*future internet*

MDPI

*Article*

# Edge-Enhanced TempoFuseNet: A Two-Stream Framework for Intelligent Multiclass Video Anomaly Recognition in 5G and IoT Environments

**Gulshan Saleem [1], Usama Ijaz Bajwa [1,*], Rana Hammad Raza [2] and Fan Zhang [3,*]**

[1] Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Lahore 54000, Pakistan; gulshnsaleem26@gmail.com
[2] Electronics and Power Engineering Department, Pakistan Navy Engineering College (PNEC), National University of Sciences and Technology (NUST), Karachi 75350, Pakistan; hammad@pnec.nust.edu.pk
[3] Ocean College, Zhejiang University, Hangzhou 316000, China
* Correspondence: usamabajwa@cuilahore.edu.pk (U.I.B.); f.zhang@zju.edu.cn (F.Z.)

**Abstract:** Surveillance video analytics encounters unprecedented challenges in 5G and IoT environments, including complex intra-class variations, short-term and long-term temporal dynamics, and variable video quality. This study introduces Edge-Enhanced TempoFuseNet, a cutting-edge framework that strategically reduces spatial resolution to allow the processing of low-resolution images. A dual upscaling methodology based on bicubic interpolation and an encoder–bank–decoder configuration is used for anomaly classification. The two-stream architecture combines the power of a pre-trained Convolutional Neural Network (CNN) for spatial feature extraction from RGB imagery in the spatial stream, while the temporal stream focuses on learning short-term temporal characteristics, reducing the computational burden of optical flow. To analyze long-term temporal patterns, the extracted features from both streams are combined and routed through a Gated Recurrent Unit (GRU) layer. The proposed framework (TempoFuseNet) outperforms the encoder–bank–decoder model in terms of performance metrics, achieving a multiclass macro average accuracy of 92.28%, an F1-score of 69.29%, and a false positive rate of 4.41%. This study presents a significant advancement in the field of video anomaly recognition and provides a comprehensive solution to the complex challenges posed by real-world surveillance scenarios in the context of 5G and IoT.

**Keywords:** edge intelligence; anomaly identification; super resolution; video classification; two-stream architecture; StyleGAN; IoT environment



**Citation:** Saleem, G.; Bajwa, U.I.; Raza, R.H.; Zhang, F. Edge-Enhanced TempoFuseNet: A Two-Stream Framework for Intelligent Multiclass Video Anomaly Recognition in 5G and IoT Environments. *Future Internet* **2024**, *16*, 83. https://doi.org/10.3390/fi16030083

Academic Editors: Yuezhi Zhou and Xu Chen

Received: 5 January 2024
Revised: 4 February 2024
Accepted: 27 February 2024
Published: 29 February 2024



## 1. Introduction

In the era of 5G and IoT, video surveillance is a critical component of modern security and monitoring strategies. This surveillance relies on advanced camera technology to observe and analyze diverse environments and contributes to applications such as security, crime prevention, safety, emergency response, traffic monitoring, and behavior analysis [1–3]. Video surveillance contributes significantly to theft prevention, traffic management, and overall safety in the residential, commercial, and industrial sectors.

The incorporation of technology and machine learning [4,5] into video surveillance, particularly in 5G and IoT environments, initiates unprecedented possibilities. Automated video surveillance systems controlled by computer vision algorithms [6–8] detect anomalies, changes in motion, and intrusions in real-time, reducing reliance on human monitoring [9]. However, challenges persist, such as operator errors, false alarms, and limitations in contextual information within video footage [10–12].

In the context of 5G and IoT, this study addresses technical limitations associated with low-quality videos: specifically, poor lighting and low spatial resolution. These difficulties have an impact on the perceptual quality of video streams [13–15] and introduce factors

such as poor lighting, camera noise, low spatial resolution, and low frame rates [9,16–19]. Despite these challenges, various techniques for detecting anomalies in low-quality surveillance videos have been proposed, [20,21]. Two primary approaches are commonly used to address the challenge of detecting anomalies in low-quality videos. The first entails improving video quality with techniques like denoising, dehazing, and super-resolution [22,23]. An alternative strategy is to use deep learning methods directly for anomaly detection in low-quality videos [24,25].

