applied sciences

MDPI

Article

Three-Stage Deep Learning Framework for Video Surveillance

Ji-Woon Lee and Hyun-Soo Kang *

Department of Information and Communication Engineering, School of Electrical and Computer Engineering, Chungbuk National University, Cheongju-si 28644, Republic of Korea; dlwldns2@gmail.com
* Correspondence: hskang@cbnu.ac.kr

Abstract: The escalating use of security cameras has resulted in a surge in images requiring analysis, a task hindered by the inefficiency and error-prone nature of manual monitoring. In response, this study delves into the domain of anomaly detection in CCTV security footage, addressing challenges previously encountered in analyzing videos with complex or dynamic backgrounds and long sequences. We introduce a three-stage deep learning architecture designed to detect abnormalities in security camera videos. The first stage employs a pre-trained convolutional neural network to extract features from individual video frames. Subsequently, these features are transformed into time series data in the second stage, utilizing a blend of bidirectional long short-term memory and multi-head attention to analyze short-term frame relationships. The final stage leverages relative positional embeddings and a custom Transformer encoder to interpret long-range frame relationships and identify anomalies. Tested on various open datasets, particularly those with complex backgrounds and extended sequences, our method demonstrates enhanced accuracy and efficiency in video analysis. This approach not only improves current security camera analysis but also shows potential for diverse application settings, signifying a significant advancement in the evolution of security camera monitoring and analysis technologies.

Keywords: deep learning; Transformers; video surveillance; anomaly detection; RNN

check for updates

Citation: Lee, J.-W.; Kang, H.-S. Three-Stage Deep Learning Framework for Video Surveillance. *Appl. Sci.* **2024**, *14*, 408. https://doi.org/10.3390/app14010408

Academic Editors: Zhaoqing Pan, Bo Peng and Jinwei Wang

Received: 30 November 2023
Revised: 20 December 2023
Accepted: 22 December 2023
Published: 2 January 2024

1. Introduction

The increasing prevalence of crime, ranging from minor offenses to violent acts, as highlighted in the "2022 Crime Analysis" report by the Supreme Prosecutors' Office of Korea, and shown in Figure 1, underscores the necessity of enhanced security measures. In this regard, the field of the video-based detection of abnormal situations, as referenced in numerous studies ([1–8]), is experiencing a surge in interest. Research in this area, especially studies investigating surveillance video anomaly detection (SVAD) using deep learning [9], is progressing rapidly. This research domain has evolved to encompass a broader scope, extending beyond behavior-based detection to include advanced areas like human facial emotion recognition for anomaly detection [10], illustrating the expanding reach and depth of investigations in this field. These systems function by capturing real-time images across various indoor and outdoor settings, both public and private. However, the proliferation of security cameras leads to the generation of an immense volume of images, making manual monitoring susceptible to human error, as well as being labor-intensive and costly. Therefore, the automation of anomaly recognition in video footage is imperative.

Video-based anomaly detection, in essence, aims to identify any unusual or non-standard activities and situations captured in video data, including incidents like violence, fires, health emergencies, unauthorized entry, and abduction. Unlike still images, which contain only spatial information, video data encompass temporal elements as well, necessitating the fusion of spatial and temporal data for accurate analysis due to the correlation between neighboring frames.

Approaches to video-based anomaly detection, integrating both spatial and temporal aspects, encompass various learning paradigms: supervised, unsupervised, and semi-supervised (one-class) learning. Supervised learning, which utilizes labeled data for both normal and abnormal instances, generally achieves higher accuracy. Nonetheless, it is limited by the need for diverse atypical samples. To overcome these challenges, our approach focuses on semi-supervised learning (or one-class learning), which primarily uses normal sample data for training. Within this domain, the one-class support vector machine (OCSVM) is widely used, and recent advancements in deep support vector data description (Deep SVDD) [11] have further extended this concept into the realm of deep learning.

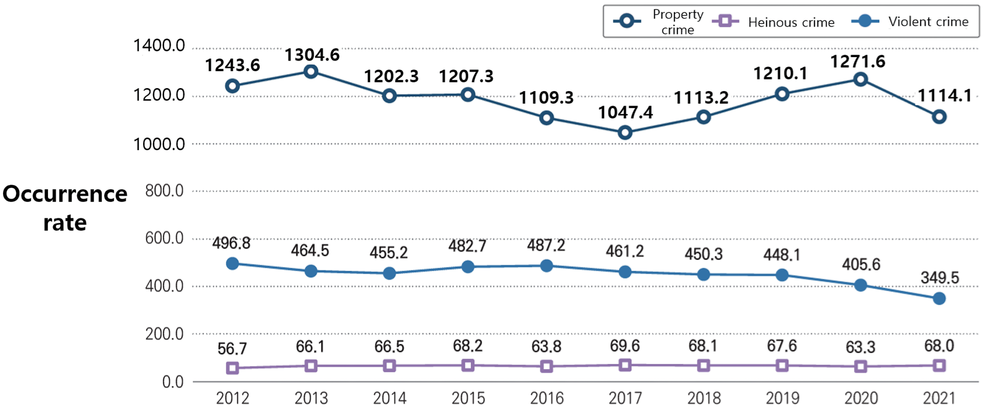

Figure 1. Trends in accrual costs by major criminal crimes of prosecution in 2012–2021. (Source: "2022 Crime Analysis" by the Supreme Prosecutors' Office of Korea).

To address the limitations of semi-supervised learning methods, which require labeled normal samples, unsupervised learning techniques are being investigated. These techniques operate under the assumption that the majority of the data consist of normal samples and can be learned without labels. A prime example of this approach is the autoencoder [12,13], featuring an encoder that compresses the input into a feature vector and a decoder that reconstructs the original input from this feature vector.

While previous studies have primarily concentrated on harnessing three-dimensional information for learning, they have struggled with reliability in analyzing complex or dynamic video backgrounds and long video sequences. To overcome these issues, our research introduces a new three-tiered network model for video anomaly detection, adept at handling both the temporal and spatial aspects of video data. The key contributions of this study are outlined as follows.

1. Introduction of a Three-Stage Network: This paper introduces a groundbreaking three-stage deep learning architecture for anomaly detection in video streams, notably in CCTV systems. This approach effectively addresses both the spatial and temporal dimensions of video data, advancing beyond the capabilities of existing methods.
2. Utilization of Pre-Trained Networks for Feature Extraction: The first phase of our approach leverages pre-trained networks, such as Vision Transformers (ViT), to extract feature vectors from video frames. This process enables an in-depth analysis of the spatial attributes of the footage, laying the groundwork for subsequent sophisticated analysis.
3. Integration of BiLSTM with Multi-Head Attention for Temporal Dynamics: The second phase combines BiLSTM with multi-head attention, focusing on the temporal relationships between frame feature vectors. This synergy enhances the model's ability to identify anomalies, leading to a marked improvement in detection accuracy.
4. Customized Transformer Module for Enhanced Pattern Recognition: In the final phase, a specially adapted Transformer module is employed. This module is proficient in handling relative positional embeddings and recognizing long-range dependencies, thereby boosting the processing efficiency. It is specifically tailored to anomaly detection in video data.

5. Superior Detection Capabilities: Through comprehensive testing and evaluation, our model demonstrates superior performance over existing leading-edge techniques, particularly in handling the complexities and variances found in CCTV footage.
6. Practical Real-World Application Potential: The proposed model exhibits significant potential for real-world applications, such as in security monitoring and automated surveillance systems. This represents a significant leap forward in the realms of video anomaly detection and deep learning technology.

