algorithms

MDPI

Article

Enhancing Video Anomaly Detection Using a Transformer Spatiotemporal Attention Unsupervised Framework for Large Datasets

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

Electrical Engineering Department, Faculty of Engineering at Shoubra, Benha University, Cairo 11629, Egypt; mohamed.rehan@feng.bu.edu.eg (M.H.H.); may.mohamed@feng.bu.edu.eg (M.S.)
* Correspondence: lamia.alrefaai@feng.bu.edu.eg

Abstract: This work introduces an unsupervised framework for video anomaly detection, leveraging a hybrid deep learning model that combines a vision transformer (ViT) with a convolutional spatiotemporal relationship (STR) attention block. The proposed model addresses the challenges of anomaly detection in video surveillance by capturing both local and global relationships within video frames, a task that traditional convolutional neural networks (CNNs) often struggle with due to their localized field of view. We have utilized a pre-trained ViT as an encoder for feature extraction, which is then processed by the STR attention block to enhance the detection of spatiotemporal relationships among objects in videos. The novelty of this work is utilizing the ViT with the STR attention to detect video anomalies effectively in large and heterogeneous datasets, an important thing given the diverse environments and scenarios encountered in real-world surveillance. The framework was evaluated on three benchmark datasets, i.e., the UCSD-Ped2, CHUCK Avenue, and ShanghaiTech. This demonstrates the model's superior performance in detecting anomalies compared to state-of-the-art methods, showcasing its potential to significantly enhance automated video surveillance systems by achieving area under the receiver operating characteristic curve (AUC ROC) values of 95.6, 86.8, and 82.1. To show the effectiveness of the proposed framework in detecting anomalies in extra-large datasets, we trained the model on a subset of the huge contemporary CHAD dataset that contains over 1 million frames, achieving AUC ROC values of 71.8 and 64.2 for CHAD-Cam 1 and CHAD-Cam 2, respectively, which outperforms the state-of-the-art techniques.

Keywords: video anomaly detection; unsupervised learning; spatiotemporal modeling; large datasets

Citation: Habeb, M.H.; Salama, M.; Elrefaei, L.A. Enhancing Video Anomaly Detection Using a Transformer Spatiotemporal Attention Unsupervised Framework for Large Datasets. *Algorithms* **2024**, *17*, 286. https://doi.org/10.3390/a17070286

Academic Editor: Arslan Munir

Received: 7 May 2024
Revised: 22 June 2024
Accepted: 23 June 2024
Published: 1 July 2024

1. Introduction

Video surveillance systems have become an essential component of our security infrastructure, playing a critical role in monitoring and ensuring safety in various settings, like public spaces, transportation systems, and commercial establishments. Since surveillance systems are used in various areas and environments, it has always been difficult to identify abnormal activity accurately and quickly in recordings. Further, more than 1 billion security cameras were used worldwide in 2023 [1], and that number is projected to exceed 2.24 billion by 2030, with a compound annual growth rate of 12.2% [2]. This growing volume of video data generated by these systems demands an efficient and reliable method for anomaly detection, which remains a challenging task due to the dynamic and complex nature of video content. Traditional approaches often rely on manual monitoring or primitive automated techniques, which are time consuming and prone to errors and inefficiencies. Since abnormal activities are rare compared to normal ones, identifying and detecting anomalies is more complicated than other video analysis forms because of the data imbalances between normal and abnormal segments. This has increased interest in leveraging advanced machine learning methods, particularly deep learning, to automate and improve anomaly detection accuracy in video surveillance.

Video anomaly detection involves identifying events or patterns in video data that deviate from the norm. These abnormalities might be anything from identifying safety risks in industrial settings to strange behaviors in crowded areas. The critical challenge in video anomaly identification is the wide range of behaviors that are considered normal in various contexts but anomalies in others; for instance, running is a typical activity for people as part of an exercise routine but running inside a bank could be perceived as anomalous or suspicious behavior. Moreover, anomalous events are complicated in nature, are often rare, and vary greatly in appearance and nature. Traditional machine learning and deep learning methods are the two main categories of anomaly detection approaches. Traditional machine learning techniques have shown impressive results in many video anomaly areas, such as abnormal human action recognition, by capturing shallow features from video data [3]. These techniques included support vector machines (SVMs) [4], Markov models [5], Bayesian networks [6], random forests (RFs) [7], probabilistic-based models [8], sparse reconstruction [9], histograms of optical flows (HOFs) [10], and histograms of oriented gradients (HOGs) features [11,12]. On the other hand, they mostly rely on handcrafted features and preprocessing, which take a lot of time and resources to complete. According to Hu et al. [13], they perform poorly in real-world scenarios and do not scale well for various datasets. The use of deep learning techniques introduced novel strategies that surpass conventional approaches and tackle the drawbacks related to traditional machine learning [14–16]. We do not address traditional machine learning techniques in this work; instead, we concentrate on the most recent advancements in deep learning models.

Recent improvements in deep learning have produced effective techniques for analyzing videos, which have significantly improved anomaly identification. These techniques, in particular, convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated remarkable potential in terms of capturing the temporal dynamics and spatial characteristics of videos. Nevertheless, they typically require an extensive amount of labeled data and have trouble capturing the nuances and variability of anomalies, which can result in errors like missed detections and false positives. Notably, the attention mechanism in deep learning allows models to focus on the most relevant parts of the input data, enhancing their ability to learn context-dependent features for tasks such as language translation, image recognition, and sequence prediction. It dynamically weighs the significance of different input components, improving the model's interpretability and performance by mimicking cognitive attention in human learning. For example, Tian et al. [17] showed how dilated convolutions and self-attention may be used in weakly-supervised video anomaly detection, emphasizing how well they can capture temporal information and enhance the discriminability of anomalies. Furthermore, the incorporation of attention mechanisms and memory units into transformer models has recently been investigated in more detail to improve the models' effectiveness in video anomaly detection. [18] used the transformer's network capability to maintain global and local associations in videos to offer a method for weakly supervised video anomaly detection utilizing "Dual Memory Units with Uncertainty Regulation".

Deep learning-based video anomaly detection (VAD) models presently employ unsupervised or weakly supervised learning, since collecting anomalous data is difficult. Unsupervised learning techniques often rely on autoencoders [19–21] or pre-trained CNNs [22,23] to extract features, and then identify anomalies by reconstructing frames or predicting future frames. For weakly supervised learning techniques, only video-level labels are needed [24–26] which use multi-instance learning to differentiate between normality and abnormality. Hence, this research focuses on unsupervised approaches since they are more appropriate for real-world application settings.

The introduction of the transformer model, basically developed for natural language processing tasks, has opened new directions in computer vision, including video anomaly detection. With their ability to focus on the most relevant portions of the input data, vision transformers (ViTs), first introduced by Dosovitskiy et al. [27], provide a paradigm change

away from the inductive biases of CNNs. ViTs can capture global context and long-range interdependence thanks to their attention mechanism, which is very useful for interpreting complex scenarios in video surveillance. In addition, ViTs learn robust and discriminative features that are not limited to local neighborhoods, as CNNs are, which accounts for their efficacy in anomaly detection.

Based on the aforementioned reasons and to take advantage of both ViTs and CNNs, we propose an unsupervised frame-based spatiotemporal video anomaly detection technique that automatically detects anomaly frames in surveillance videos by utilizing a pre-trained ViT transformer model with the attention of spatiotemporal relationships among objects (STR attention model) [28]. The proposed model was trained and validated over three video anomaly benchmark datasets in the literature; the UCSD-Ped2 [29] dataset, the CHUCK Avenue [30] dataset, and the largest benchmark available, the sophisticated ShanghaiTech [31] dataset. In addition, we challenged the proposed model against the recent, very large Charlotte Anomaly Dataset (CHAD) [32] to prove its efficacy with very large video anomaly datasets. Due to the lack of intrinsic inductive biases of CNNs in ViTs, the STR attention model allowed better capturing of the spatiotemporal relationships among objects in successive frames, including spatial locations, movement speeds and directions, and morphological changes, which enhanced the detection accuracy of abnormal frames. In addition, the existing self-attention ViT-based models use the transformers to capture relationships between long-range pixels and the global contexts of the images. Hence, we utilized the STRA model which is a spatiotemporal-based model that leverages these features by obtaining the relationships among objects in the input sequence of the fames, which strengthens the learned features of the normal videos' events and situations. This allows the proposed framework to strongly reconstruct normal frames and detect the poorly reconstructed frames as anomalous. Moreover, to tackle the problem of detection accuracy degradation when training the vision transformers on small and medium-sized datasets, an advanced training strategy was applied by dually combining the datasets and grouping all of them to form various synthetic heterogeneous datasets, then training the proposed model on the datasets individually and on the different combinations that improved anomaly detection accuracy for the models trained on the combined datasets. In addition, two different up-sampling approaches are introduced to enlarge the size of the smaller datasets, i.e., the UCSD-Ped2 and CHUCK Avenue. This strategy proved the robustness and effectiveness of the introduced spatiotemporal ViT-based model in identifying anomalies in larger heterogeneous video anomaly datasets captured in different environmental and lighting conditions and varied resolution and quality, which is a step towards developing a real-time video anomaly detection multi-modal surveillance system. The contributions of this work are summarized below:

1. An unsupervised framework for video anomaly detection-based ViT transformer and spatiotemporal block using an encoder–decoder architecture to address and effectively detect diverse video abnormalities;
2. Proposing a hybrid deep learning framework which is, to the best of our knowledge, the first one to combine the ViT model with the convolutional STR attention block for video anomaly detection;
3. Adopting a different training strategy by creating different combinations of the utilized datasets to improve the performance of the proposed model for identifying anomalies in videos, despite the diverse environmental difficulties, lighting settings, and video quality of these datasets, towards developing a video anomaly detection multi-modal surveillance system;
4. Conducting a comprehensive evaluation using publicly available video anomaly detection datasets for the introduced framework shows that when compared to cutting-edge methods, the proposed approach produces outstanding results.

The rest of this paper is structured as follows: Section 2 covers related VAD work, while Section 3 provides background information about the underlying used techniques. Details about the proposed model are provided in Section 4, while the experiments using

the introduced model, their findings on the datasets, and the discussion are shown in Section 5. Finally, the conclusion is in Section 6.

2. Related Work

In the next two subsections, we present the related work from the point of view of approaches based on vision transformer (ViT) (Section 2.1) and video anomaly detection works based on other techniques (Section 2.2). The former leverages the recent advancements in transformers applied to video processing, while the latter encompasses a variety of other methods developed for identifying anomalies in video data, such as CNNs, LSTMs, etc.

2.1. Vision Transformer-Based Approaches

Vision transformers (ViTs) have emerged as a powerful tool in video anomaly detection, offering new perspectives and capabilities. Yuan et al. [33] demonstrated the usefulness of combining the "Video Vision Transformer" (ViViT) [34] with neural architectures like U-Net [35] for enhanced performance in anomaly detection. They have shown significant improvements in handling complex video data, achieving higher accuracy in detecting anomalies. Other notable works in this area include [36,37]. The former dealt with road accident detection from dashboard cameras using ViTs, while the latter provided work on violence detection in videos. These works highlight the diverse applications of ViTs in surveillance, from detecting violent incidents to monitoring road safety, reflecting a significant shift towards more sophisticated, AI-driven surveillance systems. Tahir and Anwar [38] explored the application of vision transformers for pedestrian image retrieval and person re-identification in multi-camera systems, demonstrating the effectiveness of combining vision transformers with CNN models. Lee and Kang [39] presented AnoViT, a vision transformer-based encoder–decoder model to overcome the shortcomings of conventional convolutional encoder–decoders. The model exploited image anomaly detection and localization by learning both local and global relationships between image patches.

Leveraging vision transformers' success in deep learning and imagine classification, Berroukham et al. [40] suggested a strategy by fine-tuning a pre-trained vision transformer "ViT-B16" model to divide video frames into groups for normal and abnormal behavior. Another vision transformer-based framework for anomaly recognition in smart-city surveillance videos, called "ViT-ARN", was introduced by Ullah et al. [26]. The goal of the framework was to address the limitations of existing automated surveillance systems and it was composed of two phases: online anomaly detection, using a lightweight one-class deep neural network, and anomaly classification, using a vision transformer and a bottleneck attention mechanism. In addition, a multi-reservoir echo-state network was used to enable the assessment of real-world anomalies such as vandalism and road accidents. Furthermore, the ViT model was employed by Lee et al. [41] to carry out spatiotemporal context-based video anomaly detection in surveillance videos. The introduced model consisted of a contextual appearance module and a motion reconstruction module to concentrate on the masked, whole, and partial contextual prediction streams. An enhanced time-series vision transformer (TSViT) was developed by Lee et al. [42] to identify any irregularities in the pedestrians' (victim and follower) gait. In this model, the pedestrians' spatial information is encoded into 2D patterns, which are then passed to the TSViT as tokens. After that, the TSViT is regularized to enable training for small datasets. In addition, the use of ViT in the novelty detection of traffic scenario infrastructure was examined by Wurst et al. [43]. They converted the output of Vanilla ViT from predictive class labels to latent representations to fine-tune it. Furthermore, they used ViT to create latent representations for input road infrastructure pictures within a triplet loss-based autoencoder framework.

To alleviate the catastrophic forgetting problem in deep learning models, which causes a substantial reduction in overall performance when additional classes are added progressively during training, Fan et al. [44,45] established a contrastive learning approach for ViT. The approach used ViT as a feature extractor and performed image anomaly detection step-by-step using a contrast learning framework. A self-supervised model called UNETR

was utilized by Park et al. [46] to address the issue of insufficiently labeled data leading to inaccurate out-of-distribution detection (OOD). The model is a 3D UNET, with the ViT structure being employed as an encoder. Then, they used a skip-connection structure to connect the input pictures to the decoder after converting them into sequence representations. They noted that self-supervised learning is crucial to the study of medical OOD identification, as certain anomalous images of uncommon diseases are extremely challenging to produce. In the construction area, Lin et al. [47] rebuilt unlabeled pavement pictures using the ViT-S self-supervised learning model. They suggested a pavement anomaly detection technique based on encoding, retrieval, and matching to deal with the classification retraining issue.