This study outperforms existing approaches by introducing a new super-resolution technique called "TempoFuseNet". For enhanced anomaly detection, this innovative framework employs a two-stream architecture that combines spatial and temporal features. The spatial stream extracts features using a pre-trained Convolutional Neural Network (CNN), whereas the temporal stream captures short-term temporal characteristics efficiently using a novel Stacked Grayscale 3-channel Image (SG3I) approach. The extracted features from both streams are fused via a Gated Recurrent Unit (GRU) layer to leverage long-term temporal dependencies effectively. The contributions of this study include the identification of challenges related to intra-class and inter-class variabilities, the introduction of a super-resolution technique leveraging an encoder–bank–decoder configuration, the incorporation of a StyleGAN for feature enhancement, and the proposal of a two-stream architecture for anomaly classification.

Recognizing the nuanced landscape of automated surveillance systems is essential in the continuum of addressing challenges in video surveillance. These systems play a critical role in overcoming the limitations of manual monitoring. Despite their potential, however, these systems face challenges that necessitate strategic interventions for further refinement. One significant challenge is the generation of false alarms, which can overwhelm security personnel and undermine the effectiveness of surveillance operations. False alarms not only divert attention but also place unnecessary demands on resources. The importance of minimizing false alarms as a fundamental aspect of optimizing automated surveillance systems is acknowledged in this study.

Another problem stems from video's inherent limitation in providing comprehensive context. Surveillance videos frequently capture snippets of events, making it difficult to decipher the intentions of those being watched or comprehend the full scope of a given incident. Improving the contextual understanding of surveillance footage appears to be a critical component in addressing this challenge. Technical constraints obstruct the seamless operation of automated surveillance systems. Poor lighting, low-resolution cameras, and limited storage capacity can all have an impact on the effectiveness of these systems. To improve the robustness and reliability of automated surveillance, a comprehensive approach to addressing these technical limitations is required.

This study focuses on the technical limitations caused by low-quality videos: specifically, poor lighting and low spatial resolution. These difficulties have been identified as critical factors influencing the perceptual quality of video streams and thereby influencing the accuracy of anomaly detection systems according to [16]. The contributions of this study can be summed up as follows:

- This study meticulously identifies and articulates two critical issues inherent in surveillance videos: high intra-class variability and low inter-class variability. These challenges, which are inextricably linked to the temporal properties of video streams, both short- and long-term, are exacerbated by the prevalence of low-quality videos.
- This study makes an outstanding contribution by introducing an innovative super-resolution approach designed to mitigate the impact of low-quality videos caused by downscaling. This approach outperforms traditional bicubic interpolation by using an encoder–bank–decoder configuration to upscale videos. The primary goal is to improve the spatial resolution of videos in order to increase the accuracy of anomaly detection. The addition of a pre-trained StyleGAN as a latent feature bank is a critical step forward that enriches the super-resolution process and, as a result, improves anomaly classification accuracy.

- The study implies a two-stream architecture for anomaly classification. The spatial stream uses a pre-trained CNN model for feature extraction, whereas the temporal stream employs an innovative approach known as Stacked Grayscale Image (SG3I). SG3I effectively lowers the computational costs associated with optical flow computation while accurately capturing short-term temporal characteristics. The extracted features from both streams are concatenated and fed into a Gated Recurrent Unit (GRU) layer, which allows the model to learn and exploit long-term dependencies.
- Experiments show that the super-resolution model improves classification accuracy by 3.7% when compared to traditional bicubic interpolation methods. When combined with the encoder–bank–decoder super-resolution model, the classification model achieves an impressive accuracy of 92.28%, an F1-score of 69.29%, and a low false positive rate of 4.41%.