The rest of the paper is organized as follows. Section 2 provides the contextual background for this research while exploring the existing literature and prior studies. Section 3 begins with an explanation of the general Transformer mechanism in Section 3.1; subsequently, Section 3.2, a comprehensive explanation of the proposed overall architecture, is presented, starting with an in-depth overview. Following this, Sections 3.2.1–3.2.3 sequentially delve into each stage, providing detailed descriptions. Section 3.2.4 then offers a detailed exposition of the distinctions from the general Transformer mechanism outlined in Section 3.1. Finally, Section 3.3 concludes with an explanation of the loss function used.

Section 4 begins with an explanation of the dataset used in the research and the preprocessing stages. It continues through the final result analysis. In Section 4.1, the introduction of the dataset is provided, followed by the presentation of the experimental environment in Sections 4.2 and 4.3. Subsequently, Sections 4.4 and 4.5 conduct a detailed analysis of the performance comparison of the proposed model. Lastly, Section 4.6 conducts an ablation study for additional validation. Finally, in Section 5, we summarize the key aspects of the study, outline our contributions to the field, and suggest directions for future research.

2. Related Works

Currently, research on anomaly detection in CCTV videos is being actively conducted using various techniques. In order to compare and analyze the proposed architecture, we investigated various methodologies and confirmed their performance. In this section, we introduce diverse methodologies. First, Bilinski, P. [14] proposed an extension of Improved Fisher Vectors (IFV) for videos, which enables the spatiotemporal localization of features to increase accuracy. Additionally, they re-formalized IFV and proposed a sliding window approach that utilizes aggregated region table data structures for violence detection. Subsequent advancements in technology have enabled more sophisticated analysis. Some techniques utilize optical flow [15–18] before training the network. Optical flow refers to the visible motion pattern of an object captured in consecutive video frames. When the movement of an object occurs in an image, it is projected onto a 2D image space, where a myriad of vectors from the 3D space can be projected as vectors in the 2D image space because the dimensions are reduced. Examples of these advanced techniques include the following.

Skeleton-based anomaly detection utilizes graph convolutional networks (GCNs). A GCN predicts the label of each node in a given input graph and updates its hidden state using Equation (1), where H represents the hidden state matrix of the lth hidden layer, A represents the adjacency matrix with self-connections and an identity matrix, and W represents the weight matrix.

$$H^{(l+1)} = \sigma(AH^l W^l + b^l), \tag{1}$$

Skeleton-based anomaly detection [19] converts human images into graphs and analyzes the interactions within these graphs to detect anomalies. For example, Garcia-Cobo, G. [20] proposed an architecture using human pose extractors and ConvLSTM. The author relied on what they considered to be the most essential information to detect human bodies and their interactions, using human pose extractors and optical flow for this purpose. The architecture consists of RGB and motion detection pipelines, each of which analyzes image distributions and skeletal movements. Su, Y. [21] proposed an architecture that learns interactions between skeletal points using 3D skeleton point clouds. They aimed to represent videos as human skeleton point clouds using the multi-head Skeleton Points Interaction Learning (SPIL) module and perform inference for violence recognition in videos.

There are also studies on anomaly detection techniques using Conv3D. Typically, CNNs apply operations to images using two-dimensional (2D) kernels. However, this method can only be applied to static images. Meanwhile, CCTV image analysis includes both still images and the analysis of temporal data, making it impossible to perform analysis using a generic CNN. Therefore, the passage of time should be incorporated into the 2D CNN, and a Conv3D network has been developed for this purpose. Conv3D takes inputs in three dimensions, and its kernel is also three-dimensional (3D). The rest of the computation is the same as that of a conventional 2D CNN; however, the direction of movement of the kernel occurs along the x-, y-, and z-axes, and the convolution operation is applied to convert the $n \times n \times n$ 3D inputs into a single output.

In previous studies, several techniques have incorporated optical flow or attention modules [22]. For instance, Cheng, M. [23] proposed a flow-gated network that leveraged both 3D-CNN and optical flow. Video frames were divided into RGB and optical flow channels, each processed through Conv3D networks. The outputs from the last layers of the RGB and optical flow channels underwent ReLU and sigmoid operations before concatenation. The resulting output was used for anomaly detection. Sultani, W. [4] used multiple instance learning (MIL) and 3D convolution to differentiate between normal and abnormal scenarios. The author separated normal and abnormal video clips into positive and negative bags, respectively. Features were extracted using 3D convolution, and a multiple instance ranking objective function was proposed to rank the two instances with the highest anomaly scores within each bag. Degardin, B. [24] introduced an approach to detect abnormal events in surveillance videos using a recurrent learning framework, a random forest ensemble, and novel terms for score propagation. The author used a weakly supervised network model to classify videos into bags with positive and negative instances and detected abnormal situations in unlabeled data using 3D convolution and a Bayesian classifier. Degardin, B.M. [2] utilized a Gaussian Mixture Model (GMM) composed of a soft clustering technique to connect data points to clusters. The author introduced a new MIL solution with a loss function that incorporated both a two-kernel GMM and the estimated parameters of normal distributions. Features were extracted using 3D convolution. Mohammadi, H. [25] improved the model accuracy by adding hard attention to semi-supervised learning, utilizing the optical flow extracted from input data, and applying 3D convolution.

There are also studies on anomaly detection techniques using ConvLSTM. To understand ConvLSTM [26], we must first examine long short-term memory (LSTM) [27]. In conventional RNNs, when long intervals of data are input, backpropagation requires a long time, and slope loss occurs. To prevent this, an LSTM with a state store, an input gate, and an output gate was developed. LSTM is faster than conventional methods because each gate unit is trained to appropriately open and close the flow, which solves the slope loss problem.

ConvLSTM introduces convolutional recurrent cells into the basic LSTM structure and applies convolutional operations to the images as inputs. Consequently, it can effectively learn the spatial characteristics of the image. Moreover, because it has the characteristics of LSTM, it can also learn the temporal continuity among image frames.

Utilizing these characteristics, research has been conducted on anomaly detection techniques using ConvLSTM [28]. For example, Islam, Z. [29] proposes an efficient two-stream deep learning architecture. The architecture leverages separable convolution LSTM (sepconvlstm) and a pre-trained MobileNet. One stream takes suppressed frames as input to model the background, while the other stream processes the differences between adjacent frames. Additionally, the author presents three fusion methods to combine the output feature maps of the two streams. Sudhakaran, S. [30] introduces a combined architecture of 2D convolution and ConvLSTM. Features are extracted frame-wise using a pre-trained

AlexNet, followed by convlstm layers with 256 filters, batch normalization, and ReLU activation functions to detect anomalies.

Additionally, there are anomaly detection techniques based on Transformers, which serve as the foundation for the present study. Deshpande, K. [31] conducted anomaly detection using Videoswin Transformer feature extraction, attention layers, and the Robust Temporal Feature Magnitude (RTFM). Furthermore, Jin, P. [32] proposed a new model called ANomaly Detection with Transformers (ANDT) to detect abnormal events in aerial videos. In this approach, videos are processed into a sequence of tubes, and a Transformer encoder is employed to learn spatiotemporal features. Subsequently, the decoder combines with the encoder to predict the next frame based on the learned spatiotemporal representation.

Liu, Y. [33] provides a hierarchical GVAED taxonomy that systematically organizes the existing literature by supervision, input data, and network structure, focusing on the recent advances, such as weakly supervised, fully unsupervised, and multimodal methods. Furthermore, focal loss [34] is employed as part of the network architecture.