The authors of [48] introduced a unique model for anomaly detection called "ViV-Ano" that is based on an encoder–decoder design, where the encoder comprises a variational autoencoder (VAE) and ViT to extract both local and global characteristics from the images for anomaly identification in industrial processes. Work for low-resolution image-based anomaly detection and localization in industrial fields was proposed by Smith et al. [49]. The work investigated the use of ViT in the classification and localization of surface defects in leather. Furthermore, Yao et al. [50] presented an approach to identifying logical flaws under sophisticated semantic conditions in the industry area, particularly for defect inspection industrial tasks involving printed circuit boards (PCBs). The provided method used a pre-trained network for multi-scale prior embeddings, followed by a vision transformer with dual attention mechanisms for global–local two-level reconstruction. A masked multi-head self-attention method was used by De Nardin et al. [51] to enable the model to understand the link between various input picture patches for anomaly detection in the industrial quality control field. To accomplish high-precision picture anomaly detection, they added a new masking component to the ViT architecture and adjusted the attention between patches of varying forms. A hybrid transformer model called ViTALnet was introduced by Tao et al. [52]. It was constructed based on fine-grained feature reconstruction. To leverage the global semantic capturing capacity, ViTALnet used the vision transformer (ViT) to derive the local discriminating characteristics as feature representation. Then, a pyramidal design and global attention mechanisms were combined to create an anomaly estimation module that improves contextual data for the localization of fine-grained industrial anomalies. By using a ViT to encode the input image that has been split into fragments, VT-ADL [53] was able to obtain the features that represent the normal image and perform the image abnormality identification task for industrial textured surfaces from various fine-grained levels in a reconstruction-based manner. In addition, a Gaussian mixture density function was also introduced for pixel-level picture anomaly localization. Hence, the distribution of ViT-encoded features is modeled using the Gaussian mixture density function to model the distribution of normal data in the potential space.

2.2. Other Techniques in Video Anomaly Detection

Beyond ViTs, various other techniques are explored in video anomaly detection. This includes research on using neural networks, CNNs, and LSTM networks for anomaly detection in video surveillance. Franklin and Dabbagol [54] achieved a high accuracy rate in their neural network-based anomaly detection system, indicating the potential for real-time monitoring applications. Ullah et al. [55] presented a framework using CNN features and bi-directional LSTM, showing improved accuracy in benchmark datasets. Moreover, Qi et al. [56] introduced a dual-generator generative adversarial network method for detecting anomalies by learning the anomaly distribution in advance. It employs a noise generator for creating pseudo-anomaly frames and a reconstruction generator to learn normal video frame distribution. In addition, a second-order channel attention module enhances learning capacity. Furthermore, a self-supervised predictive architectural building block was developed by Ristea et al. [57]. The block has a convolutional layer that masks the central region of the receptive field using dilated filters. Then, the generated activation maps are transmitted through a channel attention module. The block has a loss function that reduces the reconstruction error for the masked region in the receptive field.

To replace the conventional two-stream network, a convolutional recurrent autoencoder (CR-AE) integrating an attention-based convolutional long short-term memory (ConvLSTM) network with a convolutional autoencoder to simultaneously capture both temporal and spatial irregularities was introduced by Wang and Yang [58]. Another attention encoder–decoder-based framework was developed by Li et al. [59], which used the corresponding intermediate layers from the encoder and decoder to train the model simultaneously. The authors pointed out that this improved the model's ability to detect anomalies by enabling it to capture more complex characteristics. Besides, the motion feature extraction was performed using a motion loss function that uses the actual video frames instead of optical flow with a parameter-free variance attention method that concentrates attention on moving objects. Furthermore, Wang et al. [28] utilized a fully convolutional encoder–decoder network with symmetric skip connections to discover the spatiotemporal relationships among objects. The authors implemented an attention mechanism during the encoding phase to improve the understanding of the spatiotemporal relationships among different object types. In addition, a dynamic pattern generator in the decoding phase was designed to memorize these relationships, thereby improving the reconstruction of normal samples and making abnormal sample reconstructions more challenging. Another encoder–decoder framework called SMAMS was proposed by Fu et al. [60] and built using memory modules and a spatiotemporal masked autoencoder. It utilized spatiotemporal cubes and extracted the spatiotemporal properties of the video events by using a spatiotemporal masked autoencoder. Then, memory modules were used to maintain unmasked video patches of various feature layers with skip connections to recompense for crucial information loss.

A method that used parallel spatial-temporal CNNs to address the unusual distribution of information in video frames was introduced by Hu et al. [61]. They utilized an optical flow algorithm combined with a varied-size cell structure to segment spatial-temporal interest blocks containing moving objects. A parallel 3D-CNN was utilized to describe the same behavior at different temporal lengths, ensuring comprehensive capture of behavior information while reducing irrelevant data. Another 3D-CNN-based work proposed by Hwang and Kang [62] focused on both spatial and temporal data for detecting anomalies, specifically violent behaviors, in videos. It incorporates a convolutional block attention module (CBAM) to improve interpretability and focus on crucial information in the video data. Unlike traditional 3D-CNNs that use multiple frames simultaneously, this method merges various frames into one image, reducing memory usage to enable more precise detection of anomalies.

A weakly supervised video anomaly detection technique to bridge the gap between normal and abnormal instances and tackle the problem of false alarms was proposed by Lee et al. [63]. They utilized a unique multiple instance learning (MIL) framework based on a memory unit that comprised a multi-head attention feature augmentation (MHFA) module, a loss function with KL divergence, and Gaussian distribution estimation. Kotkar and Sucharita [64] proposed a classification approach called MST-RNN-LSTM, "Modified Spatiotemporal Recurrent Neural Network using Long Short-Term Memory", for detecting anomalies in video surveillance as part of IoT-based smart city initiatives. Cuboids were created by processing the normalized frames for motion tracking, and a discrete wavelet transform (DWT) with principal component analysis (PCA) was applied to the cuboids. The training of the features and the classification were performed by an RNN–LSTM model. A similar work for video anomaly detection in surveillance systems, proposed by Taghinezhad and Yazdi [65], utilized a 2D CNN-based decoder, a time-distributed 2D encoder, and an architecture resembling U-Net. In addition, a memory module was included to save the most pertinent prototypical patterns of normal activities to facilitate inaccurate predictions for abnormal inputs. In summary, the video anomaly detection domain is rapidly evolving, with significant contributions from transformer-based and other advanced techniques. These developments are substantial for enhancing the capabilities of surveillance systems and contributing to public safety and security.

A weakly supervised data augmentation network was conducted by Lei et al. [66] to support attention-guided data augmentation and enrich the input pictures. The method used the multi-scale feature extraction technique to extract visual information from video footage at different scales. Then, the model was trained by incorporating the enhanced convolutional block attention module (CBAM) into the base U-Net architecture to suppress interference from non-anomalous areas in the video, enabling the model to focus on anomaly-relevant regions.

The state of the art shows that recent research focuses on vision transformers as leading models for detecting video anomalies in different fields, but there is a great demand for intelligent surveillance systems for preserving public safety and security.

3. Background

There is an increasing need for reliable anomaly detection systems in the rapidly developing industry of video surveillance. Conventional methods have frequently relied on techniques that work well for identifying local patterns but poorly in terms of fully understanding complex and dynamic situations. This research investigates a novel hybrid deep learning model that combines the advantages of autoencoders, CNNs, and vision transformers (ViTs) to overcome these drawbacks. Through the integration of these technologies, the proposed model is able to improve the accuracy and reliability of anomaly identification in a variety of surveillance contexts, while also capturing the complex spatiotemporal correlations present in videos. This background serves as the foundation upon which our hybrid model is built, enabling it to achieve superior performance on benchmark datasets and offering a comprehensive solution to the challenges faced in video anomaly detection.

3.1. Vision Transformers

The ViT emerged as a novel deep learning model initially developed for natural language processing (NLP) tasks [67], which has acquired attention recently in the field of image analysis, especially for anomaly detection in images. ViTs deviate from the conventional deep learning models by employing self-attention mechanisms rather than the traditional convolutional layers, enabling them to capture the global dependencies among image patches effectively. This model processes an image by dividing it into non-overlapping patches, which are then projected into vectors through linear projection. The integration of positional encodings with these patch embeddings ensures that the model retains the spatial context of the image, which is crucial for understanding the global layout of the image components. In contrast to traditional methods, a ViT's self-attention mechanism in the transformer encoder enables it to concentrate on various regions of the picture according to contextual importance, which enables a more comprehensive and nuanced analysis.

After the preliminary processing, the vision transformer improves the feature extraction process by using residual connections, layer normalization, and multi-head self-attention processes. This enhances the model's ability to identify complex patterns and relationships in the analyzed image. This advanced architecture enables the transformer to process the image patches in a manner that captures both local and global contextual information, leading to highly accurate image classification outcomes. With a multi-layer perceptron (MLP) head acting as the classification layer at the end of the model, the input images are given class labels according to the characteristics that the transformer encoder retrieved. The ViT's unique approach, which makes use of the self-attention mechanism for image classification tasks, represents a significant shift from traditional convolution-based methods, offering a promising new direction for advancing the capabilities of deep learning models in image analysis and other fields.

The vision transformer (ViT) architecture, as shown in Figure 1 [68], could be summarized in several key stages, described as follows:

1. Patch extraction: the input image is divided into fixed-size patches (e.g., 16 × 16 pixels). This step transforms the 2D image into a sequence of flattened 2D patches;
2. Patch embedding: via a trainable linear projection, each extracted patch is projected linearly into a higher-dimensional vector (embedding). This procedure modifies the image patches so the transformer can process them, like embedding tokens in NLP tasks;
3. Positional encoding: positional encodings are added to the patch embeddings to retain the positional information of each patch. This step ensures the model can recognize where each patch is located in the image, which is crucial for understanding the overall structure of the image;
4. Transformer encoder: this is the core of the ViT. This encoder consists of multiple layers, each containing two main components, a multi-head self-attention mechanism and a position-wise feed-forward network. Layer normalization is applied before each component, and residual connections are used after each component;
5. Multi-head self-attention: this mechanism allows the model to weigh the importance of different patches relative to each other, enabling the capture of global dependencies across the entire image;
6. Feed-forward network: after attention aggregation, each patch embedding is processed independently by a feed-forward network, which allows for nonlinear transformations of the patch representations;
7. Classification head: finally, the output from the transformer encoder is passed through a classification head, typically a multi-layer perceptron (MLP), to produce the final class predictions. This head processes the global representation of the image derived from the concatenated patch embeddings.

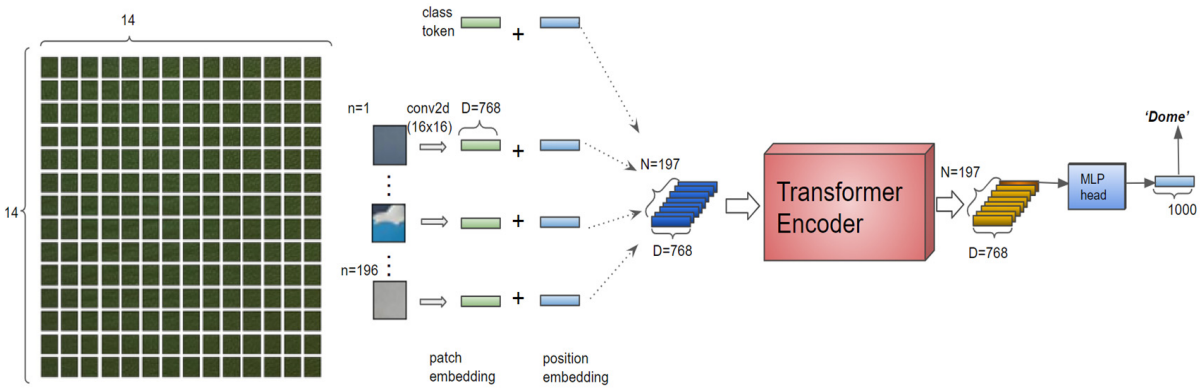


Figure 1. The vision transformer workflow [68].

3.2. CNNs and Autoencoders for Video Anomaly Detection

CNNs were designed to learn the spatial hierarchies of data automatically and adaptively. Furthermore, the convolutional layers apply filters to picture areas to preserve the spatial relationships between pixels and capture local characteristics. Due to their design, CNNs are superior in image classification and recognition tasks because they can learn progressively complicated patterns as input data flow through successive layers [69]. Although CNNs utilize filters on different picture regions, this naturally restricts the network's capacity to capture global relationships, even if it works incredibly well for identifying local characteristics. Thus, their capacity to comprehend the full context of an image is still restricted.

Using a nonlinear mapping function, a deep AE transforms the input data into their hidden low-dimensional representation. The objective is to train the AE to recreate the input patterns at the network's output. AEs cannot replicate anomalous data samples when trained on regular data samples. Therefore, deviations from the training model caused by abnormal activities lead to poor reconstruction (high reconstruction error).

4. Methodology

4.1. Model General Structure

In this work, a spatiotemporal attention encoder–decoder-based approach is explored. It comprises a model structure that can automatically perform video anomaly detection. The proposed structure is shown in Figure 2 where the encoder uses ViT-B_16 [27]. The multi-head self-attention (MSA) of the ViT is used to capture the relationships between image patches in the representation. The spatiotemporal attention mechanism is placed after the ViT to gather the spatiotemporal dependencies of the obtained features. The decoder consists of four CNN layers to allow fast reconstruction for the video frames with an adaptive pooling layer that provides an output reconstructed frame with the same size as the input.