To sum up, this research not only identifies and understands the difficulties that are associated with surveillance footage, but it also introduces novel approaches to deal with those difficulties. The end result of these efforts is observable improvements in performance and accuracy for the classification of anomalies in surveillance videos.

## 2. Related Work

Video anomaly detection is critical in the domain of surveillance systems, as it addresses the need to identify anomalous segments within video streams. Over time, two major streams of methodologies have emerged for this critical task: handcrafted approaches and deep-learning-based methods. The former employs manual feature engineering techniques such as STIP, SIFT-3D, and optical flow histograms, whereas the latter makes use of the power of deep neural networks such as VGG and ResNet to process spatiotemporal data efficiently. The introduction of two-stream Convolutional Neural Networks (CNNs) for improved activity recognition and novel approaches to modeling long-term temporal dependencies are notable advancements. The literature includes a wide range of deep learning models, from ConvLSTM to attention-based architectures, all of which contribute to the improvement of anomaly detection in videos. Furthermore, weakly supervised techniques, generative models, and recent efforts to address anomalies in low-resolution videos have significantly expanded the scope of this evolving field. In the midst of these advances, our research focuses on a novel problem: detecting anomalies in multi-class scenarios in low-quality surveillance videos. We present a unified methodology that combines novel super-resolution techniques with a two-stream architecture, providing a comprehensive solution to the complexities of real-world surveillance scenarios.

Manual feature engineering methods such as STIP, SIFT-3D, and optical flow histograms involve human intervention [26,27]. While insightful, improved dense trajectory approaches like the one by Wang et al. [28] surpass earlier handcrafted techniques. The advent of deep learning has revolutionized video anomaly identification, with networks like VGG and ResNet efficiently processing spatiotemporal data in videos [29,30]. Noteworthy in this domain is the introduction of two-stream Convolutional Neural Networks (CNNs), which combine spatial and temporal inputs for improved activity recognition [31,32].

Advancements in modeling long-term temporal dependencies have been achieved through techniques like temporal segment networks and 3D convolutional filters. Wang et al. [33] introduced a temporal segment network that exhibited robust performance on benchmark datasets. The C3D method by Tran et al. [34] addressed challenges in modeling temporal information and inspired subsequent work by Maqsood et al. [35] for anomaly classification. Among deep-learning-based models, significant strides have been made, particularly in domains involving nonlinear, high-dimensional data. Luo et al. [36] proposed a Convolutional Long Short-Term Memory (ConvLSTM) model for encoding video frames and identifying anomalies. Ullah et al. [37] introduced a Convolution-Block-Attention-based LSTM model that enhances spatial information accuracy. Riaz et al. [38] combined human posture estimation with a densely connected fully Convolutional Neural Network (CNN) for anomaly identification. Hasan et al. [1] utilized a recurrent neural net-

work (RNN) and a convolutional autoencoder for anomaly detection, while Liu et al. [39] integrated temporal and spatial detectors for anomaly identification.

Weakly supervised techniques, including C3D and MIL, have been employed for anomaly detection. Sultani et al. [2] combined weak video labels with Multiple Instance Learning (MIL). Landi et al. [40] used a coordinate-based regression model for tube extraction. Generative models like GANs have been explored, with Sabokroul et al. [41] training GANs for visual anomaly detection. BatchNorm into Weakly Supervised Video Anomaly Detection (BN-WVAD) [42] has been used to capitalize on the statistical insight that temporal features of abnormal events often behave as outliers; BN-WVAD leverages the Divergence of Feature from Mean vector (DFM) function from BatchNorm. This DFM criterion serves as a robust abnormality indicator and identifies potential abnormal snippets in videos. It enhances anomaly recognition, proves to be more resistant to label noise, and provides an additional anomaly score to refine predictions from classifiers that are sensitive to noisy labels. In [43], a Temporal Context Aggregation (TCA) module for efficient context modeling and a Prompt-Enhanced Learning (PEL) module for enhanced semantic discriminability are demonstrated. The TCA module captures complete contextual information, while the PEL module incorporates semantic priors using knowledge-based prompts to improve discriminative capacity and maintain separability between anomaly sub-classes. Additionally, a Score Smoothing (SS) module is introduced in the testing phase to reduce false alarms. In [44], a U-Net-like structure is implemented to effectively capture both local and global temporal dependencies in a unified manner. The encoder hierarchically learns global dependencies on top of local ones, and the decoder propagates this global information back to the segment level for classification.

