (Fan et al., 2020) used a Gaussian Mixture Fully Convolutional Variational Autoencoder. Despite the authors (Deepak et al., 2021) utilizing Spatiotemporal Residual Autoencoder, the developed spatial-only residual autoencoder surpasses it by +12.4 % in the AUC term. In addition, the introduced model outperforms the Spatio-temporal two-stream autoencoder and its individual spatial and temporal streams that were introduced in (Feng et al., 2021) by achieving +11.6 %, +11.4 %, and +10.9 % versus their spatially, temporally, and Spatio-temporal streams, respectively.

For the CUHK Avenue dataset, the proposed model has better results than the Incremental Spatio-Temporal Learner introduced in (Nawaratne et al., 2019) by +4.1 % and the temporal, spatial, and spatiotemporal streams in (Feng et al., 2021) by +3.1 %, +2.8 %, and +0.6 %, respectively. It is noted that the other techniques have better results than the introduced model for this dataset, but the proposed lightweight model still has promising results as it depends only on spatial features, while all other techniques utilize Spatio-temporal features. Moreover, the obtained results are not far from theirs.

The effectiveness of the suggested model in terms of both the number of model parameters and Flops measures is demonstrated in Table 6 by comparing the proposed lightweight model and state-of-the-art pre-trained models trained on ImageNet-1K. It is worth noting that all pretrained models in Table 6 were finetuned on ImageNet-1K for the supervised classification task rather than the proposed model which is an unsupervised detection task. In conclusion, the comparison with the most advanced pretrained models highlighted the model's computational efficiency more comprehensively.

Table 6. Comparison between the flops and parameters of models trained on ImageNet-1K.

<table>
<thead>
<tr>
<th>Model</th>
<th>Flops (G)</th>
<th>Parameters (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG-13</td>
<td>11.34</td>
<td>133.05</td>
</tr>
<tr>
<td>VGG-16</td>
<td>15.5</td>
<td>138.36</td>
</tr>
<tr>
<td>VGG-19</td>
<td>19.67</td>
<td>143.67</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>3.68</td>
<td>21.8</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>4.12</td>
<td>25.56</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>7.85</td>
<td>44.55</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>11.58</td>
<td>60.19</td>
</tr>
<tr>
<td>Inception-v3</td>
<td>5.75</td>
<td>23.83</td>
</tr>
<tr>
<td>Swin-tiny</td>
<td>4.36</td>
<td>28.29</td>
</tr>
<tr>
<td>Swin-small</td>
<td>8.52</td>
<td>49.61</td>
</tr>
<tr>
<td>Swin-base</td>
<td>15.14</td>
<td>87.77</td>
</tr>
<tr>
<td>Swinv2-tiny-w8</td>
<td>4.35</td>
<td>28.35</td>
</tr>
<tr>
<td>Swinv2-tiny-w16</td>
<td>4.40</td>
<td>28.35</td>
</tr>
<tr>
<td>Swinv2-small-w8</td>
<td>8.45</td>
<td>49.73</td>
</tr>
<tr>
<td>Swinv2-small-w16</td>
<td>8.57</td>
<td>49.73</td>
</tr>
<tr>
<td>Swinv2-base-w8</td>
<td>14.99</td>
<td>87.92</td>
</tr>
<tr>
<td>Swinv2-base-w16</td>
<td>15.14</td>
<td>87.92</td>
</tr>
<tr>
<td>The proposed model</td>
<td>1.65</td>
<td>4.85</td>
</tr>
</tbody>
</table>
applications, the computation efficiency of a model is just as important as accuracy, especially when the application requires real-time processing, such as in video surveillance. Through an assessment of the number of parameters and Flops, the model could be verified to be computationally efficient and correct.

The main difference between the proposed spatial-only model and recent models in the literature is its low complexity due to its simple architecture, as it incorporates a 3-layer Conv. Autoencoder with only one residual block. The spatiotemporal model proposed by (Deepak et al., 2021) has 3-residual blocks with multiple Conv-3D and ConvLSTM layers, compared with 1-residual block in the introduced model, which achieved very high performance for the UCSD-Ped 2 dataset and a comparable one for the CUHK Avenue dataset. The model in (Feng et al., 2021) adopted a two-stream autoencoder with spatiotemporal features in which the first stream is a spatial 3-layer Conv. Autocoder, and the second one is a 3-layer autoencoder with 3 ConvLSTM layers, and late fusion is then applied. Despite the complex architecture of this model, the introduced lightweight model beat it in the performance of anomaly detection for both datasets. Moreover, the model in (Wang and Yang, 2022) used a 4-layer attention-based ConvLSTM network and 4-layer Conv. Autoencoder to extract the spatiotemporal features. The detection performance for the UCSD-Ped2 dataset is almost the same as the obtained one from the proposed model, and the authors didn’t utilize the CUHK Avenue dataset in their experiments. In (Zhang et al., 2023), the authors adopted DPU as the baseline model with coding and decoding phases and dual attention modules, which are the main time-consuming parts, to obtain video sequences’ contextual information. Although the model has many stages and outperforms the proposed model’s performance, the gained results here are not far from theirs. The framework developed by (Taghinezhad and Yazdi, 2023) was composed of four main components that work together to predict future frames in a video sequence. The first component is a Time-Distributed-2D encoder of 4 Conv. blocks that extracts spatial features from stacked frames. The second component, called Multi-Scale-Memorizer, stores and retrieves semantic information at different scales. The third component, the predictor, is composed of 3 Conv. GRU to preserve temporal patterns in the sequence. Finally, the decoder with three Conv. blocks generates the predicted output frame. This spatiotemporal complex framework was called MsMp-net, and its performance was promising in both datasets, but the introduced simple and lightweight model was competitive, especially for the UCSD-Ped2 datasets.