3. Proposed Method

The network proposed in this paper is based on Transformers [35] and has been restructured to improve the detection accuracy according to the objectives. This section begins by elucidating the architectural frameworks of the Transformer and ViT, which serves as the primary feature extractor. Subsequently, we delve into a comparative analysis, highlighting the distinctions between the conventional Transformers and the novel Transformer proposed in this research. Finally, we provide a comprehensive overview of the entire architectural framework.

3.1. Transformer Mechanism

Before the advent of Transformers, the "Seq2Seq model" [36–39] based on RNNs was recognized as the dominant methodology for the processing of sequence data.

However, these RNN models, like the Seq2Seq model, have several limitations. First, the model must process the inputs sequentially, which limits computation parallelization. Second, the window size of the model is fixed, which renders it difficult to learn the relationships between items that are far apart in a sequence, particularly because of the problem of gradient decay or explosion.

A proposed solution to this problem is an attention mechanism. Recently, improved RNNs, such as LSTM and GRU, have improved the window size to increase the relationships between distant items to a certain extent; however, the window size of these models is still limited. In contrast, the attention mechanism can theoretically have an infinite window size under the premise that computing resources are available, it can be parallelized for fast learning speeds, and it can be effectively applied to very large datasets.

In this attention mechanism, self-attention plays a crucial role in considering the relationships between words in the input sequence. Self-attention introduces the concepts of positional encoding: the query, key, and value. Since self-attention does not explicitly consider sequential order information, it adds vectors containing positional information of the same dimension as the input used in the model. This is known as positional encoding.

Once positional encoding is completed, self-attention utilizes three elements: the query, key, and value. The query extracts information about the given input and is used to evaluate the relationships with other input elements. The key determines how much attention each element of the input sequence should receive, and it combines the query and key to calculate attention scores, which represent how much attention each element should receive. These scores contain the actual information and, by weighting according to attention scores, form the final output. The value of each element is weighted based on its importance and attention, and these values are combined to generate the final output of attention.

The self-attention mechanism is particularly well suited for dealing with sequential data where information changes over time or where correlations are important, such as videos, which are the type of data used in this study.

However, in practice, the data that we deal with have complex structures with many different attributes and perspectives. This diversity cannot be fully captured by a single attentional mechanism because it operates on a limited set of information and is unlikely to reflect the interactions between the various attributes of complex data. To address these issues, the multi-head attention technique is used, which employs multiple attention heads in parallel. The advantage of this structure is that each attention head captures information from an independent perspective, and these diverse perspectives are integrated to form a richer representation.

Transformers are architectures that are not based on an RNN structure but introduce self-attention and multi-head attention to the encoder and decoder parts. This integrates the advantages of self-attention and multi-head attention and has exhibited excellent performance in various natural language processing tasks, such as translation, document summarization, question and answer, and time series data processing. The basic structure of the Transformer is illustrated in Figure 2.

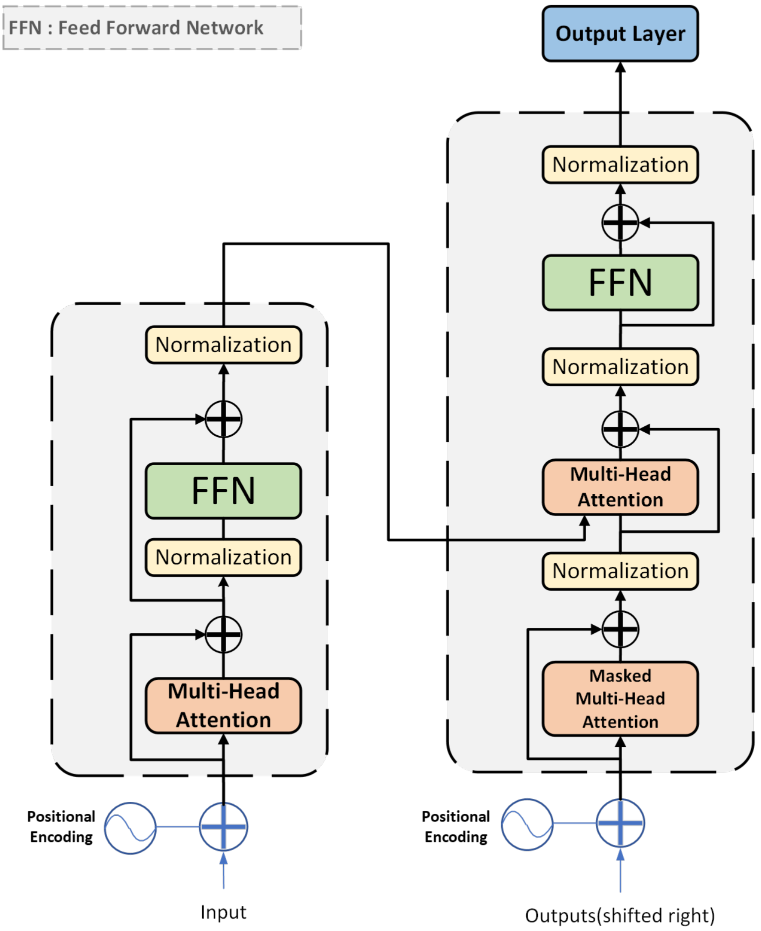

Figure 2. Architecture of Transformer.

3.2. Proposed Overall Architecture

Anomaly detection using the networks described in Section 2 has generally been performed by incorporating temporal features into the input data using optical flow or by configuring the network itself to learn both spatial and temporal features. However, the network introduced in this study first obtained the spatial characteristics and then used them to learn and compute the temporal characteristics, based on which the three-stage framework is proposed. The proposed three-stage framework structure is depicted in Figure 3.

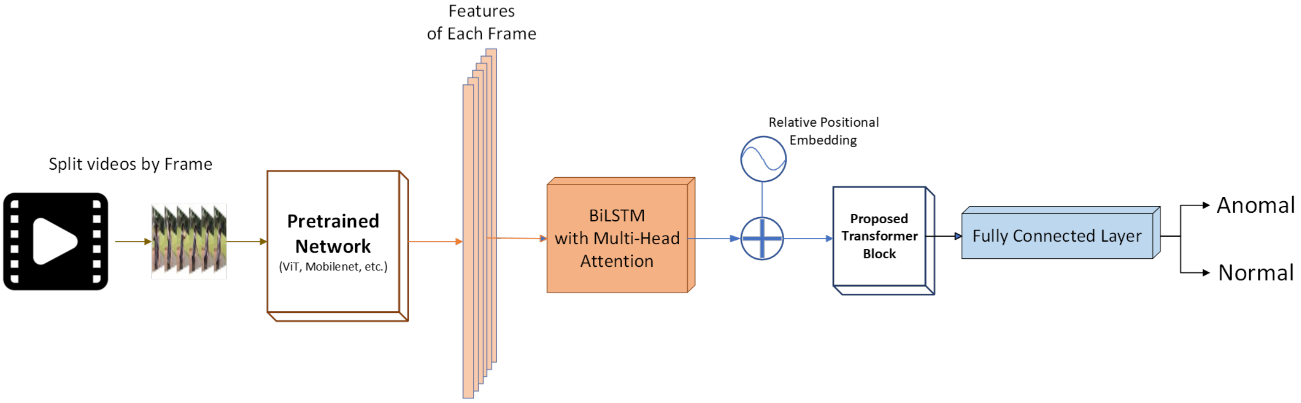

Figure 3. Proposed overall framework architecture.