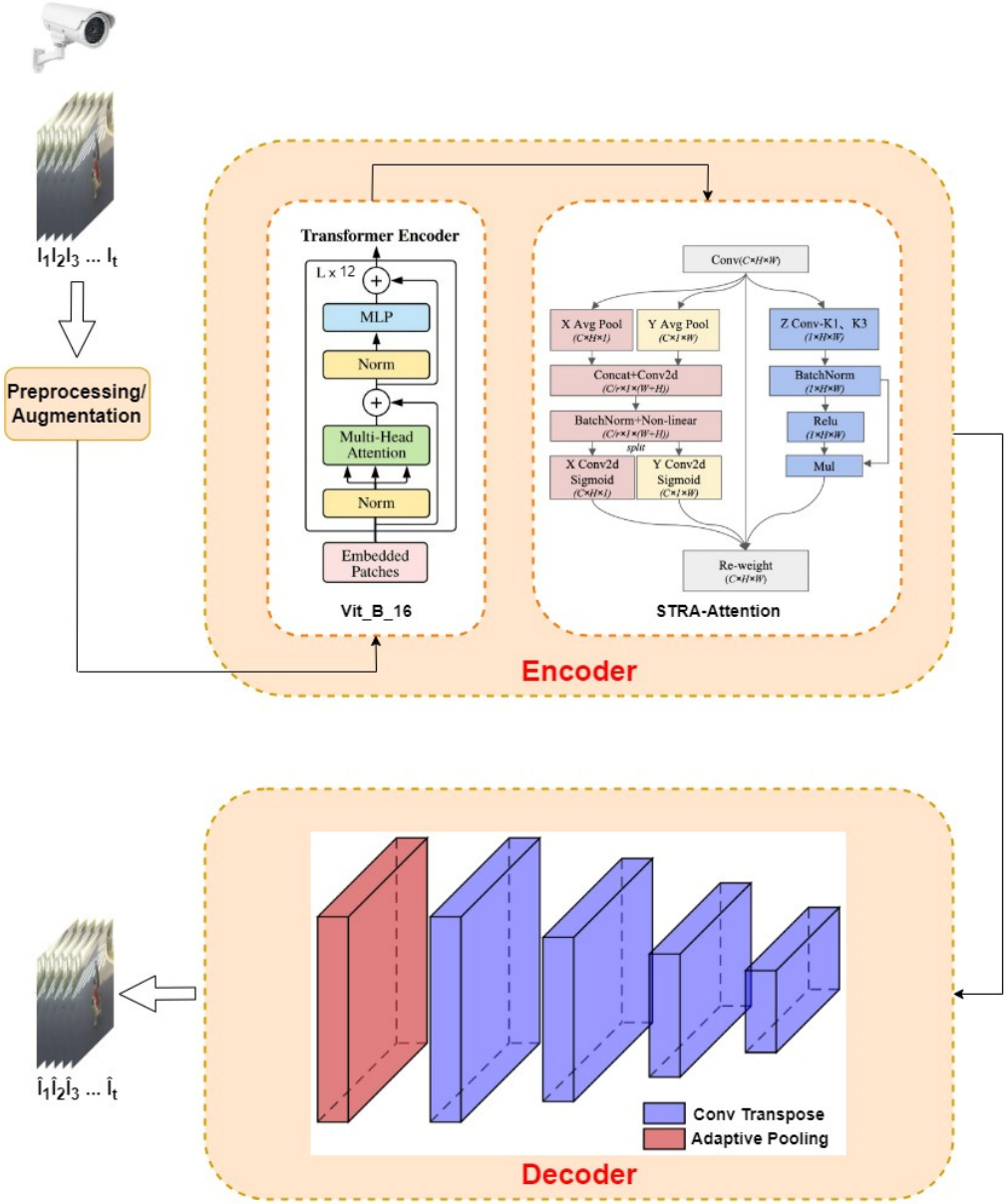


Figure 2. Architecture of the proposed spatiotemporal ViT-based model.

4.2. Model Operation

The proposed approach learns the normal data distribution and minimizes the reconstruction errors, which are the differences between the input images and the reconstructed ones. The model accepts the input frames sequence ($I_1 I_2 \ldots I_t$) and produces the recon-

structed output sequence ($\hat{I}_1\hat{I}_2 \ldots \hat{I}_t$), as depicted in Figure 2. An anomaly score is calculated at the output of each frame. The score is then compared to a threshold value. If the score is more than the threshold then the frame is anomalous, otherwise it is anomaly free.

4.2.1. ViT-Based Encoder–Decoder

As mentioned in Section 3.1, the ViT is useful for capturing the global context dependencies for images, unlike CNNs which focus on local characteristics. The model utilizes the ViT-B_16 pre-trained model in the encoder part with its classification head dropped, which was designed originally for image classification. This allows the model to be used as a feature extractor, as shown in Figure 3. The ViT-B_16 receives the input frame and divides it into 16 × 16 patches, which are subsequently linearly embedded into vectors. Then, positional encodings are employed to preserve the spatial information in the obtained vectors. In ViT-B_16, each patch is represented by a 768-dimensional vector. The ViT-B_16 model has twelve transformer encoder layers and each encoder has twelve attention heads. Thus, the input patches are processed through these multi-head self-attention mechanisms, followed by feed-forward networks that allow the ViT to focus on different parts of the input image simultaneously, capturing dependencies and relationships across the entire image. Finally, the multi-layer perceptron (MLP) head is utilized to aggregate the information from all patches. The obtained features are fed to the spatiotemporal attention model to capture the spatiotemporal relationships among objects in the video frames. Subsequently, the decoder part uses those features to reconstruct the input frames and minimize the reconstruction loss. The decoder consists of four CNN layers to allow fast reconstruction of the video frames with an adaptive pooling layer that provides an output reconstructed frame with the same size as the input.

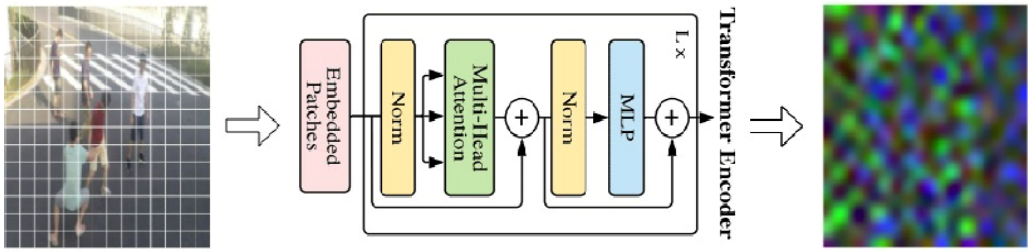


Figure 3. Example of features map obtained using ViT.

In the proposed framework, each video frame was encoded into a compact and informative representation using the ViT-B_16 model in an unsupervised manner for video anomaly detection. The network was trained on normal frames only to preserve their normal distribution, which enables the normal and anomaly samples to be distinguished. After that, the STRA model is applied to capture the spatiotemporal dependencies and relationships between objects among the input frames sequence to leverage the features of the normal frames. A CNN decoder then reconstructs the original video frames using the features obtained from STRA model. In this way, the transformer is capable of capturing global dependencies and intricate patterns within the frame, which is crucial for identifying anomalies that often appear as unusual patterns or objects in a video sequence.

The proposed model configuration becomes efficient at reconstructing normal frames during testing but finds it difficult to rebuild anomalous frames, as it does not learn to cope with such patterns. Hence, the reconstruction error can then be used as a signal to detect anomalies.

4.2.2. Spatiotemporal Relation Attention Model

Current attention-based techniques use global average pooling to help the model gather global information. However, the spatiotemporal connections among objects are more significant for video anomaly identification than the information in a layer's channels [28]. Accordingly, we used the STR attention block, as shown in Figure 4, to capture

the spatiotemporal interconnections among objects in the video frames. In contradistinction to the framework proposed by Wang et al. [28], the ViT transformer was adopted as the encoder part of the proposed model to capture the global context between the image patches instead of using CNN modules, which enable the model to obtain both global image context features and spatiotemporal relationships between objects to enhance the process of video anomaly detection. The STR attention block uses both temporal and spatial attention to identify complex correlations in video data. The spatial attention is attained by using convolutional layers and pooling techniques (with kernel sizes of H × 1 and 1 × W) to aggregate information along horizontal and vertical axes. The spatial attention scores are then generated by concatenating the obtained features and passing them through a shared convolutional layer to emphasize important spatial components. The temporal attention is performed through temporal pooling and convolutional processes to find relationships between frames and produce the temporal attention scores. Furthermore, a spatiotemporal attention map is created by combining these spatial and temporal scores, weighting the initial feature maps to increase attention to areas with important spatiotemporal correlations.

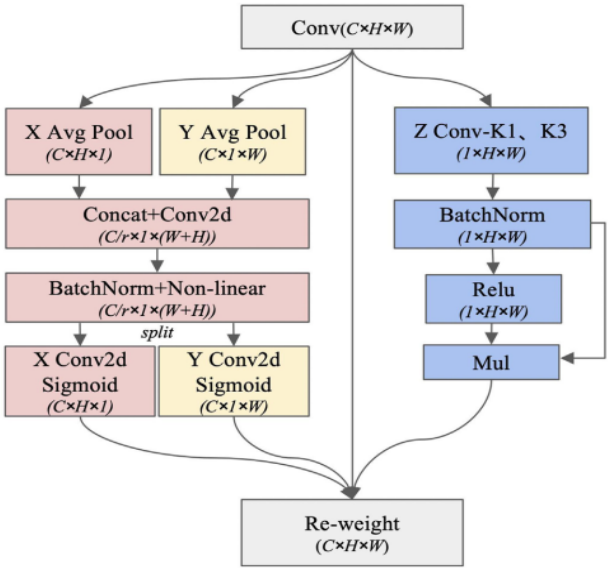


Figure 4. The spatiotemporal relationships attention (STR attention) module [28].

According to Wang et al. [28], the introduced attention method combines temporal and spatial information to obtain the spatiotemporal relationship information, which allows it to extract and aggregate dimensional components independently. The STR attention architecture consists of input processing, dimensional feature encoding, feature aggregation, an attention mechanism, and attention weighted output. The operation details are as follows:

A one-to-one dimensional feature encoding procedure uses two pooling kernels with spatial sizes of (H,1) and (W,1) to encode each channel along the horizontal and vertical axes, respectively, and encode the input $X \in \mathbb{R}^{(C \times H \times W)}$. The following is the output of the H dimension's *C*-th channel:

$$Y_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i) \tag{1}$$

where the features of the *H* dimension are represented by the input X, which typically comes from the fixed kernel size convolution layer. Additionally, the following formula is used to determine the output of the *W* dimension's *C*-th channel:

$$Y_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \tag{2}$$

Then, the features of the video frame sequence at the time t are derived from every channel feature in the following manner:

$$Y_t(X) = F^{3\times3}\sigma\left(F^{1\times1}(X)\right),\ X \in\ \mathbb{R}^{C/r\times(H+W)} \tag{3}$$

where the normalization process is represented by σ and the convolution operation is represented by F. The convolution kernel sizes are represented by 3×3 and 1×1, while the reduction ratio that regulates the change in the number of channels is denoted by r. A collection of spatiotemporal-aware feature mappings is produced by aggregating the outputs of the three transformations mentioned above along each of the three dimensions. The aggregation process is performed firstly by concatenating feature maps obtained from Equations (1) and (2) as follows:

$$Y^{hw} = \delta\left(F^{1\times1}\left[Y^h, Y^w\right]\right) \tag{4}$$

where [.,.] represents the concatenation of two tensors, $F^{1\times1}$ indicates the convolutional transformation through a convolutional kernel of size 1×1, and d represents the nonlinear activation function. Then, the output Y is split into two tensors with the same dimension, i.e., $Y^h \in \mathbb{R}^{C/r\times H}$ and $Y^w \in \mathbb{R}^{C/r\times W}$. After that, the spatial and temporal attentions are obtained by applying the activation function to each of the split tensors as follows:

$$g^h = \text{Sigmoid}\left(F_h\left(Y^h\right)\right) \tag{5}$$

$$g^w = \text{Sigmoid}\left(F_w\left(Y^w\right)\right) \tag{6}$$

where F_h and F_w indicate two different 1×1 convolutional kernels. Then, $Y_t(X)$ is obtained by the nonlinear activation function to represent the region feature weight tensor g^t that is sensitive to time change as follows:

$$g^t = \text{Relu}(Y_t(X)) \tag{7}$$

Finally, the attention weighted output of the spatiotemporal relationship attention module Y is as follows:

$$Y_c(i,j) = X_c(i,j) \times g_c^h(i) \times g_c^w(i) \times g_t^h(i,j) \tag{8}$$

Hence, the network can more precisely determine the spatiotemporal links among objects because the three transformations capture the long-range temporal dependencies and spatial relationships in the feature space. Accordingly, the introduced attention mechanism captured the region of interest by using spatial connections and temporal interdependence among objects. Moreover, the STR attention module could be plugged into the model without altering the input–output structure.

The mean squared error (MSE) is applied as a loss function by computing the average squared difference between the input and output frames to estimate how far the reconstructed frame is from the original. Because the model has only been trained for normalcy, the MSE values will be higher for anomalous frames and lower for regular ones. The goal is to reconstruct the normal video segments in a meaningful way. Thus, stable structures must be used to minimize the reconstructive loss between the ground truth Y and decoder output $\hat{Y}$. Furthermore, the obtained reconstruction error is used as an anomaly score to distinguish between the normal and anomaly frames. The reconstruction error could be calculated as follows:

$$L_{rec} = \left\|\hat{Y} - Y\right\|_2 \tag{9}$$

In addition, an experimental threshold is established, meaning that if a frame's anomaly score exceeds a threshold value, it is regarded as abnormal, and vice versa.

5. Experiments

5.1. Datasets

The proposed framework was evaluated using three challenging, publicly available anomaly datasets that are commonly used as benchmark datasets; the UCSD-Ped2, CUHK Avenue, and ShanghaiTech, and the extra-large CHAD anomaly dataset.

- UCSD-Pedestrian 2 (Ped2) [29] is considered one of the most popular datasets for unsupervised video abnormality detection. Researchers at the University of California, San Diego (UCSD) produced this dataset in 2010 to record usual pedestrian activity and rare anomalies in a controlled setting. The videos were taken from the same area at various times of the day, with varying lighting and shadow effects, which complicate the anomaly identification procedure. In this dataset, pedestrians crossed in front of the camera. Normal circumstances usually involve people strolling along walkways, but abnormalities are distinguished by the appearance of strange objects (like carts, bicycles, skateboards, etc.), strange motion patterns (like skating on a board), and people walking on grass. Ped2 consists of 16 videos with 2550 frames for training and 12 videos with 2010 frames for testing, each with a size of 240 by 360 pixels;
- The CUHK Avenue [30]: the Chinese University of Hong Kong (CUHK) produced the CUHK Avenue dataset in 2013, which provides a broader variety of anomalies in an open campus setting. The videos are captured from a single scene of the campus avenue. With 21 test films and 16 training videos, there are a total of 15,324 frames for testing and 15,328 frames for training in this dataset. The resolution of each frame is 360 by 640 pixels, and around 47 anomalous behaviors, such as loitering, throwing objects, and running through the gate, were noted. The CUHK Avenue dataset's training and testing clips are no longer than two minutes;
- ShanghaiTech [31]: the most extensive and available unsupervised dataset for video anomaly detection, consisting of 437 clips from 13 cameras positioned across the ShanghaiTech campus at a frame resolution of 856 $\times$ 480 pixels. It has 107 test videos with both normal and abnormal occurrences; there are a total of 130 abnormal events, and 330 training clips containing 274,515 frames with only normal events. Every scenario in the videos has a different cast of individuals, difficult lighting, and unusual camera angles. Among the human anomalies in the dataset are activities like skateboarding, riding motorcycles and bikes, chasing, fighting, and jogging. All videos have frame and pixel annotations;
- The CHAD [32] is a multi-camera anomaly dataset set in a commercial parking lot, which includes 420 videos with over 1.15 million high-resolution frames from four camera views. Scenes 1–3 and Scene 4 are captured at 30 frames per second, and at 1920 $\times$ 1080 resolution and 1280 $\times$ 720 resolution, respectively. According to its authors, it is the largest fully annotated anomaly detection dataset, offering detailed person annotations from continuous video streams captured by stationary cameras. The CHAD provides human detection, tracking, and pose annotations encompassing four types: frame-level anomaly labels, person bounding boxes, person ID labels, and human key points. It includes 59,172 anomalous frames representing 22 distinct behaviors categorized into group and individual activities, and 1,093,477 normal frames. The group activities comprise fighting, punching, pushing, pulling, slapping, strangling, theft, etc., while the individual activities include throwing, running, riding, falling, littering, etc.