Regarding the performance aspect, it could be derived that the introduced model has about 4.8 M parameters and 1.6G FLOPs, which is much less in comparison to the pretrained models in Table 6, taking into consideration that those pretrained models probably incorporated in the literature as backbone models or even as preprocessing steps besides other architectures parts or streams, which proves how the proposed model is lightweight with efficient performance in comparison to the state-of-the-art. Moreover, the robustness of the model in recognizing anomalous occurrences in recordings is rigorously evaluated using two challenging datasets, revealing reliable performance across various scenarios, including congested situations. Furthermore, the model achieved significant AUC-ROC percentages for the UCSD-Ped 2 and CUHK Avenue datasets using a 3-layer spatial residual encoder-decoder architecture, which is not complex architecture such as architecture in recent studies.

On the other hand, there are some limitations regarding performing this study. The model’s ability to identify anomalies in unknown or varying scenarios may be compromised if the training data isn’t representative of real-world circumstances. This is because autoencoders rely heavily on the training data. Furthermore, deep learning models such as autoencoders might operate as black boxes. Consequently, it can be hard to figure out why they indicate something as abnormal, which may not be desired in situations where interpretability is essential. Moreover, the concept of an ‘anomaly’ can evolve in dynamic environments. Hence, certain events might be regarded as abnormal in some contexts but not in others. As a result, conventional autoencoders may not cope with these changes if they are not frequently retrained.

The comparison with state-of-the-art approaches reveals that using residual blocks is more successful than using deep extra layers to train a spatiotemporal autoencoder. In addition, the introduced model employed residual block to address the problem of vanishing gradients, which helps the model to generalize well during training. Moreover, the recent studies focused on the accuracy detection of anomalies in videos, ignoring the performance of the model, which is in high demand for the time being due to the necessity for real-time video anomaly detection systems that work in a fast and efficient manner with low hardware capabilities to exploit the massive video data generated by CCTV cameras. Therefore, the proposed lightweight end-
to-end unsupervised system for anomaly identification in videos using a deep residual autoencoder is emphasized as a promising contribution with enhanced anomaly detection accuracy and high performance.

7. Conclusion and future work

In this work, an unsupervised lightweight technique for video anomaly identification based on residual autoencoder architecture is introduced. The suggested model can extract and learn the spatial features of normal data that are different from anomalous data. Thus, the (low/high) reconstruction error between the input and reconstructed frames is what determines the (regular/anomalous) events. The study's analysis reveals that the system's performance has been vastly enhanced by using the residual connection with a doubled number of units for the decoder in comparison to the encoder's units. The residual connection proved its efficiency in the proposed model by overcoming the degradation problem that makes the accuracy get saturated when the network's depth increases and also allows the model to be lightweight with a minimum depth level. To rigorously evaluate the robustness of recognizing anomalous occurrences in recordings, several experiments were conducted using two challenging datasets that revealed that the suggested technique reliably works for a variety of scenarios, including congested situations. The model achieved 95.41% and 80.9% in terms of AUC-ROC for the UCSD-Ped 2 and CUHK Avenue datasets, respectively. Based on the experimental results, we can confidently assert that the proposed approach is both precise and robust, and it is comparable to state-of-the-art approaches. Furthermore, because the proposed approach is a lightweight and entirely spatial-based model, it has the potential to serve as the basis for a real-time anomaly detection system. To be precise, the two benchmark-used datasets don't reflect some surveillance scenarios, such as crimes, so the introduced model may or may not detect anomalies in these scenarios. Hence, the matter needs more investigations for datasets that include such situations. Moreover, it is noted that the videos in the datasets suffer from low-quality resolution which in turn could negatively affect the performance of the models, so we believe that enhancing the resolution using one of the image super-resolution techniques will enhance the results of the proposed model or any other model that utilizes these datasets.

In future work, we intend to take advantage of the temporal features and combine them with the spatial features, investigate emerging technologies such as vision transformers and skeletal deep learning video anomaly detection methods to enhance the abnormality detection efficiency, especially for the CUHK Avenue dataset, and expand the work on another benchmark anomaly dataset. Moreover, we plan to use explainable deep learning techniques, which provide exceptional advantages for understanding and dealing with anomalies and even preventing anomalous events from taking place in advance.

Authors contributions

Conceptualization, L.A.E., M.A.S., and M.H.H.; Methodology/study design, L.A.E., M.A.S., and M.H.H; Software, M.H.H; Validation, L.A.E., M.A.S., and M.H.H; Formal analysis, L.A.E., M.A.S., and M.H.H; Investigation, L.A.E., M.A.S., and M.H.H; Writing-original draft, M.H.H; Writing-review and editing, L.A.E., and M.A.S.; Visualization, L.A.E., and M.H.H.; Supervision, L.A.E., and M.A.S. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflicts of interest

The authors declare no conflict of interest.

References