3.2.1. First Stage

The first stage involves extracting features from individual still-frame images. In this study, learning was not conducted separately to obtain spatial characteristics. Using networks that learned well with large amounts of data, features were quickly extracted for each frame image, and various pretrained networks were used with ViT [40] as the main focus in this study.

ViT was developed to apply the Transformer architecture, as explained in Section 3.1, to image data analysis. The core concept of ViT, as depicted in Figure 4, involves dividing the input image into a series of fixed-size patches. Before being used as input to the Transformer, each patch is flattened into a single dimension. Subsequently, the flattened patches have a 'class token' added, similar to the 'CLS token' that is added to sentences in Transformers. Additionally, position embeddings are added to each patch image to recognize the relative positions of the patches. These preprocessed patches are then fed into the Transformer encoder, which learns the relationships between the image features and the patches.

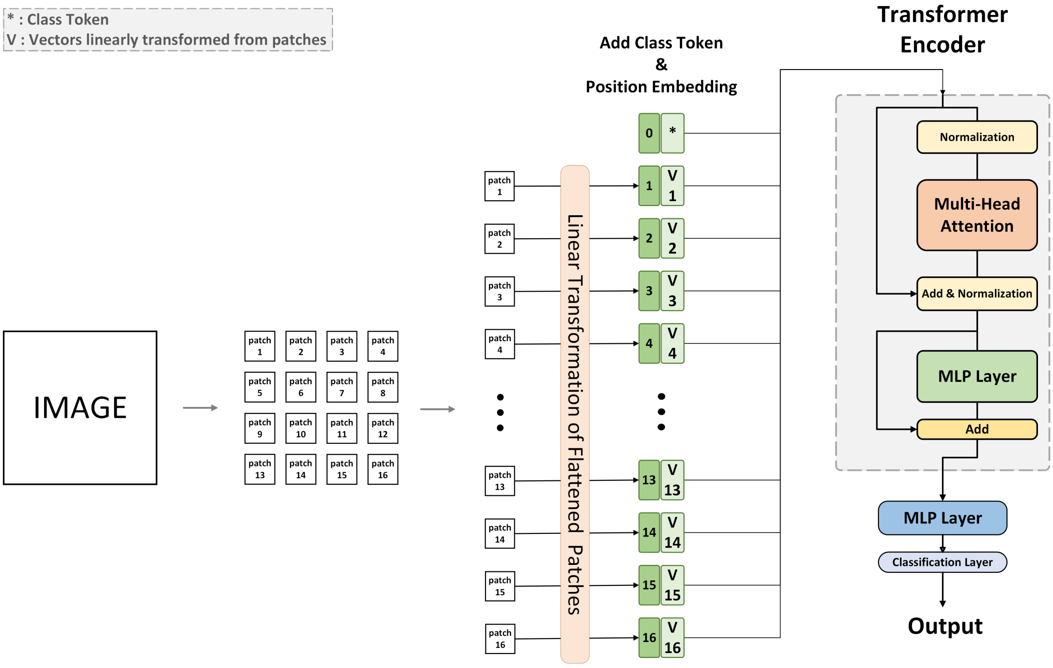

Figure 4. Architecture of Vision Transformer.

Thus, ViT applies the characteristics of Transformers to image data, shows excellent performance when used with large datasets, and is used in various vision tasks.

When the image feature extraction was completed, 768 features were outputted for each frame image, and when applied to each video, the final output was in the form of (number of videos, frame length, 768). The output data from the first stage are further processed through the combination layer of BiLSTM and multi-head attention in the second stage.

3.2.2. Second Stage

The bidirectional LSTM (BiLSTM) emerges as a significant improvement in the development of the long short-term memory (LSTM) networks mentioned in Section 2. BiLSTM augments the fundamental LSTM architecture by incorporating bidirectional processing capabilities. This modification allows for the simultaneous analysis of sequence data in both the forward and backward directions, enabling the consideration of past and future information at each time step. Consequently, BiLSTM surpasses unidirectional LSTM in capturing sequence-dependent characteristics, particularly excelling in extracting more accurate insights from temporally correlated data by considering both past and future contexts.

In the second stage, the BiLSTM and multi-head attention techniques are integrated to effectively model the temporal characteristics of sequential data. BiLSTM integrates a bidirectional processing capability into LSTM, allowing it to simultaneously analyze sequence data in both the forward and backward directions, thus understanding the overall temporal flow of the sequence. On the other hand, multi-head attention processes and integrates information from various points within the sequence in parallel. This combination proves particularly beneficial in handling complex, temporally evolving information, such as video data.

The integration of bidirectional LSTM (BiLSTM) with multi-head attention layers is known for its enhanced capability in handling long-term dependencies in sequence data. However, it still encounters limitations in maintaining and transmitting information from the distant past when dealing with exceedingly long sequences, as the length of the sequence increases. To overcome these limitations, a third stage is introduced, where the data processed through the second stage are further processed by Transformer encoder layers. This additional step aids in addressing the difficulties associated with lengthy sequence data, enhancing the model's ability to effectively retain and utilize historical information.

3.2.3. Third Stage

For the utilization of data processed through the BiLSTM and multi-head attention layers as input for the Transformer encoder layer, it is essential to perform a position embedding operation on the data. Typically, data with a temporal dimension, such as videos, are subjected to absolute position embeddings to encode unique positional information. However, in this study, relative position embeddings are employed. Relative position embeddings are particularly beneficial for video data processing, as they better grasp the dynamic context by considering the relative temporal distances between frames. This approach is suitable for videos of varying lengths, given its basis in relative relationships. Additionally, it has the advantage of effectively capturing the interactions between adjacent and distant frames, enhancing the understanding of the inter-frame dynamics.

The data passed through the relative position embedding layer were used as input to the Transformer encoder architecture, specifically tailored to anomaly detection, as proposed in this paper, and the Transformer encoder architecture is shown in Figure 5.

The output from the Transformer encoder layer structure undergoes further refined processing. The global maxpooling layer plays a crucial role in reducing the complexity of the data by extracting the most significant information from each feature map while preserving key features. This is followed by a dropout layer to prevent overfitting and enhance the model's generalization ability. Finally, a sigmoid operation is employed to determine whether the input video data are normal or exhibit an anomaly.

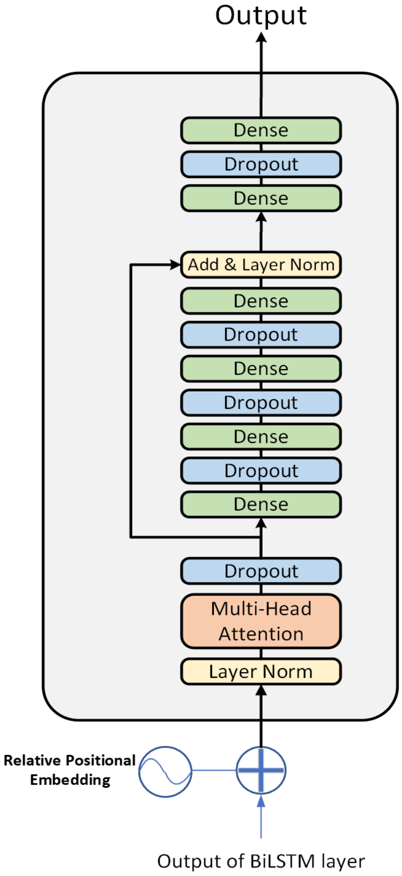

Figure 5. Architecture of proposed Transformer block.