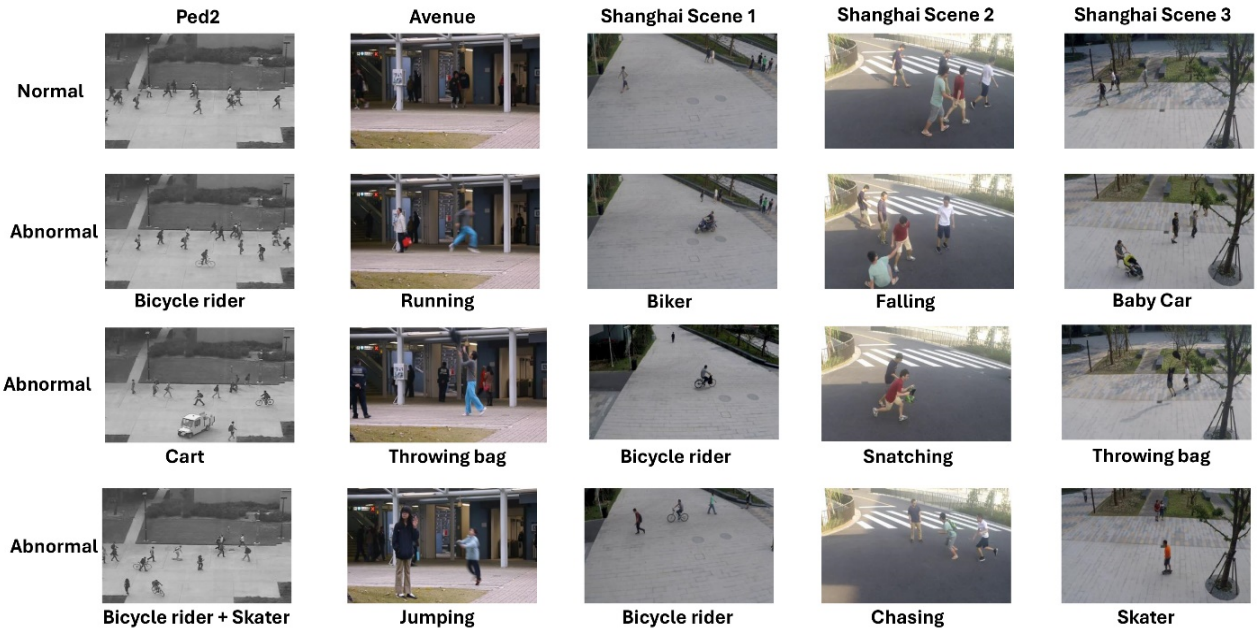
The summary for the four video anomaly datasets is shown in Table 1. Different samples from the datasets displayed in Figure 5 vary from normal to anomaly, with short descriptions for anomaly samples.

To add another complexity level to the model to prove its effectiveness in detecting abnormalities in videos, we opted to make new mixed datasets by combining the utilized datasets in permutations, as shown in Table 2. This mechanism will allow the model to be

trained on heterogeneous and very large-sized data with extremely diverse environmental conditions and data qualities.

Table 1. Summary of video anomaly datasets.

Dataset	#Training Videos	#Training Frames	#Test Videos	#Test Frames	Anomaly Types
UCSD-Ped2	16	2550	12	2010	• Carts; • Bicycles; • Skateboards, etc.
CUHK Avenue	16	15,328	21	15,324	• Loitering; • Throwing; • Running; • Jumping, etc.
ShanghaiTech	330	274,515	107	40,791	• Skateboarding; • Motorcycles; • Bikes; • Chasing; • Fighting; • Robbing; • Jogging, etc.
CHAD	278	1,026,174	134	126,475	Group activities • Fighting; • Punching; • Pulling; • Slapping; • Strangling; • Theft; • Pick-pocketing; • Chasing, etc. Individual activities • Throwing; • Running; • Riding; • Falling; • Jumping, etc.



(**a**)

Figure 5. *Cont.*

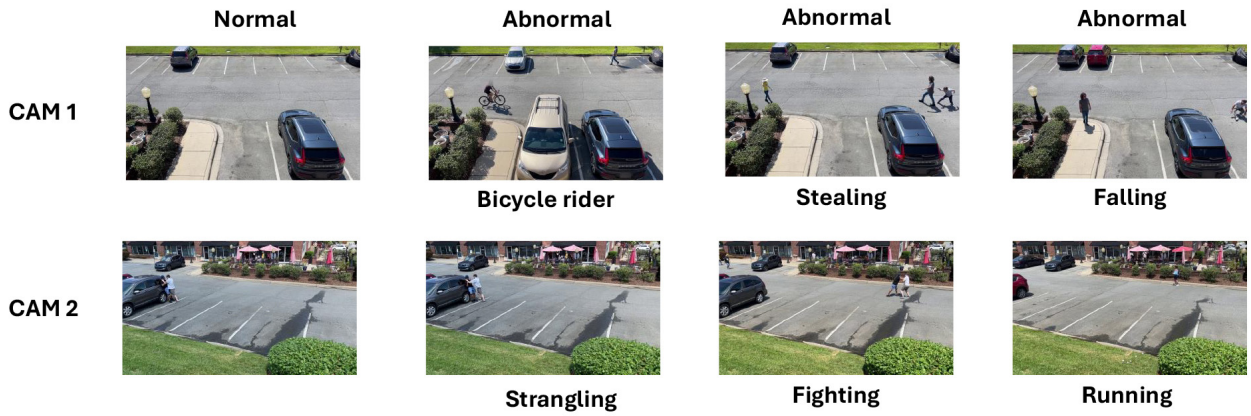


Figure 5. Different anomaly samples for (**a**) UCSD, Avenue, and Shanghai datasets. (**b**) Cam 1 and Cam 2 from the CHAD dataset.

Table 2. Size of the combined video anomaly datasets.

Dataset	#Training Frames
UCSD + Avenue	17,878
UCSD + Shanghai	277,065
Avenue + Shanghai	289,843
UCSD + Avenue + Shanghai	292,393

5.2. Model Implementation and Assessment Metrics

The proposed framework was implemented using Python 3 and Pytorch backend on a machine equipped with an NVIDIA GeForce RTX 3080 GPU and 16 GB of memory for the experiments. The input video frame was resized to $224 \times 224 \times 3$ dimensions to capture its characteristics from the ViT. The Adam optimizer [70] was used to train and optimize the model with a learning rate of $1e^{-4}$, batch size 16, and a five-frame input sequence length. The hidden dimension in MLP in ViT was set to 128, 512, 1024 and 1024, for the UCSD, Avenue, Shanghai, and CHAD, respectively.

The performance of the model was assessed using the area under the receiver operating characteristic (ROC) curve, following the other research such as [26,49,56,64,71,72]. The AUC ROC was employed to determine the capability of the introduced method to discriminate between normal and abnormal frames. Accordingly, the existence of abnormalities in the recordings at the frame level was assessed following the assessment methodologies [73–75]. The true positive rate (TPR) against the false positive rate (FPR) curve constitutes the ROC curve. The TPR and FPR are computed using Equations (10) and (11), respectively.

$$TPR = \frac{TP}{TP + FN} \tag{10}$$

$$FPR = \frac{FP}{FP + TN} \tag{11}$$

The abbreviations *TP* and *FN* stand for true positive and false negative, respectively, whereas *FP* and *TN* are abbreviations for false positive and true negative.

5.3. Preprocessing

Due to the remarkable difference in sizes among the four utilized datasets, as shown in Table 1, two different up-sampling approaches were applied as preprocessing steps to the UCSD and Avenue datasets to enlarge their sizes and study the effect of augmentation on the learning process of the proposed model. To add another level of complexity, the three datasets, UCSD-Ped2, CUHK Avenue, and ShanghaiTech, were merged interchange-

ably, and the model was trained using the merged versions to prove the effectiveness and robustness of the proposed model for large datasets. The up-sampling step was conducted using two different approaches.

- Approach (A) aims to increase the size of the dataset by five times, where each original frame is boosted by five augmented frames. The applied transformation operations were flip and crop. In the flip operation, the image is randomly flipped horizontally with a 50% chance, while in crop the image is randomly cropped to 224 × 224 pixels;
- Approach (B) involves applying a series of transformations, both individually and in dual combinations, to each frame to obtain a 1:10 original–augmented frame ratio. These transformations include random cropping to a size of 224 × 224 pixels, random horizontal flipping, random rotation within a range of 0 to 60 degrees, and random affine transformations that include slight translations (up to a 10% shift in x and y directions), scaling (between 90% and 110% of the original size), and shearing (distortion) by up to 10 degrees.

These approaches significantly increase the diversity of the training dataset by introducing various geometric changes in perspective that generally improve the robustness and generalization ability of the training models. Table 3 shows the number of frames in each dataset after applying different up-sampling approaches.

Table 3. The augmented dataset sizes.

Augmentation Approach	#Frames UCSD	#Frames Avenue
Approach (A)	15,300/2550	91,968/2550
Approach (B)	28,050/15,328	168,608/15,328

5.4. Results and Discussion

This section gives a thorough investigation about our experimental evaluation of the model using a variety of large-scale anomaly detection datasets. Different experiments based on the encoder–decoder architecture were launched, where the pre-trained ViT model is considered an encoder with the decoder constituting the "baseline" architecture. Then, the STR attention block was involved after the ViT to show the usefulness of using the spatiotemporal characteristics for video anomaly detection.

The proposed model architecture for anomaly identification produced outstanding results, as shown in Table 4, which demonstrates the effectiveness of using the STR attention model for improving anomaly detection accuracy. Table 4 compares the results of using the STR attention model to the results of using the framework without it. Accordingly, the obtained results show that the proposed architecture can successfully detect anomalies with high detection rates by utilizing the local characteristics and global dependencies (context) of the video frames captured by the proposed model for the video anomaly detection process. Moreover, the detection accuracy results depicting the efficacy of the proposed model are shown in Figure 6. The figure shows the ROC results for different test videos from the four datasets, which demonstrate the effectiveness of using the ViT transformer coupled with the spatiotemporal attention model to enhance the detection accuracy of normal/abnormal frames for surveillance videos.

Table 4. The effectiveness of the model for the different datasets.

Model\Dataset	UCSD%	Avenue%	Shanghai%
Baseline	94.5	84.9	77.3
Baseline + STR attention	95.6	85.2	81.9

The effect of using augmentation and combined datasets is presented in Table 5, which shows that the augmentation techniques improved the ViT + STR attention model's anomaly detection accuracy for the Avenue datasets when strategy 'B' was applied. This

demonstrated the usefulness of the model in identifying anomalies when trained on large datasets. Furthermore, merging the datasets improves the individual and overall detection rate when the STR attention model is applied, even though each dataset contains movies of varying qualities and a range of shooting scenarios, colors, and angles. Additionally, by combining the larger dataset with other datasets, as shown in Table 6, the accuracy of anomaly identification significantly improves. The combination strategy is considered a leading idea for developing automated real-time video surveillance anomaly detection systems. It is well known that those surveillance systems have heterogeneous captured videos with different environmental circumstances and various anomaly types.

Table 5. AUC ROCs (%) for different up-sampling approaches for the UCSD and Avenue datasets.

Approach / Dataset	Approach A		Approach B	
	Baseline	**Baseline + STRA**	**Baseline**	**Baseline + STRA**
UCSD	92.4	92.8	91.0	90.1
Avenue	84.5	84.6	84.8	**86.8**

Table 6. AUC ROCs (%) for different combinations for the UCSD, Avenue, and Shanghai datasets.

Model Trained on	Model Tested on	Baseline	Baseline + STR Attention
Avenue + UCSD	UCSD	94.6	94.5
	Avenue	84.8	85.0
	Avenue + UCSD	87.5	87.7
Shanghai + UCSD	UCSD	89.3	89.9
	Shanghai	80.2	**82.1**
	Shanghai + UCSD	80.9	82.5
Shanghai + Avenue	Avenue	86.6	86.1
	Shanghai	79.1	80.6
	Shanghai + Avenue	80.3	81.5
Shanghai + Avenue + UCSD	UCSD	90.2	92.3
	Avenue	85.8	**86.8**
	Shanghai	79.4	80.3
	Shanghai + Avenue + UCSD	81.4	81.8

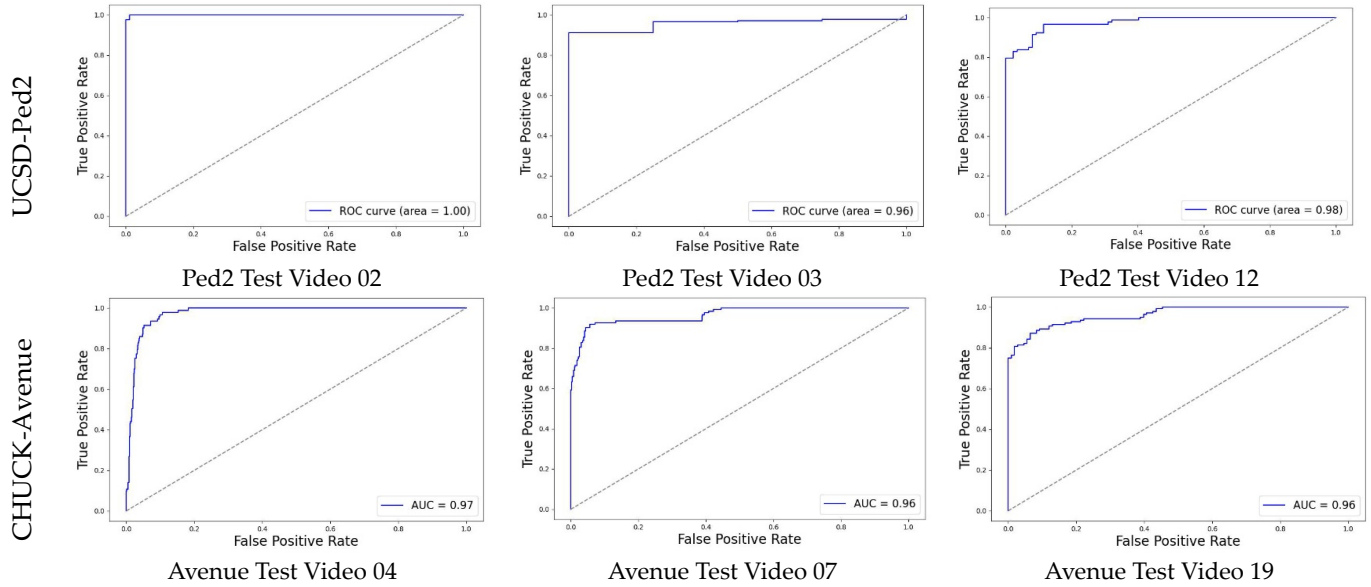


Figure 6. *Cont.*

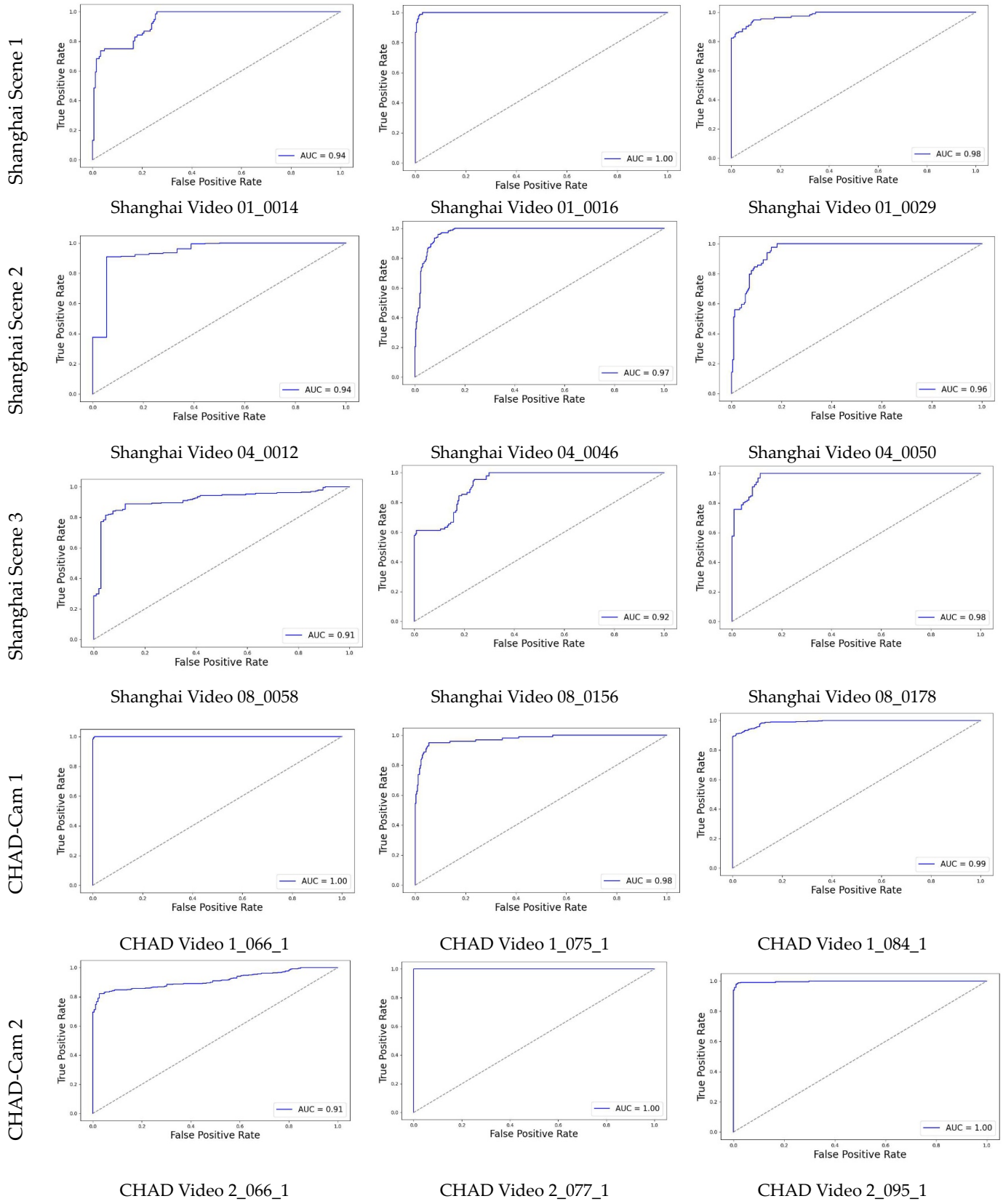


Figure 6. Different AUC ROCs for the four datasets.

Based on the results obtained from the ShanghaiTech dataset, which proves the ability of the proposed model, ViT + STRA, to detect anomalies in large datasets and to leverage

our findings about the efficiency of the proposed model when trained on larger datasets, we trained the model on the CHAD dataset. Precisely, we utilized a subset of the CHAD dataset of videos captured by Cam 1 and Cam 2 to train and evaluate the proposed model. The new subset comprises 492,671 normal frames for training and 62,879 frames for testing, totaling 555,550 frames, which is much bigger than the ShanghaiTech and considered an extra-large dataset. The model achieved an impressive average ROC of 68% for Cam 1 and Cam 2 combined as a test set.

5.5. Comparative Methods

As demonstrated in Table 7, the proposed method achieved comparable performance for both the UCSD and Avenue datasets compared to the state-of-the-art (SOTA) approaches. Notably, the two datasets are characterized by small and medium size, respectively, and the model was designed to deal with the large datasets. For the Shanghai University of Science and Technology dataset, there are thirteen distinct complicated contexts and situations, including dense crowds, varied movement patterns, and unusual occurrences. Nevertheless, the proposed model showed a 7.3% increase in detection performance over existing VAD methods on this dataset, which is considered the large-scale benchmark dataset currently available in the video anomaly detection domain. The proposed method's power is demonstrated by its comparison with other recent SOTA methods, such as the spatiotemporal convolutional autoencoder model introduced by Kommanduri and Ghorai [76] and the transformer memory autoencoder approach used by Wang et al. [77], which yielded detection superiorities of 8.4% and 9.6% for the Shanghai dataset, respectively. In addition, it achieved a superior result compared to the recent work in [66] which employed an attention U-Net based on multi-scale feature extraction with a data augmentation network, achieving +0.6% for the Avenue dataset, but the authors did not use the Shanghai dataset in their experiments. As a result, the proposed model made progress towards finding a high-detection performance solution to the problem of finding abnormalities in large, diverse, and heterogeneous anomaly datasets. Therefore, this research may contribute to the development of reliable real-time video anomaly detection systems.

Table 7. Frame-level AUC performance comparison of the proposed framework against state-of-the-art anomaly detection methods for the UCSD, Avenue, and Shanghai datasets.

Method		Model	UCSD-Ped2 (%)	Avenue (%)	ShanghaiTech (%)
Non-Recon.	[78]	Multivariate Gaussian Fully Convolution Adversarial Autoencoder (MGFC-AAE)	91.6	84.2	-
	[79]	FSM-GAN	98.1	80.1	73.5
	[80]	Two U-Net generators + discriminator	96.3	85.1	73.0
	[19]	Memory-guided AE	97.0	88.5	70.5
	[81]	Spatiotemporal consistency enhanced network (STCEN)	96.9	86.6	73.8
	[75]	Attention residual AE	97.4	86.7	73.6
Recon.	[82]	Memory-augmented AE	94.1	83.3	71.2
	[83]	AE + Probability distribution density estimator	95.4	-	72.5
	[84]	Gaussian Mixture Fully Convolutional Variational Autoencoder (GMFC-VAE)	92.2	83.4	-
	[85]	Two-stream clustering Autoencoder	96.5	86.0	73.3
	[86]	Residual autoencoder	83.0	82.0	-

Table 7. *Cont.*

Method		Model	UCSD-Ped2 (%)	Avenue (%)	ShanghaiTech (%)
Recon.	[87]	Two-stream autoencoder	84.5	80.3	-
	[58]	Attention-based ConvLSTM network þ Conv Autoencoder	95.6	-	73.1
	[71]	AE + DPU + Attention module	97.9	85.9	-
	[88]	Implicit two-path AE + normalizing flow	97.3	85.8	74.7
	[56]	Noise AE + AE + Second-order channel attention	97.9	86.2	-
	[21]	Residual AE	95.4	80.9	-
	[76]	ConvLSTM-AE	97.9	**89.8**	73.7
	[77]	Transformer memory + AE	98.1	88.5	72.5
	[66]	Multi-scale feature attention U-Net + weakly supervised data augmentation	97.9	86.2	-
	Ours	ViT + STR attention	95.6	86.8	**82.1**

Table 8 shows a comparison of the introduced framework against the state of the art. There are few works that employed the CHAD for video anomaly detection because the CHAD is a recent dataset, for example, Yao et al. [89] evaluated some pose-based techniques on this dataset. From Table 8, it is obvious that our ViT + STRA framework outperforms these techniques, taking into consideration that the TSGAD [90], GEPC [91], and STG-NF [92] techniques were trained on Cam 1 or Cam 2 individually. On the other hand, our model was trained on a combined set of Cam 1 and Cam 2, which makes it more complex and presents more challenges for the proposed model to detect anomalies.

Table 8. A comparison of AUC ROCs (%) for different models trained on the CHAD dataset.

Model	Cam 1	Cam 2
TSGAD [90]	63.2	60.1
GEPC [91]	63.0	63.0
STG-NF [92]	61.5	56.2
Proposed [1]	71.8	64.2

[1] Our model was trained on a combined set of Cam 1 and Cam 2.

Based on Tables 4–8, we could conclude that: (1) the proposed model obtains a higher AUC of 82.1% on the complicated ShanghaiTech; (2) on the UCSD, Avenue, and ShanghaiTech databases, it produces an AUC performance boost over the baseline of 1.1%, 0.3%, and 4.6%, respectively; (3) the combination of the various training datasets demonstrates the efficacy of the proposed method for learning and extracting spatiotemporal characteristics and its robustness for identifying anomalies under different conditions and scenarios for very large datasets; (4) the recent video anomaly CHAD dataset achieved the highest average AUC of 68% on Cam 1 and Cam 2, with a superiority of +8.6% and +1.2% for Cam 1 and Cam 2, respectively, over SOTA, and (5) the model demonstrated an outstanding ability to identify abnormal events and objects in videos compared to SOTA.

5.6. Case Studies with Visualizations

To gain further insight into the effectiveness of the proposed model, Figures 7–11 display the anomaly scores for different videos from the three datasets with some key normal and abnormal frames, with the anomaly ground truth shaded in pink. Two dif-

ferent anomaly scores from each of the Ped2 and Avenue datasets are displayed, as the two datasets represent single-scene videos (taken from a fixed position camera from a fixed view angle). For the ShanghaiTech dataset, there are thirteen distinct scenarios; hence, we display three different scenes with two situations for each. In addition, we show two different videos for each Cam 1 scene and Cam 2 scene from the CHAD dataset. The proposed approach responds successfully to anomaly event occurrences that are shown by the rising anomaly scores during the anomaly periods and the falling anomaly scores for the regular situations.

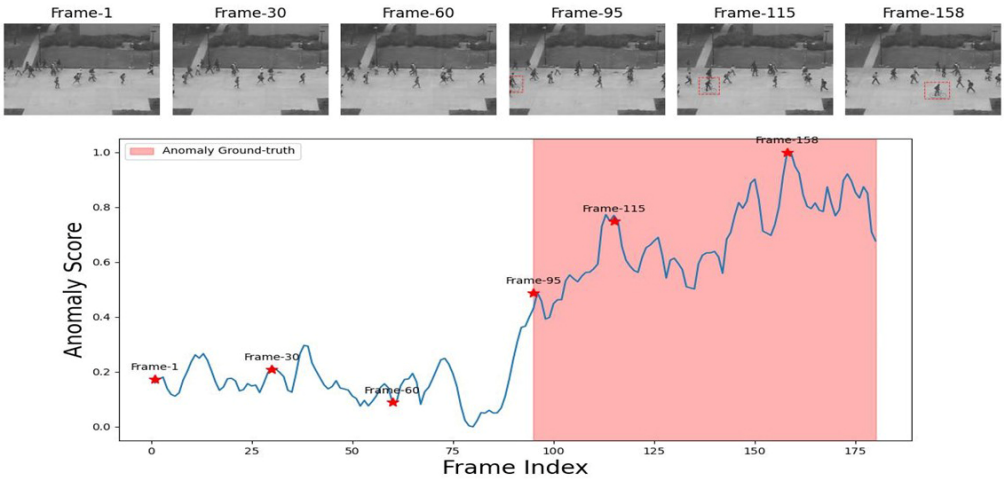


Figure 7. Anomaly score curves of two different videos for UCSD-Ped2 dataset: (**a**) Test 02 (anomaly type: cyclist); (**b**) Test 04 (anomaly type: cart and cyclist).

For instance, Figure 7 depicts the anomaly scores for test videos 02 and 04 from the UCSD-Ped dataset. When the cyclist's anomalous event, shaded in pink, happened in Figure 7a, the score increased and remained high until it reached the end of the anomaly period. Furthermore, Figure 7b shows how the anomaly score rises as soon as the anomaly object, the cart, enters the picture and stays high for the duration that the anomaly objects, the cart and the cyclist, remain in the scene. Moreover, Figure 8 displays the anomaly scores for test clips 05 and 06 from the Avenue dataset. The anomaly score in Figure 8a indicates that the anomalous event "playing with the bag" is correctly identified. The anomalous events in test video 06 (throwing bag and wrong direction) may be seen in Figure 8b during four periods, three of which are very short and represent a guy throwing a bag. Despite this, the proposed model was able to identify the anomalous times.