The introduction of a fully connected layer is pivotal in effectively classifying any anomalies within the input video data based on the output of the proposed architecture. This layer simplifies complex features and focuses on critical information for classification tasks, achieving high accuracy and efficiency. This structural approach is key to the precise analysis of video data and the reliable detection of anomalous states.

This framework was trained independently for each video, taking into account the correlations between the features of frames within each video. Such an approach enables the more precise and effective analysis of complex and diverse video data, offering potential applications in various real-world scenarios.

3.2.4. Proposed Transformer Architecture

The modified Transformer encoder block introduces a novel approach optimized for the processing of video data. Compared to the conventional Transformer architecture, this new design focuses more deeply on analyzing the intricate relationships between video frames. Key improvements include the application of layer normalization and multi-head attention in the initial stages, along with a more complex, dense projection layer configuration. These modifications enable the model to process and integrate the features of each frame more efficiently.

The first dense projection layer consists of four dense layers and three dropout layers, which help to prevent overfitting and enhance the model's generalization capabilities while also extracting more complex features. The process of combining the input and output of this layer, followed by additional layer normalization, smooths the information flow and increases the network stability, contributing to the model's ability to learn and integrate finer details from the input data.

Subsequently, the second dense projection layer, comprising two dense layers and one dropout layer, further refines and compresses the features extracted by the first layer. This layer plays a crucial role in effectively condensing and extracting essential information from the final output, thereby optimizing the network's predictive performance. The use of these two stages of dense projection layers allows for the efficient handling of the complex nature of video data, enabling more precise analysis and learning, and is expected to yield superior performance in areas such as video analysis, object tracking, and action recognition.

Overall, the innovative structure of this Transformer encoder block focuses on deeply analyzing and processing the complex characteristics of video data, offering superior performance compared to traditional methods.

3.3. Loss Function

To train the Transformer-based framework proposed in this study, binary focal loss [34] was adopted and used as the loss function. Focal loss was first developed for object detection. Object detection is mainly divided into two types: two-stage and one-stage detectors. The difference is that a two-stage detector performs localization and classification, whereas a one-stage detector processes localization and classification simultaneously. Two-stage detectors have better accuracy performance, but they have the disadvantage of a lower speed. However, one-stage detectors have the opposite disadvantage of being faster but less accurate, and focal loss is designed to improve the accuracy. The main problem with one-stage detectors is class imbalance during training. For example, the difference between the proportions of real objects and background in an image is very large, with real objects accounting for a very small percentage of the background. Because of this imbalance, the training process frequently incorrectly predicts the presence of an object against a background. Outputting an object box in the background is learned as an error, but its frequency is so high that it interferes with learning and reduces the efficiency.

Traditionally, cross-entropy, which focuses on penalizing incorrect predictions, has been adopted and used. However, focal loss reduces the weight of the loss by multiplying the basic cross-entropy loss by a specific weighting term. This specific weighting term is based on the model-generated prediction probabilities and dynamically adjusts the weight of the prediction errors in each class. This dynamically adjusted weighting term allows the model to focus on a smaller number of real-world entities, thus mitigating the class imbalance. The mathematical expression for this focal loss is given by Equation (2).

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t), \tag{2}$$

where p_t is the probability that the model makes a correct prediction. Further, the $(1 - p_t)^{\gamma}$ term is a specific weighting term that dynamically adjusts the weights for prediction errors in each class based on the prediction probabilities generated by the model. In this term, the hyperparameter γ controls the model's focus on difficult examples. As the value of γ increases, the model penalizes challenging examples more, being more focused on learning these challenging examples. In Equation (2), a γ value of 0 implies a standard cross-entropy loss.

In this study, we utilized the characteristics of focal loss to overcome the class imbalance problem that occurs in anomaly detection problems, where there are significantly fewer anomalies than in normal situations, and we obtained a high-performance anomaly detection model.

4. Experiment and Results

4.1. Dataset Introduction

Real-world video data from CCTV and other security cameras are required for the performance verification and learning of the architecture proposed in this study. However, there are limitations to collecting and organizing large amounts of video data manually, and video data from security cameras, such as roads, individuals, institutions, and companies, are limited owing to various legal and ethical issues related to privacy and security. Therefore, we use the UBI-Fights [24], Hockey Fights [41], Movie Fights [42], and Crowd [43] datasets, which are widely used in anomaly detection research for training and verification purposes. UBI-Fights is a large dataset consisting of 80 h of videos, which includes 216 anomaly videos and 784 everyday scenes, totaling 1000 30 fps videos. The videos were stripped of unnecessary video segments (video introductions, news, etc.) that interfered with the learning process. The UBI-Fights dataset encompasses videos from diverse locations and includes various situations, captured across different times of the day and night. An example video is shown in Figure 6.

Figure 6. Gallery of the UBI-Fights dataset.

In addition to the UBI-Fights dataset, the other three major datasets—Hockey Fights, Movie Fights, and Crowd—each offer unique and varied environments. Hockey Fights includes various conflicts and anomalous situations in hockey games, while Movie Fights features scenes from actual movies. Crowd includes both normal and abnormal situations in crowd settings.

All three datasets consist of video data with a frame rate of 25 fps, as listed in Table 1. The Crowd dataset includes videos ranging from 3 to 5 s in length, while the others are all 1 s long. This format is conducive to accurately analyzing rapidly evolving anomalous situations, and the structure of these datasets is deemed crucial in assessing the applicability and efficiency of the proposed framework in a variety of contexts.

Table 1. Dataset information.

Dataset	Video Length (s)	Fps	Number of Anomalies	Number of Normal Instances
UBI-Fights	16–6416	30	216	784
Hockey Fights	1	25	500	500
Movie Fights	1	25	100	100
Crowd	3–5	25	123	123

All datasets were divided into training and validation datasets to validate the performance of the network. For the UBI-Fights dataset, 40 abnormal and 27 normal situations were used for validation, totaling 67 data points, referring to the list of validation files provided. For all other datasets, 30% of the total dataset was used as validation data.

For the training process, each frame image was resized to a resolution of 800 × 800 pixels. Due to the sufficient length of videos in the UBI-Fights dataset, 150 frames per video were utilized. However, for the remaining datasets, each consisting of approximately one second of footage, nearly all frames within a video, 30 frames, were used. A challenge arose in certain instances, including specific videos from the UBI-Fights dataset, where the total number of frames was less than the required frame count. To resolve this, when the total number of frames in a video was fewer than the set number, the last saved frame was duplicated in reverse order up to the designated frame count, effectively reconstructing the video to simulate a reverse playback effect.

4.2. Optimizer

Various optimization algorithms have been developed, and, in this study, we conducted a comprehensive evaluation of different optimization techniques to measure the peak performance of the proposed model. The results indicated that, among these techniques, AdamW demonstrated the highest performance in terms of generalization capabilities and was also effective in preventing overfitting. AdamW, an enhancement of the Adam optimization method, separates weight decay from the optimization process, thereby creating a more stable learning environment. This separation has been proven to provide better generalization performance and is particularly effective in preventing overfitting. The experimental results of this study highlight the significant role of AdamW in enhancing the model's generalization ability and learning stability.

4.3. Experimental Environment

For the training purposes of this study, a high-performance graphics processing unit (GPU), the RTX 4090, along with a 13th generation Intel i7 CPU, was utilized to ensure rapid computation speeds and efficient data processing capabilities. This setup was designed to provide ample memory capacity to smoothly handle large datasets.

Regarding the software environment, the study operated under the Windows 11 operating system(version: 22H2, OS build: 22621.2861 from Microsoft Corporation, Executed in Cheong-Ju Si, Republic of Korea), with Python (version 3.10.12 from python.org, Executed in Cheong-Ju Si, Republic of Korea)as the primary programming language. The processes of data preprocessing, model building, training, and testing were conducted using functions provided by TensorFlow Keras.

4.4. Results

We named our proposed baseline framework "BiLT", which indicates the combination of BiLSTM, multi-head attention, and the Transformer Encoder. We proposed five different network architectures based on BiLT and compared their performance to existing methodologies. Each network was designed considering specific objectives and environments, and their characteristics are as follows.

1. VBiLT (Vision Transformer-based BiLT): This architecture uses Vision Transformer as the feature extractor. It effectively captures the global context and can learn features by emphasizing important parts through the attention mechanism.
2. MobileBiLT (MobileNet-based BiLT): In this architecture, MobileNet [44] is used as the feature extractor. It inherits MobileNet's characteristics to provide reasonable accuracy with reduced computational complexity.
3. DenseBiLT (DenseNet-based BiLT): This architecture utilizes DenseNet [45] as the feature extractor. It takes advantage of DenseNet's feature reuse mechanism, allowing it to preserve the initial features by utilizing information from previous layers.
4. CvtBiLT (ConvNeXt Tiny-based BiLT): ConvNeXt Tiny is used as the feature extractor in this architecture. ConvNeXt [46] is inspired by Transformer models and enhances the performance of traditional convolutional networks. This architecture provides high performance at a reasonable computational cost.
5. CvlBiLT (ConvNeXt Large-based BiLT): In this architecture, ConvNeXt Large is employed as the feature extractor. ConvNeXt Large has more parameters and deeper layers compared to ConvNeXt Tiny, making it suitable for handling complex and diverse data.

The performance evaluation in this study was conducted through a comparison with existing networks. To facilitate this comparison, we adopted the same performance metrics as those used in the previous networks investigated in this work, ensuring consistency in the evaluation criteria. In all studies utilizing the UBI-Fights dataset, due to its imbalanced nature, model performance comparisons were made using the Area Under the Curve (AUC) metric, which better reflects this imbalance, instead of accuracy. For studies employing the Hockey Fights, Movie Fights, and Crowd datasets, model performance comparisons were conducted using accuracy, as it effectively measures the classification precision. As mentioned in Section 4.1, the proposed model underwent training and validation using the divided training and validation datasets.

The results of validating the proposed architecture using the UBI-Fights dataset, followed by a comparison of the AUC performance with existing networks, are presented in Table 2.

Table 2. Comparison of result on UBI-Fights dataset.

Method	UBI-Fights
	AUC
Sultani et al. [4]	89.2
SS-Model [24]	81.9
SS-Model + WS-Model + Sultani et al. [24]	84.6
GMM [2]	90.6
VBiLT (Ours)	96.33
MobileBiLT (Ours)	97.96
DenseBiLT (Ours)	96.16
CvtBiLT (Ours)	97.87
CvlBiLT (Ours)	98.56

Additionally, the outcomes of training and validating the proposed architecture with the Hockey Fights, Movie Fights, and Crowd datasets, followed by a comparison of the accuracy (ACC) performance with existing networks, are shown in Table 3.

Table 3. Comparison of result on each dataset.

Method	Hockey Fights	Movie Fights	Crowd
	ACC	ACC	ACC
Semi-Supervised Hard Attention (SSHA (RGB only)) [25]	98.7	100	-
SepConvLSTM-A [29]	99.0	100	-
SepConvLSTM-C [29]	99.5	100	-
SepConvLSTM-M [29]	99.0	100	-
STIFV [14]	93.4	99.0	-
ConvLSTM [30]	97.1	100	-
Human Skeletons + Change Detection [20]	94.5	98.5	94.3
SPIL Convolution [21]	96.8	98.5	94.5
Flow Gated Network [23]	98.0	100	88.87
VBiLT (Ours)	97.67	100	94.0
MobileBiLT (Ours)	96.33	100	93.24
DenseBiLT (Ours)	97.67	100	93.24
CvtBiLT (Ours)	97.0	100	95.95
CvlBiLT (Ours)	98.0	100	95.95

According to the performance evaluation results, all five proposed frameworks demonstrated superior performance on the UBI-Fights dataset compared to existing networks. Additionally, in alignment with the main objective of this study—detecting anomalies in CCTV footage—it was confirmed that CvlBiLT exhibited the most impressive classification performance on the Movie Fights and Crowd datasets. This underscores the capability of the proposed frameworks to accurately distinguish between normal and abnormal situations in both class-balanced and class-imbalanced datasets, thereby proving their effectiveness across various environments and conditions. This emphasizes the effectiveness of the proposed modified Transformer for anomaly detection and the flexibility and accuracy of the proposed three-stage framework for anomaly detection.

In a more detailed examination of the results on the Movie Fights dataset, all the proposed frameworks achieved detection accuracy of 100%. On the Crowd dataset, CvtBiLT and CvlBiLT, which use ConvNeXt as the backbone with high model complexity, performed the best. For the Hockey Fights dataset, CvlBiLT achieved commendable performance of 98%. On the UBI-Fights dataset, which was measured using AUC as the primary metric, MobileBiLT, with the lightest model complexity, outperformed all other networks, followed by CvlBiLT, with the highest model complexity. In summary, CvlBiLT achieved the most superior performance on the three datasets closely related to CCTV footage, UBI-Fights, Movie Fights, and Crowd, demonstrating its effectiveness in environments most relevant to the intended application. Furthermore, it also exhibited the best performance among the proposed architectures in the additional validation dataset, Hockey Fights.

When examining the average anomaly detection accuracy (AUC for UBI-Fight and ACC for the others) across the four datasets for the five proposed frameworks, the results were as follows: VBiLT achieved 97%, MobileBiLT achieved 96.88%, DenseBiLT achieved 96.77%, CvtBiLT achieved 97.71%, and CvlBiLT achieved 98.13%. This shows that CvlBiLT and CvtBiLT, which have high model complexity, perform the best. However, from a model complexity perspective, the VBiLT and MobileBiLT frameworks, which use lightweight models like ViT or MobileNet as backbones, are more effective choices. This indicates that these models can achieve high accuracy with low computational complexity.

These results demonstrate that the proposed three-stage framework is highly useful for image analysis in various conditions and environments. The performance evaluation in this study provides an important benchmark for future research and development.

4.5. Exploring Limitations

Our approach demonstrates excellent performance in anomaly detection in everyday CCTV footage. However, it is important to understand that it faces challenges in videos with rapid motion (significant frame-to-frame differences). The primary objective of our proposed network is anomaly detection in everyday CCTV footage. To validate this objective, we utilized three datasets: UBI-Fights, Movie Fights, and Crowd. Additionally, we included one dataset, Hockey Fights, with a different nature to examine the additional detection accuracy. As evident from Tables 2 and 3, our proposed model outperforms existing networks in three datasets (UBI-Fights, Movie Fights, and Crowd). However, for the supplementary dataset, Hockey Fights, our model did not achieve better performance compared to existing networks. This result arises from the rapid pace inherent in the hockey games captured in the Hockey Fights dataset, where significant inter-frame disparities facilitate effective detection, even in the absence of intricate models. These limitations provide valuable insights for future research and improvements.

4.6. Ablation Study

The key strength of this architecture lies in the potent combination of the BiLSTM, multi-head attention layer, and Transformer encoder. The BiLSTM and multi-head attention layer excels in effectively handling short-term dependencies and intricately capturing the local features of sequences. This enables a deeper understanding of each part of the sequence, which is particularly useful in capturing dynamic elements, such as changes over time.