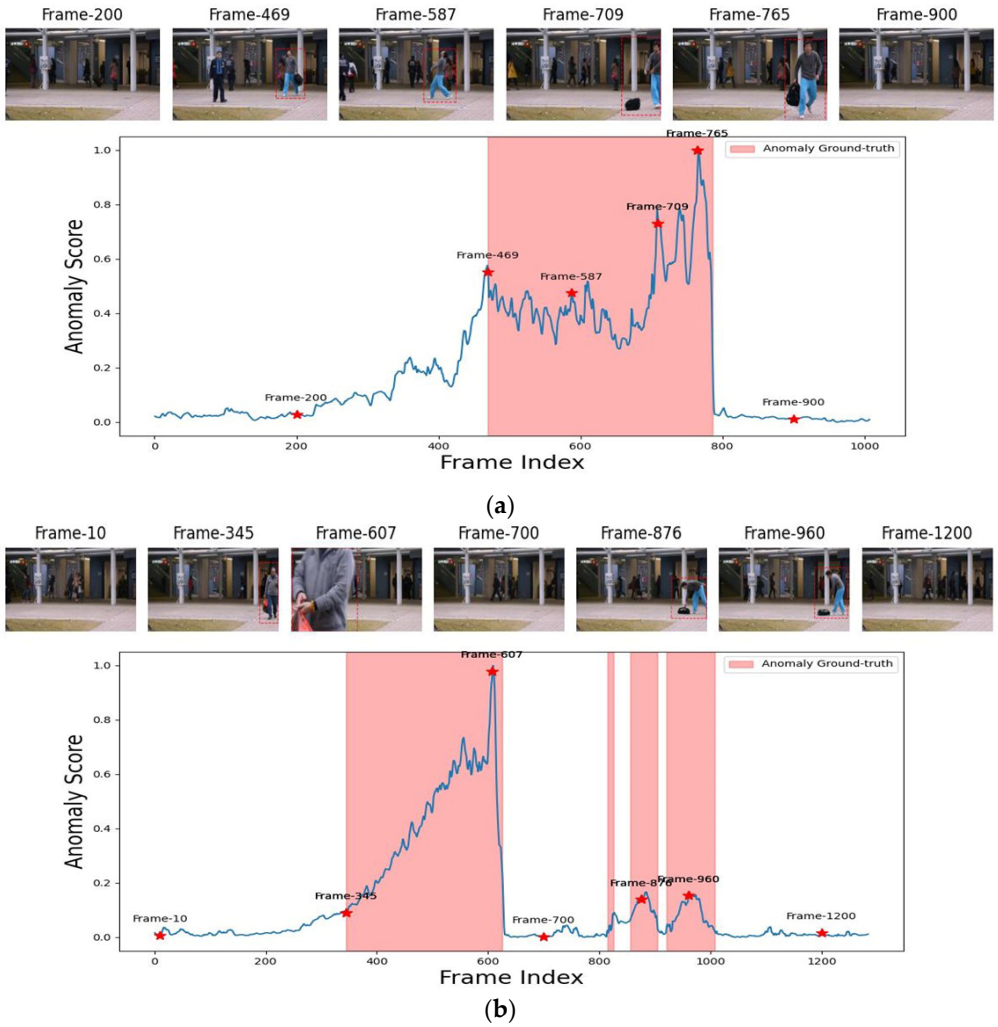
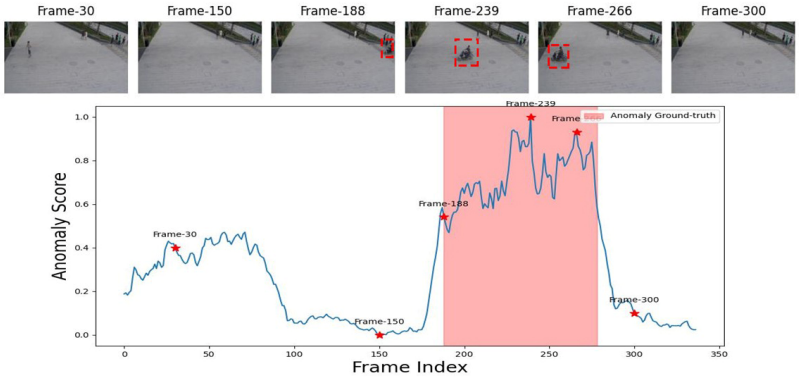
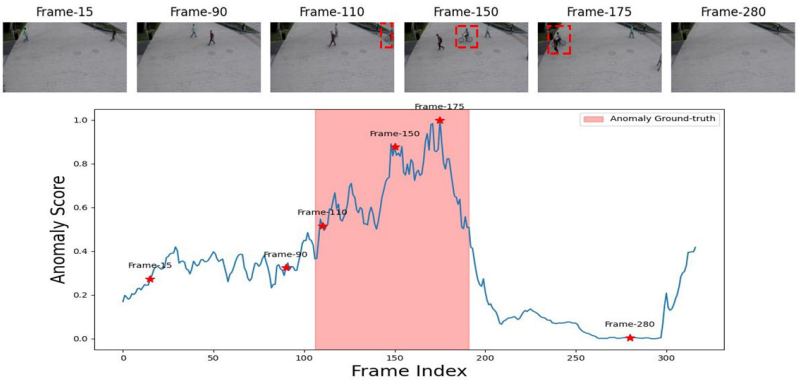


Figure 8. Anomaly score curves of two different videos from the CHUCK Avenue dataset: (**a**) Test 05 (anomaly type: throwing bag); (**b**) Test 06 (anomaly type: wrong direction and throwing bag).

Figures 9–11 represent the anomaly scores for test videos from three distinct scenes from the challenging ShanghaiTech dataset. The anomaly scores in the figures demonstrate how the anomaly events/periods in every video were identified successfully with high scores. The presence of the motorcyclist in the scene raised the abnormality score in test video 0016 from scene 01, as shown in Figure 9a, which is the same behavior that occurred in Figure 9b with the biker's existence in test video 0177 from the same scene. For scene 03, two test video anomaly scores, 0031 and 0032, are shown in Figure 10. The first shows a guy who hijacks the bag from his colleague and then starts chasing. In the latter, the anomalous event was represented by a person falling to the ground. As seen in Figure 10a, the model was able to correctly identify the complex abnormal situation with a high anomaly score, in the same manner as for the anomalous period in Figure 10b. In Figure 11, which exhibits the anomaly scores with shaded ground truth for videos 0144 and 0147 from the Shanghai scene 06, the model optimally recognized the unusually moving car on the sidewalk with extremely high scores, as shown in Figure 11a. On the other hand, Figure 11b depicts a cyclist with an umbrella and shows that the anomaly score was extremely high when this anomalous event occurred. As a result, the above figures show that the model performed exceptionally well in detecting distinct anomalous events and objects in the three different datasets. It demonstrated its efficacy in differentiating between normal and abnormal events in videos by producing high anomaly scores for anomalous intervals and low scores for normal ones.

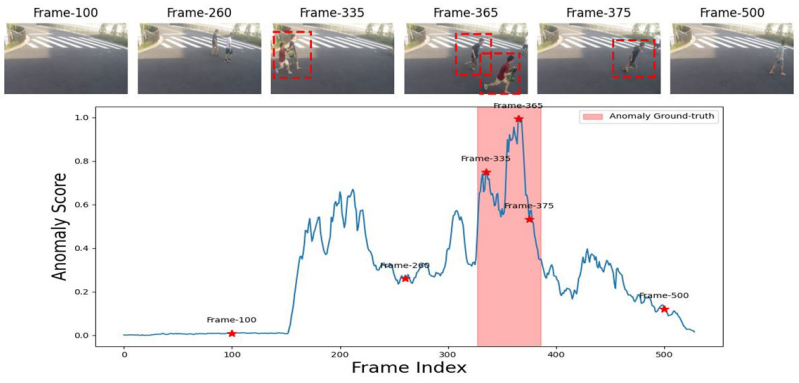


(**a**)

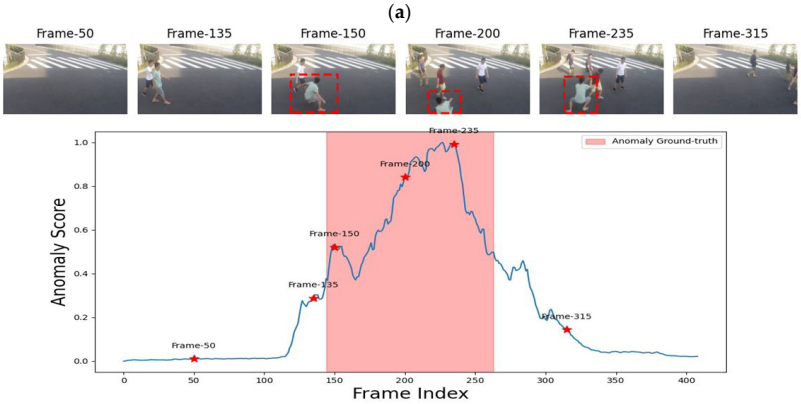


(**b**)

Figure 9. Anomaly score curves of two different videos for Scene 01 of Shanghai dataset: (**a**) Test01_0016 (anomaly type: motorcyclist); (**b**) Test01_0177 (anomaly type: biker).

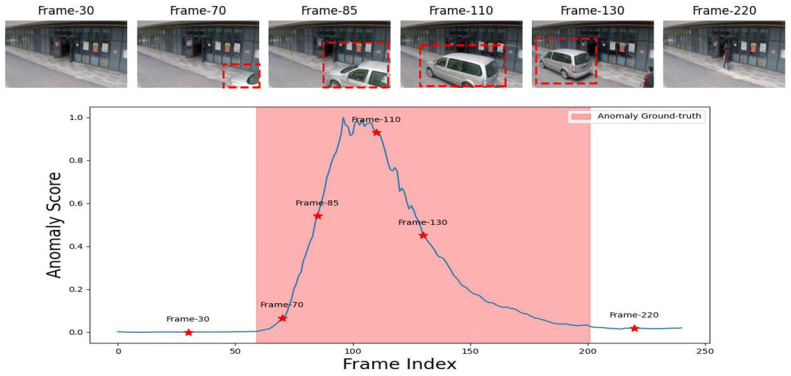


(**a**)

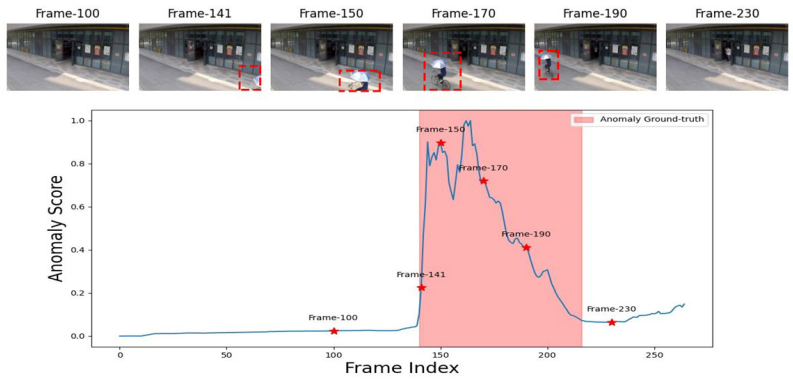


(**b**)

Figure 10. Anomaly score curves of two different videos for Scene 03 of Shanghai dataset: (**a**) Test03_0031 (anomaly type: robbing and chasing); (**b**) Test03_0032 (anomaly type: falling).



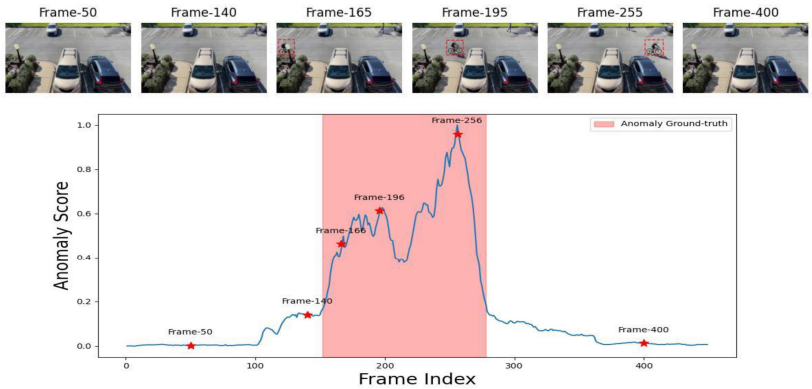
(**a**)



(**b**)

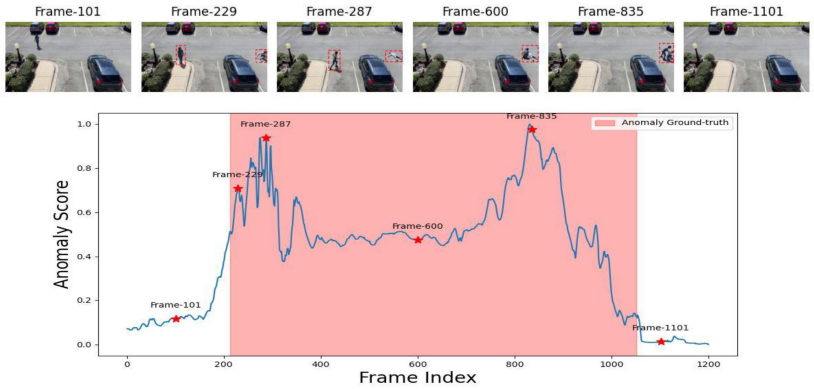
Figure 11. Anomaly score curves of two different videos for Scene 06 of Shanghai dataset: (**a**) Test06_0144 (car in pedestrian walkway); (**b**) Test06_0147 (biker with the umbrella in pedestrian walkway).

The anomaly scores for four test videos with two scenarios from the CHAD dataset are displayed in Figures 12 and 13. The figures depict that the proposed model was able to identify different abnormal events with high score values. The abnormal event of "biking" from the Test 1_066_1 video of the CHAD-Cam 1 scenario is represented in Figure 12a, which shows that the abnormal event was successfully recognized. The "falling" and "running" events from the Test 1_084_1 movie of the same situation, as seen in Figure 12b, were also recognized with strong anomaly scores in the same way. For Test 2_077_1 from CHAD-Cam 2, two people appeared to fight, "one is strangling the other", then they chased, and the model was able to detect the whole complicated event, as shown in Figure 13a. A runner who appeared in Figure 13b, from the Test 2_095_1 video, was identified correctly with high anomaly scores, even though it appears as a small object, far from the camera position, and intertwined with objects in the scene background.



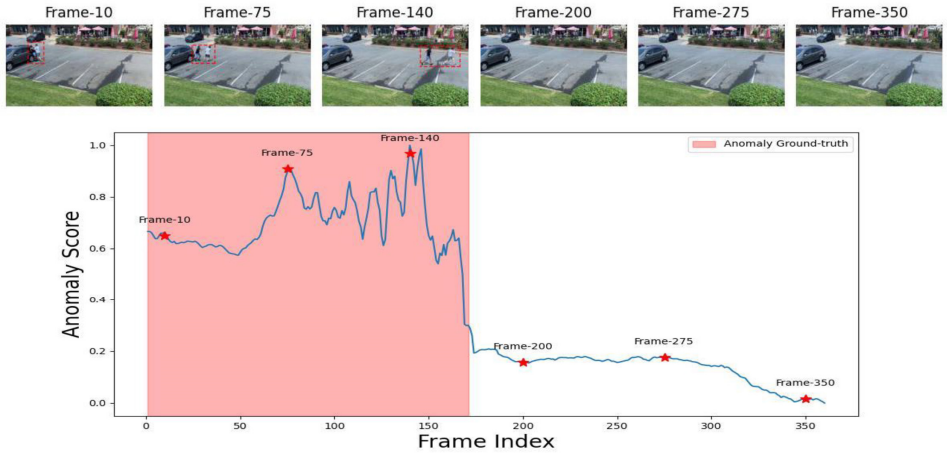
(**a**)

Figure 12. *Cont.*

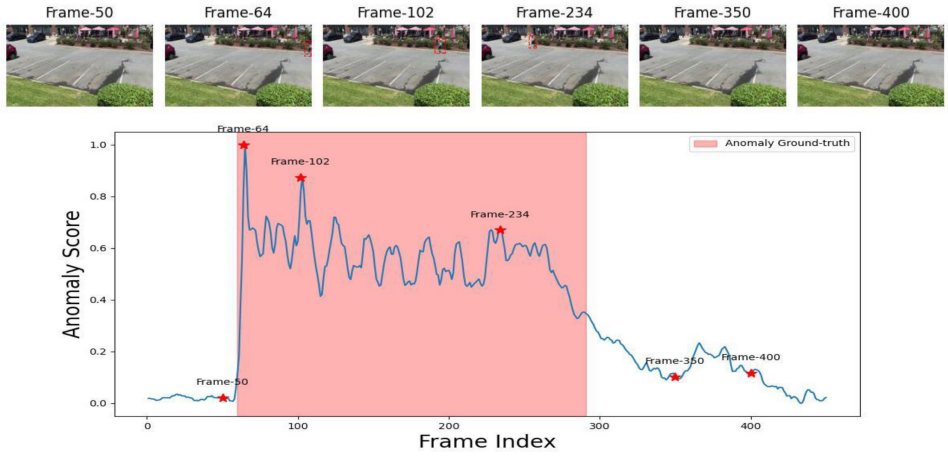


(**b**)

Figure 12. Anomaly score curves of two different videos for Cam 1 of CHAD dataset: (**a**) Test 1_066_1 (anomaly type: biker); (**b**) Test 1_084_1 (anomaly type: falling).



(**a**)



(**b**)

Figure 13. Anomaly score curves of two different videos for Cam 2 of CHAD dataset: (**a**) Test 2_077_1 (anomaly type: strangling and chasing); (**b**) Test 2_095_1 (anomaly type: running).

On the other hand, it was found that some ShanghaiTech test videos contain badly annotated frames (minor falsely annotated frames), such as frames No. 151, 157, and 162 in video 06_0144, frame No. 184 in video 06_0147, and frame No. 72 in video 06_0150. In addition to frame No. 161 in video 03_0033 and frame No. 202 in video 02_0164, falsely annotated frames were excluded from the test. Accordingly, these videos need to be revised by the dataset authors.

It is worth noting that the introduced approach does not perform optimally on all three datasets for the following reasons: the three datasets differ significantly in terms of shooting situations, colors, and perspectives, making it challenging for a system to produce comparable results across all datasets. Moreover, ViTs do not have CNNs' inherent inductive biases, and they frequently need a massive amount of data to be trained efficiently. However, they can outperform CNNs when they are trained, particularly on large-scale datasets. The proposed ViT + STRA-attention model proved this point when mixed combinations of the datasets were used as the detection accuracy improved despite the diverse natures of the three datasets, which include various illumination and environmental conditions, various view angles, different scenes with different quality and resolutions, and the large number and diversity of anomalous events and objects in the datasets. Moreover, CNNs use convolutional layers, assuming the spatial hierarchies and localizations in images. This strong inductive bias allows CNNs to learn well from very limited datasets by using the intrinsic structure of visual data. In contrast, ViTs use self-attention processes to identify associations between patches in pictures by treating them as a succession of patches. Because it makes no assumptions about the intrinsic order or locality among the patches, this method has less inductive bias concerning the spatial structure of images. This gives ViTs great flexibility and not only enables them to extract intricate relationships, but also implies that for them to learn these patterns efficiently, more data are needed. This reduced inductive bias might cause underfitting for the smaller training datasets, whereby the ViT might not pick up enough information about the relevant characteristics of the pictures to operate effectively on unobserved data. In addition, we believe that the proposed model is more sophisticated and thus able to perform on small datasets such as the UCSD-Ped2, since it consists of low-quality grey frames, however, it achieved promising results of 95.6%. On the other hand, a better performance was obtained with the much larger Avenue dataset, making it superior to the Shanghai dataset which is considered the largest dataset available for VAD. Additionally, VIT + STR attention enhances the detection of larger datasets (i.e., Shanghai and Avenue) compared to the baseline model for the different combinations of the datasets, as shown in Table 6. We call the combined versions 'very large datasets'. In addition, it outperforms the SOTA techniques for the Cam 1 and Cam 2 data from the largest CHAD dataset with AUC scores of 71.8% and 64.2%, respectively. This shows the ability of the proposed model to extract robust, effective, and complex features and successfully identify anomalies with high scores from extremely diverse videos captured from different environments, with different qualities, and in different conditions that contain a wide range of various anomaly events and types.

However, we think that by taking into account the following issues and constraints, our framework might be improved in the future:

1. We could benefit from the recent memory-augmented neural networks (MANNs) to improve the suggested approach. MANNs are networks that preserve a larger collection of representations throughout time by selectively storing and updating only the relevant information. Hence, these networks may be augmented in our model to improve its capacity to recognize more subtle and complicated patterns. Memory networks have efficient scaling abilities and adaptability to various scenarios and circumstances. Further, this could lead to high-accuracy anomaly detection for larger datasets, which is our focus in this research, and enhance anomaly detection in videos for real-time systems. As in real-world applications, the characteristics of anomalies could vary over time; hence, scaling capacity and adaptability are essential. Consequently, normal and abnormal behaviors may be distinguished more precisely;
2. It is important to note that the thorough results of each threshold are the basis for the proposed model's performance evaluation. Nevertheless, only the optimal threshold may be chosen for real-world use. A dynamic threshold-based anomaly detection technique was presented by Jia et al. [93], which offers an alternative viewpoint on how to improve our approach;

3. Since not every frame in a video is worth detecting, frame summarization may also improve. More video data can be analyzed simultaneously and processing power can be conserved if the frames that are more likely to be abnormal can be identified. The literature [94,95] motivated our investigation into future work.

6. Conclusions

The proposed framework leverages a novel integration of vision transformers with the spatiotemporal relationship attention technique to address the challenges of video anomaly detection in surveillance systems. In addition, this work applied a novel training strategy by combining different benchmark datasets to prove its robustness in dealing with large and distinct data. By utilizing the advantages of ViTs for capturing global context and the dynamic processing of STR attention for analyzing motions and interactions within frames, the unsupervised method produces a high-performance model suitable for large-scale and heterogeneous environments. This improvement illustrates the appropriateness of our technique for real-world scenarios, where diverse environmental conditions and data qualities often challenge existing systems. Our experiments show that our model not only performs superiorly to state-of-the-art anomaly detection techniques, but it also effectively adapts to the intricate variability of surveillance video data. Future research will examine how to scale this approach and improve the feature extraction capabilities using different ViTs, if we access more powerful resources, to handle even more complex and varied datasets, such as the whole 1.15 million frame CHAD dataset. Moreover, the memory mechanism could be incorporated to increase the capacity of the model for feature acquisition. Additionally, further adjustments and optimizations could improve the framework's application, potentially setting a new standard for automated surveillance systems and making them more reliable, efficient, and capable of handling the intricacies of real-world anomalies.

Author Contributions: Conceptualization, L.A.E., M.S. and M.H.H.; methodology, L.A.E., M.S. and M.H.H.; software, M.H.H.; validation, L.A.E. and M.S.; formal analysis, L.A.E., M.S. and M.H.H.; investigation, L.A.E., M.S. and M.H.H.; writing—original draft preparation, M.H.H.; writing—review and editing, L.A.E., M.S. and M.H.H.; visualization, M.H.H.; supervision, L.A.E., M.S. and M.H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Information Technology Industry Development Agency (ITIDA)–Information Technology Academia Collaboration (ITAC) program under grant number CFP242.

Data Availability Statement: The data presented in this work are openly available in "UCSD Anomaly Detection Dataset" at: [http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm], accessed on 18 May 2023, reference number [29], "Avenue Dataset for Abnormal Event Detection" at [https://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html], accessed on 18 May 2023, reference number [30], "ShanghaiTech Campus dataset (Anomaly Detection)" at [https://svip-lab.github.io/dataset/campus_dataset.html], accessed on 29 November 2023, reference number [31], and "CHAD: Charlotte Anomaly Dataset" at [https://github.com/TeCSAR-UNCC/CHAD], accessed on 2 June 2024, reference number [32].

Acknowledgments: The authors gratefully acknowledge the Information Technology Industry Development Agency (ITIDA)–Information Technology Academia Collaboration (ITAC) program technical and financial support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sirisha 10 Helpful Surveillance Camera Market Statistics in 2023. Available online: https://dataprot.net/statistics/surveillance-camera-statistics/ (accessed on 18 January 2024).
2. Research, G.V. Surveillance Camera Market Size & Outlook. Available online: https://www.grandviewresearch.com/horizon/outlook/surveillance-camera-market-size/global (accessed on 18 June 2024).
3. Duong, H.-T.; Le, V.-T.; Hoang, V.T. Deep learning-based anomaly detection in video surveillance: A survey. *Sensors* **2023**, *23*, 5024. [CrossRef] [PubMed]

4. Abidine, B.M.; Fergani, L.; Fergani, B.; Oussalah, M. The joint use of sequence features combination and modified weighted SVM for improving daily activity recognition. *Pattern Anal. Appl.* **2018**, *21*, 119–138. [CrossRef]
5. Sok, P.; Xiao, T.; Azeze, Y.; Jayaraman, A.; Albert, M.V. Activity recognition for incomplete spinal cord injury subjects using hidden Markov models. *IEEE Sens. J.* **2018**, *18*, 6369–6374. [CrossRef]
6. Xiao, Q.; Song, R. Action recognition based on hierarchical dynamic Bayesian network. *Multimed. Tools Appl.* **2018**, *77*, 6955–6968. [CrossRef]
7. Hu, C.; Chen, Y.; Hu, L.; Peng, X. A novel random forests based class incremental learning method for activity recognition. *Pattern Recognit.* **2018**, *78*, 277–290. [CrossRef]
8. Saligrama, V.; Chen, Z. Video anomaly detection based on local statistical aggregates. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2112–2119.
9. Mo, X.; Monga, V.; Bala, R.; Fan, Z. Adaptive sparse representations for video anomaly detection. *IEEE Trans. Circuits Syst. Video Technol.* **2013**, *24*, 631–645.
10. Dalal, N.; Triggs, B.; Schmid, C. Human detection using oriented histograms of flow and appearance. In *Computer Vision—ECCV* 2006: 9th European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; Proceedings, Part II 9; Springer: Berlin/Heidelberg, Germany, 2006; pp. 428–441.
11. Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning temporal regularity in video sequences. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 733–742.
12. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
13. Hu, X.; Hu, S.; Huang, Y.; Zhang, H.; Wu, H. Video anomaly detection using deep incremental slow feature analysis network. *IET Comput. Vis.* **2016**, *10*, 258–267. [CrossRef]
14. Kiran, B.R.; Thomas, D.M.; Parakkal, R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *J. Imaging* **2018**, *4*, 36. [CrossRef]
15. Nayak, R.; Pati, U.C.; Das, S.K. A comprehensive review on deep learning-based methods for video anomaly detection. *Image Vis. Comput.* **2021**, *106*, 104078. [CrossRef]
16. Abdalla, M.; Javed, S.; Radi, M.A.; Ulhaq, A.; Werghi, N. Video Anomaly Detection in 10 Years: A Survey and Outlook. *arXiv* **2024**, arXiv:2405.19387.
17. Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J.W.; Carneiro, G. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 4975–4986.
18. Zhou, H.; Yu, J.; Yang, W. Dual Memory Units with Uncertainty Regulation for Weakly Supervised Video Anomaly Detection. *arXiv* **2023**, arXiv:2302.05160. [CrossRef]
19. Park, H.; Noh, J.; Ham, B. Learning memory-guided normality for anomaly detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14372–14381.
20. Lv, H.; Chen, C.; Cui, Z.; Xu, C.; Li, Y.; Yang, J. Learning normal dynamics in videos with meta prototype network. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15425–15434.
21. Habeb, M.H.; Salama, M.A.; Elrefaei, L.A. Video Anomaly Detection using Residual Autoencoder: A Lightweight Framework. *Mansoura Eng. J.* **2023**, *49*, 10. [CrossRef]
22. Smeureanu, S.; Ionescu, R.T.; Popescu, M.; Alexe, B. Deep appearance features for abnormal behavior detection in video. In *Image Analysis and Processing-ICIAP 2017: 19th International Conference, Catania, Italy, 11–15 September 2017*; Proceedings, Part II 19; Springer: Berlin/Heidelberg, Germany, 2017; pp. 779–789.
23. Hinami, R.; Mei, T.; Satoh, S. Joint detection and recounting of abnormal events by learning deep generic knowledge. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3619–3627.
24. Feng, J.-C.; Hong, F.-T.; Zheng, W.-S. Mist: Multiple instance self-training framework for video anomaly detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14009–14018.
25. Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J.W.; Carneiro, G. Weakly-supervised video anomaly detection with contrastive learning of long and short-range temporal features. In Proceedings of the 2021 18th IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021.
26. Ullah, W.; Hussain, T.; Baik, S.W. Vision transformer attention with multi-reservoir echo state network for anomaly recognition. *Inf. Process. Manag.* **2023**, *60*, 103289. [CrossRef]
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
28. Wang, Y.; Liu, T.; Zhou, J.; Guan, J. Video anomaly detection based on spatio-temporal relationships among objects. *Neurocomputing* **2023**, *532*, 141–151. [CrossRef]