On the other hand, the Transformer encoder is adept in processing long-term dependencies in sequence data and is exceptional at capturing global patterns across the entire sequence. This is beneficial in understanding the relationships between distant elements in a sequence and analyzing data in a broader context.

When these two layers are used in conjunction, the model can effectively capture both short-term and long-term dependencies within sequence data, effectively integrating them. This approach allows for a more accurate understanding of complex patterns and relationships within the input sequence and prevents an over-reliance on specific patterns, thus avoiding overfitting and enabling the model to make decisions based on finer details. Additionally, this architecture enhances the model's generalization capability, allowing for more accurate predictions across various types of data and scenarios, especially in processing time series data like videos.

To validate our approach, we separately configured and trained the BiLSTM and multi-head attention layer (the second stage) and the modified Transformer encoder layer (the third stage). The results of these ablation training sessions are presented in Tables 4 and 5. Table 4 demonstrates the measurement of the AUC using the UBI-Fights dataset, where we first evaluated all three stages together, followed by configurations of only stages 1 and 3, and then stages 1 and 2. Table 5 assesses the accuracy using the Crowd dataset in a similar manner.

As evidenced in Tables 4 and 5, the performance was most superior when all stages were integrated, with the combination of stages 1 and 2 following closely. This indicates that the temporal insight and multi-dimensional feature capture capabilities of the BiLSTM and multi-head attention, when combined with the pattern analysis prowess of the modified Transformer encoder, enabled a deeper understanding of the sequence data and more accurate predictions. The results of these ablation experiments underscore the exceptional performance of the combined layers, suggesting their utility in various sequence data processing tasks, including video analysis and time series data handling.

Table 4. Comparative analysis of training results for networks configured with different stages using the UBI-Fights dataset.

Network	Whole Architecture	Without Second Stage	Without Third Stage
VBiLT	96.33	95.97	96.06
MobileBiLT	97.96	96.76	97.18

Table 5. Comparative analysis of training results for networks configured with different stages using the Crowd dataset.

Network	Whole Architecture	Without Second Stage	Without Third Stage
VBiLT	94.0	91.89	93.24
MobileBiLT	93.24	89.19	91.89

Additionally, to validate the effectiveness of the proposed Transformer encoder block, we conducted further ablation studies. Experiments were performed on the UBI-Fights dataset by varying the number of heads in the CvlBiLT architecture's Transformer block. The results are presented in Table 6. The best performance was observed with eight heads, and a trend of diminishing performance was noted when the number of heads was either too low or too high. Furthermore, to demonstrate the efficacy of relative position embeddings, we conducted comparative experiments using the best-performing CvlBiLT architecture, applying absolute position embeddings for comparison. The outcomes of these experiments are detailed in Table 7. The results indicated that the application of relative position embeddings outperformed the use of absolute position embeddings.

Table 6. Results of performance measurement according to the variation in the number of heads.

Number of Heads	AUC of CvlBiLT
2	96.76
4	98.24
8	98.56
12	97.78
16	97.73

Table 7. Performance comparison of absolute and relative position embeddings in the CvlBiLT architecture.

Position Embedding	UBI-Fights (AUC)	Crowd (ACC)
Absolute	97.36	94.59
Relative	98.56	95.95

5. Conclusions

In conclusion, this paper presents a three-stage deep learning architecture for anomaly detection in video surveillance, particularly focusing on CCTV footage. The architecture begins with the extraction of features from video frames, utilizing a range of pre-trained networks including Vision Transformer (ViT) and MobileNet. This approach allows for a comprehensive and effective analysis of the video data, leveraging the strengths of these varied networks.

The second stage integrates BiLSTM with multi-head attention, blending the benefits of both techniques for effective temporal analysis. BiLSTM enhances the model's ability to understand temporal sequences by processing data in both the forward and backward directions, while multi-head attention allows the model to concurrently process and integrate multiple facets of the input sequence. This combination is particularly adept at handling complex, time-evolving data in video sequences.

To address the challenges posed by long sequences, the third stage introduces a modified Transformer encoder layer. This layer, utilizing relative position embeddings, is adept at handling the dynamic nature of video data, capturing interactions across frames, and enhancing the understanding of inter-frame dynamics. The Transformer encoder, tailored to anomaly detection, further processes the output from the BiLSTM and multi-head attention layers, refining it for the final classification.

The architecture concludes with a global maxpooling layer to distill key features, a dropout layer to prevent overfitting, and a sigmoid operation for final anomaly classification. The introduction of a fully connected layer further ensures precise classification based on the architecture's output.

In this research, the proposed framework was validated using multiple open datasets containing various anomalous situations. The results demonstrate that it overcomes the limitations of existing video anomaly detection networks, achieving high detection accuracy and showing potential for application in real-time monitoring environments. This system enhances the monitoring accuracy and efficiency of security camera systems like CCTV and is applicable in diverse environments and situations, including scenarios without human presence in the footage. This versatility stems from the framework's ability to interpret features across the entire video and the relationships between frames, making it suitable for various scenarios and conditions.

Recognizing the need for further research in detecting different types of anomalies, future studies will focus on enhancing the scalability of the proposed framework for broader applications in various settings. This research is expected to significantly improve the accuracy and scalability of video analysis in security camera systems like CCTV.

Author Contributions: Conceptualization, J.-W.L.; Methodology, J.-W.L.; Software, J.-W.L.; Validation, J.-W.L. and H.-S.K.; Formal analysis, J.-W.L.; Investigation, H.-S.K.; Resources, H.-S.K.; Data curation, J.-W.L.; Writing—original draft, J.-W.L.; Writing—review & editing, J.-W.L. and H.-S.K.; Visualization, J.-W.L. and H.-S.K.; Supervision, H.-S.K.; Project administration, H.-S.K.; Funding acquisition, H.-S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2024-2020-0-01462) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation), and financially supported by the Ministry of Small and Medium-sized Enterprises (SMEs) and Startups (MSS), Korea, under the "Regional Specialized Industry Development Plus Program (R&D, s3363609)" supervised by the Korea Technology and Information Promotion Agency for SMEs (TIPA).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: UBI-Fights Dataset: http://socia-lab.di.ubi.pt/EventDetection/, accessed on 20 November 2023; Hockey Fight Dataset: https://www.kaggle.com/datasets/yassershrief/hockey-fight-vidoes, accessed on 20 November 2023; Movie Fight Dataset: https://academictorrents.com/details/70e0794e2292fc051a13f05ea6f5b6c16f3d3635, accessed on 20 November 2023; Crowd Dataset: https://talhassner.github.io/home/projects/violentflows/index.html, accessed on 20 November 2023.