29. Mahadevan, V.; Li, W.; Bhalodia, V.; Vasconcelos, N. Anomaly detection in crowded scenes. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1975–1981.
30. Lu, C.; Shi, J.; Jia, J. Abnormal event detection at 150 fps in MATLAB. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2720–2727.
31. Luo, W.; Liu, W.; Gao, S. A revisit of sparse coding based anomaly detection in stacked RNN framework. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 341–349.
32. Danesh Pazho, A.; Alinezhad Noghre, G.; Rahimi Ardabili, B.; Neff, C.; Tabkhi, H. Chad: Charlotte anomaly dataset. In *Scandinavian Conference on Image Analysis, 2023*; Springer: Cham, Switzerland, 2023; pp. 50–66.
33. Yuan, H.; Cai, Z.; Zhou, H.; Wang, Y.; Chen, X. Transanomaly: Video anomaly detection using video vision transformer. *IEEE Access* **2021**, *9*, 123977–123986. [CrossRef]
34. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. Vivit: A video vision transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 6836–6846.
35. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Proceedings, Part III 18, 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
36. Hajri, F.; Fradi, H. Vision Transformers for Road Accident Detection from Dashboard Cameras. In Proceedings of the 2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Madrid, Spain, 29 November–2 December 2022; pp. 1–8.
37. Singh, S.; Dewangan, S.; Krishna, G.S.; Tyagi, V.; Reddy, S.; Medi, P.R. Video vision transformers for violence detection. *arXiv* **2022**, arXiv:2209.03561.
38. Tahir, M.; Anwar, S. Transformers in pedestrian image retrieval and person re-identification in a multi-camera surveillance system. *Appl. Sci.* **2021**, *11*, 9197. [CrossRef]
39. Lee, Y.; Kang, P. AnoViT: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder. *IEEE Access* **2022**, *10*, 46717–46724. [CrossRef]
40. Berroukham, A.; Housni, K.; Lahraichi, M. Fine-Tuning Pre-trained Vision Transformer Model for Anomaly Detection in Video Sequences. In *International Conference on Big Data and Internet of Things, 2022*; Springer: Cham, Switzerland, 2022; pp. 279–289.
41. Lee, J.; Nam, W.-J.; Lee, S.-W. Multi-contextual predictions with vision transformer for video anomaly detection. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 1012–1018.
42. Lee, J.; Lee, S.; Cho, W.; Siddiqui, Z.A.; Park, U. Vision transformer-based tailing detection in videos. *Appl. Sci.* **2021**, *11*, 11591. [CrossRef]
43. Wurst, J.; Balasubramanian, L.; Botsch, M.; Utschick, W. Novelty detection and analysis of traffic scenario infrastructures in the latent space of a vision transformer-based triplet autoencoder. In Proceedings of the 2021 IEEE Intelligent Vehicles Symposium (IV), Nagoya, Japan, 11–17 July 2021; pp. 1304–1311.
44. Fan, W.; Shangguan, W.; Chen, Y. Transformer-based contrastive learning framework for image anomaly detection. *Int. J. Mach. Learn. Cybern.* **2023**, *14*, 3413–3426. [CrossRef]
45. Fan, W.; Shangguan, W.; Bouguila, N. Continuous image anomaly detection based on contrastive lifelong learning. *Appl. Intell.* **2023**, *53*, 17693–17707. [CrossRef]
46. Park, S.; Balint, A.; Hwang, H. Self-supervised medical out-of-distribution using U-Net vision transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention, 2021*; Springer: Cham, Switzerland, 2021; pp. 104–110.
47. Lin, Z.; Wang, H.; Li, S. Pavement anomaly detection based on transformer and self-supervised learning. *Autom. Constr.* **2022**, *143*, 104544. [CrossRef]
48. Choi, B.; Jeong, J. ViV-Ano: Anomaly detection and localization combining vision transformer and variational autoencoder in the manufacturing process. *Electronics* **2022**, *11*, 2306. [CrossRef]
49. Smith, A.D.; Du, S.; Kurien, A. Vision transformers for anomaly detection and localisation in leather surface defect classification based on low-resolution images and a small dataset. *Appl. Sci.* **2023**, *13*, 8716. [CrossRef]
50. Yao, H.; Luo, W.; Yu, W.; Zhang, X.; Qiang, Z.; Luo, D.; Shi, H. Dual-attention transformer and discriminative flow for industrial visual anomaly detection. *IEEE Trans. Autom. Sci. Eng.* **2023**, 1–15. [CrossRef]
51. De Nardin, A.; Mishra, P.; Foresti, G.L.; Piciarelli, C. Masked transformer for image anomaly localization. *Int. J. Neural Syst.* **2022**, *32*, 2250030. [CrossRef]
52. Tao, X.; Adak, C.; Chun, P.-J.; Yan, S.; Liu, H. ViTALnet: Anomaly on industrial textured surfaces with hybrid transformer. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 5009013. [CrossRef]
53. Mishra, P.; Verk, R.; Fornasier, D.; Piciarelli, C.; Foresti, G.L. VT-ADL: A vision transformer network for image anomaly detection and localization. In Proceedings of the 2021 IEEE 30th International Symposium on Industrial Electronics (ISIE), Kyoto, Japan, 20–23 June 2021; pp. 1–6.
54. Franklin, R.J.; Dabbagol, V. Anomaly detection in videos for video surveillance applications using neural networks. In Proceedings of the 2020 Fourth International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 8–10 January 2020; pp. 632–637.

55. Ullah, W.; Ullah, A.; Haq, I.U.; Muhammad, K.; Sajjad, M.; Baik, S.W. CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks. *Multimed. Tools Appl.* **2021**, *80*, 16979–16995. [CrossRef]
56. Qi, X.; Hu, Z.; Ji, G. Improved Video Anomaly Detection with Dual Generators and Channel Attention. *Appl. Sci.* **2023**, *13*, 2284. [CrossRef]
57. Ristea, N.-C.; Madan, N.; Ionescu, R.T.; Nasrollahi, K.; Khan, F.S.; Moeslund, T.B.; Shah, M. Self-supervised predictive convolutional attentive block for anomaly detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13576–13586.
58. Wang, B.; Yang, C. Video anomaly detection based on convolutional recurrent AutoEncoder. *Sensors* **2022**, *22*, 4647. [CrossRef] [PubMed]
59. Li, S.; Cheng, Y.; Zhang, L.; Luo, X.; Zhang, R. Video anomaly detection based on a multi-layer reconstruction autoencoder with a variance attention strategy. *Image Vis. Comput.* **2024**, *146*, 105011. [CrossRef]
60. Fu, Y.; Yang, B.; Ye, O. Spatiotemporal Masked Autoencoder with Multi-Memory and Skip Connections for Video Anomaly Detection. *Electronics* **2024**, *13*, 353. [CrossRef]
61. Hu, Z.-p.; Zhang, L.; Li, S.-f.; Sun, D.-g. Parallel spatial-temporal convolutional neural networks for anomaly detection and location in crowded scenes. *J. Vis. Commun. Image Represent.* **2020**, *67*, 102765. [CrossRef]
62. Hwang, I.-C.; Kang, H.-S. Anomaly Detection Based on a 3D Convolutional Neural Network Combining Convolutional Block Attention Module Using Merged Frames. *Sensors* **2023**, *23*, 9616. [CrossRef] [PubMed]
63. Lee, J.; Koo, H.; Kim, S.; Ko, H. Cognitive Refined Augmentation for Video Anomaly Detection in Weak Supervision. *Sensors* **2023**, *24*, 58. [CrossRef] [PubMed]
64. Kotkar, V.A.; Sucharita, V. Fast anomaly detection in video surveillance system using robust spatiotemporal and deep learning methods. *Multimed. Tools Appl.* **2023**, *82*, 34259–34286. [CrossRef]
65. Taghinezhad, N.; Yazdi, M. A new unsupervised video anomaly detection using multi-scale feature memorization and multipath temporal information prediction. *IEEE Access* **2023**, *11*, 9295–9310. [CrossRef]
66. Lei, S.; Song, J.; Wang, T.; Wang, F.; Yan, Z. Attention U-Net based on multi-scale feature extraction and WSDAN data augmentation for video anomaly detection. *Multimed. Syst.* **2024**, *30*, 118. [CrossRef]
67. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
68. Honda, H. Vision Transformer Pipeline (Image). 2022. Available online: https://github.com/hirotomusiker/schwert_colabdata_stor-age/blob/master/images/vit_demo/vit_input.png (accessed on 28 January 2024).
69. Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional neural networks: An overview and application in radiology. *Insights Into Imaging* **2018**, *9*, 611–629. [CrossRef] [PubMed]
70. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
71. Zhang, Q.; Wei, H.; Chen, J.; Du, X.; Yu, J. Video Anomaly Detection Based on Attention Mechanism. *Symmetry* **2023**, *15*, 528. [CrossRef]
72. Wang, Z.; Chen, Y. Anomaly detection with dual-stream memory network. *J. Vis. Commun. Image Represent.* **2023**, *90*, 103739. [CrossRef]
73. Chen, D.; Wang, P.; Yue, L.; Zhang, Y.; Jia, T. Anomaly detection in surveillance video based on bidirectional prediction. *Image Vis. Comput.* **2020**, *98*, 103915. [CrossRef]
74. Chang, Y.; Tu, Z.; Xie, W.; Luo, B.; Zhang, S.; Sui, H.; Yuan, J. Video anomaly detection with spatio-temporal dissociation. *Pattern Recognit.* **2022**, *122*, 108213. [CrossRef]
75. Le, V.-T.; Kim, Y.-G. Attention-based residual autoencoder for video anomaly detection. *Appl. Intell.* **2023**, *53*, 3240–3254. [CrossRef]
76. Kommanduri, R.; Ghorai, M. DAST-Net: Dense visual attention augmented spatio-temporal network for unsupervised video anomaly detection. *Neurocomputing* **2024**, *579*, 127444. [CrossRef]
77. Wang, Z.; Gu, X.; Gu, X.; Hu, J. Enhancing video anomaly detection with learnable memory network: A new approach to memory-based auto-encoders. *Comput. Vis. Image Underst.* **2024**, *241*, 103946. [CrossRef]
78. Li, N.; Chang, F. Video anomaly detection and localization via multivariate gaussian fully convolution adversarial autoencoder. *Neurocomputing* **2019**, *369*, 92–105. [CrossRef]
79. Zhang, Z.; Zhong, S.-h.; Fares, A.; Liu, Y. Detecting abnormality with separated foreground and background: Mutual generative adversarial networks for video abnormal event detection. *Comput. Vis. Image Underst.* **2022**, *219*, 103416. [CrossRef]
80. Tang, Y.; Zhao, L.; Zhang, S.; Gong, C.; Li, G.; Yang, J. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognit. Lett.* **2020**, *129*, 123–130. [CrossRef]
81. Hao, Y.; Li, J.; Wang, N.; Wang, X.; Gao, X. Spatiotemporal consistency-enhanced network for video anomaly detection. *Pattern Recognit.* **2022**, *121*, 108232. [CrossRef]
82. Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M.R.; Venkatesh, S.; Hengel, A.v.d. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1705–1714.
83. Abati, D.; Porrello, A.; Calderara, S.; Cucchiara, R. Latent space autoregression for novelty detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 481–490.

84. Fan, Y.; Wen, G.; Li, D.; Qiu, S.; Levine, M.D.; Xiao, F. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. *Comput. Vis. Image Underst.* **2020**, *195*, 102920. [CrossRef]
85. Chang, Y.; Tu, Z.; Xie, W.; Yuan, J. Clustering driven deep autoencoder for video anomaly detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XV 16; Springer: Cham, Switzerland, 2020; pp. 329–345.
86. Deepak, K.; Chandrakala, S.; Mohan, C.K. Residual spatiotemporal autoencoder for unsupervised video anomaly detection. *Signal Image Video Process.* **2021**, *15*, 215–222. [CrossRef]
87. Feng, J.; Liang, Y.; Li, L. Anomaly detection in videos using two-stream autoencoder with post hoc interpretability. *Comput. Intell. Neurosci.* **2021**, *2021*, 7367870. [CrossRef] [PubMed]
88. Cho, M.; Kim, T.; Kim, W.J.; Cho, S.; Lee, S. Unsupervised video anomaly detection via normalizing flows with implicit latent features. *Pattern Recognit.* **2022**, *129*, 108703. [CrossRef]
89. Yao, S.; Noghre, G.A.; Pazho, A.D.; Tabkhi, H. Evaluating the Effectiveness of Video Anomaly Detection in the Wild: Online Learning and Inference for Real-world Deployment. *arXiv* **2024**, arXiv:2404.18747.
90. Noghre, G.A.; Pazho, A.D.; Tabkhi, H. An Exploratory Study on Human-Centric Video Anomaly Detection through Variational Autoencoders and Trajectory Prediction. In Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 1–6 January 2024; pp. 995–1004.
91. Markovitz, A.; Sharir, G.; Friedman, I.; Zelnik-Manor, L.; Avidan, S. Graph embedded pose clustering for anomaly detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10539–10547.
92. Hirschorn, O.; Avidan, S. Normalizing flows for human pose anomaly detection. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 13545–13554.
93. Jia, D.; Zhang, X.; Zhou, J.T.; Lai, P.; Wei, Y. Dynamic thresholding for video anomaly detection. *IET Image Process.* **2022**, *16*, 2973–2982. [CrossRef]
94. Kumar, K.; Shrimankar, D.D.; Singh, N. Eratosthenes sieve based key-frame extraction technique for event summarization in videos. *Multimed. Tools Appl.* **2018**, *77*, 7383–7404. [CrossRef]
95. Jadon, S.; Jasim, M. Unsupervised video summarization framework using keyframe extraction and video skimming. In Proceedings of the 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 30–31 October 2020; pp. 140–145.