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

References

1. Popoola, O.P.; Wang, K. Video-based abnormal human behavior recognition—A review. *IEEE Trans. Syst. Man Cybern. Part (Appl. Rev.)* **2012**, *42*, 865–878. [CrossRef]
2. Degardin, B.M. Weakly and Partially Supervised Learning Frameworks for Anomaly Detection. Ph.D. Thesis, Universidade da Beira Interior (Portugal), Beira, Portugal, 2020.
3. Wang, T.; Qiao, M.; Lin, Z.; Li, C.; Snoussi, H.; Liu, Z.; Choi, C. Generative neural networks for anomaly detection in crowded scenes. *IEEE Trans. Inf. Forensics Secur.* **2018**, *14*, 1390–1399. [CrossRef]
4. Sultani, W.; Chen, C.; Shah, M. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6479–6488.
5. Ravanbakhsh, M.; Nabi, M.; Sangineto, E.; Marcenaro, L.; Regazzoni, C.; Sebe, N. Abnormal event detection in videos using generative adversarial nets. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 1577–1581.
6. Flaborea, A.; Collorone, L.; di Melendugno, G.M.D.; D'Arrigo, S.; Prenkaj, B.; Galasso, F. Multimodal Motion Conditioned Diffusion Model for Skeleton-based Video Anomaly Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Vancouver, BC, Canada, 18–22 June 2023; pp. 10318–10329.
7. Rodrigues, R.; Bhargava, N.; Velmurugan, R.; Chaudhuri, S. Multi-timescale trajectory prediction for abnormal human activity detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 2–5 March 2020; pp. 2626–2634.
8. Flaborea, A.; di Melendugno, G.M.D.; D'arrigo, S.; Sterpa, M.A.; Sampieri, A.; Galasso, F. Contracting Skeletal Kinematic Embeddings for Anomaly Detection. *arXiv* **2023**, arXiv:2301.09489.
9. Şengönül, E.; Samet, R.; Abu Al-Haija, Q.; Alqahtani, A.; Alturki, B.; Alsulami, A.A. An Analysis of Artificial Intelligence Techniques in Surveillance Video Anomaly Detection: A Comprehensive Survey. *Appl. Sci.* **2023**, *13*, 4956. [CrossRef]
10. Kalyta, O.; Barmak, O.; Radiuk, P.; Krak, I. Facial Emotion Recognition for Photo and Video Surveillance Based on Machine Learning and Visual Analytics. *Appl. Sci.* **2023**, *13*, 9890. [CrossRef]
11. Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S.A.; Binder, A.; Müller, E.; Kloft, M. Deep one-class classification. In Proceedings of the International conference on machine learning. *PMLR* **2018**, *80* , 4393–4402.
12. Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning temporal regularity in video sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 733–742.
13. Chong, Y.S.; Tay, Y.H. Abnormal event detection in videos using spatiotemporal autoencoder. In Proceedings of the Advances in Neural Networks-ISNN 2017: 14th International Symposium, ISNN 2017, Sapporo, Hakodate, and Muroran, Hokkaido, Japan, 21–26 June 2017; Proceedings, Part II 14; Springer: Berlin/Heidelberg, Germany, 2017; pp. 189–196.
14. Bilinski, P.; Bremond, F. Human violence recognition and detection in surveillance videos. In Proceedings of the 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Colorado Springs, CO, USA, 23–26 August 2016; pp. 30–36.

15. Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; Brox, T. Flownet 2.0: Evolution of optical flow estimation with deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 2462–2470.
16. Xu, H.; Zhang, J.; Cai, J.; Rezatofighi, H.; Yu, F.; Tao, D.; Geiger, A. Unifying flow, stereo and depth estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13941–13958. [CrossRef] [PubMed]
17. Weinzaepfel, P.; Lucas, T.; Leroy, V.; Cabon, Y.; Arora, V.; Brégier, R.; Csurka, G.; Antsfeld, L.; Chidlovskii, B.; Revaud, J. CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Vancouver, BC, Canada, 18–22 June 2023; pp. 17969–17980.
18. Zhao, S.; Sheng, Y.; Dong, Y.; Chang, E.I.; Xu, Y. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6278–6287.
19. Hachiuma, R.; Sato, F.; Sekii, T. Unified keypoint-based action recognition framework via structured keypoint pooling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 22962–22971.
20. Garcia-Cobo, G.; SanMiguel, J.C. Human skeletons and change detection for efficient violence detection in surveillance videos. *Comput. Vis. Image Underst.* **2023**, *233*, 103739. [CrossRef]
21. Su, Y.; Lin, G.; Zhu, J.; Wu, Q. Human interaction learning on 3d skeleton point clouds for video violence recognition. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part IV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 74–90.
22. Zhu, B.; Hofstee, P.; Lee, J.; Al-Ars, Z. An attention module for convolutional neural networks. In Proceedings of the Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, 14–17 September 2021; Proceedings, Part I 30; Springer: Berlin/Heidelberg, Germany, 2021; pp. 167–178.
23. Cheng, M.; Cai, K.; Li, M. RWF-2000: An open large scale video database for violence detection. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 4183–4190.
24. Degardin, B.; Proença, H. Iterative weak/self-supervised classification framework for abnormal events detection. *Pattern Recognit. Lett.* **2021**, *145*, 50–57. [CrossRef]
25. Mohammadi, H.; Nazerfard, E. Video violence recognition and localization using a semi-supervised hard attention model. *Expert Syst. Appl.* **2023**, *212*, 118791. [CrossRef]
26. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
27. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
28. Abdali, A.M.R.; Al-Tuma, R.F. Robust real-time violence detection in video using cnn and lstm. In Proceedings of the 2019 2nd Scientific Conference of Computer Sciences (SCCS), Baghdad, Iraq, 27–28 March 2019; pp. 104–108.
29. Islam, Z.; Rukonuzzaman, M.; Ahmed, R.; Kabir, M.H.; Farazi, M. Efficient two-stream network for violence detection using separable convolutional lstm. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8.
30. Sudhakaran, S.; Lanz, O. Learning to detect violent videos using convolutional long short-term memory. In Proceedings of the 2017 14th IEEE international Conference on Advanced Video and Signal Based Surveillance (AVSS), Honolulu, HI, USA, 21–26 June 2017; pp. 1–6.
31. Deshpande, K.; Punn, N.S.; Sonbhadra, S.K.; Agarwal, S. Anomaly detection in surveillance videos using transformer based attention model. In Proceedings of the International Conference on Neural Information Processing, Virtual Event, 22–26 November 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 199–211.
32. Jin, P.; Mou, L.; Xia, G.S.; Zhu, X.X. Anomaly detection in aerial videos with transformers. *IEEE Trans. Geosci. Remote. Sens.* **2022**, *60*, 1–13 [CrossRef]
33. Liu, Y.; Yang, D.; Wang, Y.; Liu, J.; Song, L. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *arXiv* **2023**, arXiv:2302.05087.
34. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in nEural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
36. Bonetto, R.; Soldan, M.; Lanaro, A.; Milani, S.; Rossi, M. Seq2Seq RNN based gait anomaly detection from smartphone acquired multimodal motion data. *arXiv* **2019**, arXiv:1911.08608.
37. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
38. Kandoor, A. Tiny Neural Models for Seq2Seq. *arXiv* **2021**, arXiv:2108.03340.
39. Kong, L.; Alberti, C.; Andor, D.; Bogatyy, I.; Weiss, D. Dragnn: A transition-based framework for dynamically connected neural networks. *arXiv* **2017**, arXiv:1703.04474.

40. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
41. Mukherjee, S.; Saini, R.; Kumar, P.; Roy, P.P.; Dogra, D.P.; Kim, B.G. Fight detection in hockey videos using deep network. *J. Multimed. Inf. Syst.* **2017**, *4*, 225–232.
42. Nievas, E.B.; Suarez, O.D.; Garcia, G.B.; Sukthankar, R. Movies Fight Detection Dataset. In Proceedings of the Computer Analysis of Images and Patterns, Seville, Spain, 29–31 August 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 332–339.
43. Hassner, T.; Itcher, Y.; Kliper-Gross, O. Violent flows: Real-time detection of violent crowd behavior. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 1–6.
44. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
45. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
46. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.