Recent research has focused on addressing anomalies in extremely low-resolution videos [25,45–47]. Techniques such as Inverse Super-Resolution (ISR), initially introduced by Ryoo et al. [45], aim to identify optimal image modifications for extracting additional information from low-resolution images. Additionally, multi-Siamese loss functions have been proposed to maximize data utilization from a collection of low-resolution images. Chen et al. [46] developed a semi-coupled two-stream network that leverages high-resolution images to assist with training a low-resolution network. Xu et al. [47] demonstrated that using high-resolution images improves low-resolution recognition by incorporating a two-stream neural network architecture that takes high-resolution images as inputs. Their approach, sharing convolutional filters between low- and high-resolution networks, significantly enhanced performance. In addition, Demir et al. [48] proposed the TinyVIRAT dataset for natural low-resolution videos and presented a gradual generative technique for enhancing the quality of low-resolution events. Super-resolution techniques have also found success in various applications such as low-resolution face verification, small object detection, person re-identification, and activity recognition [49–52]. For instance, Ataer et al. [50] introduced an identity-preserving super-resolution approach for face verification at very low resolutions, and Bai et al. [51] developed a multitask generative adversarial network for small object detection.

In summary, the field of video anomaly detection has witnessed diverse advancements, from Bayesian deep learning to convolutional models, recurrent neural networks, and spatial–temporal graph attention networks. Our study addresses the challenge of detecting anomalies in multi-class scenarios within low-quality surveillance videos and showcases improved classification performance compared to interpolation-based strategies. The integration of novel super-resolution techniques and a two-stream architecture forms the backbone of our methodology and contributes to the evolution of video anomaly recognition in complex real-world scenarios. While the literature review reflects significant progress in video anomaly detection, there is a significant research gap that our study seeks to fill. Existing approaches have primarily focused on either high-quality video scenarios or have addressed anomalies in a binary manner, both of which are insufficient for real-world applications. The combination of novel super-resolution techniques and a two-stream architecture, as proposed in our methodology, represents a novel approach to closing this

gap. Our research contributes to the evolving landscape of video anomaly recognition by providing a tailored solution to the complexities of multi-class scenarios and low-quality surveillance videos within 5G and IoT environments.

## 3. Materials and Methods

The effectiveness of anomaly detection in surveillance videos is inextricably linked to the quality of the input data. In this methodology, we address the challenges posed by low-quality surveillance videos; we focus on issues such as poor lighting and spatial resolution. Our method combines advanced video resizing techniques with deep-learning-based super-resolution methods to improve the overall quality of video streams. The initial stages of our methodology include a meticulous video resizing process in which we experiment with various interpolation methods to upscale low-resolution videos. We then present a novel video super-resolution strategy that takes advantage of GLEAN, a framework that uses Generative Adversarial Networks (GANs) for latent feature extraction. Unlike traditional GAN-based models, our implementation uses a streamlined process that requires only one forward pass to generate high-resolution images. The use of a StyleGAN, which has been fine-tuned on a dataset containing both low- and high-resolution representations of surveillance video frames, is critical to our super-resolution strategy. This pre-trained StyleGAN acts as a latent feature bank by providing rich priors for creating realistic, high-quality, high-resolution videos. The proposed framework "TempoFuseNet" is presented in Figure 1, and the specifics of all stages are discussed, including the dataset, pseudocode for the super-resolution algorithm, and an explanation of our two-stream architecture for anomaly classification. The goal is not only to improve the spatial resolution of surveillance videos but also to provide a solid framework for accurately detecting anomalies in challenging real-world scenarios within 5G and IoT environments.

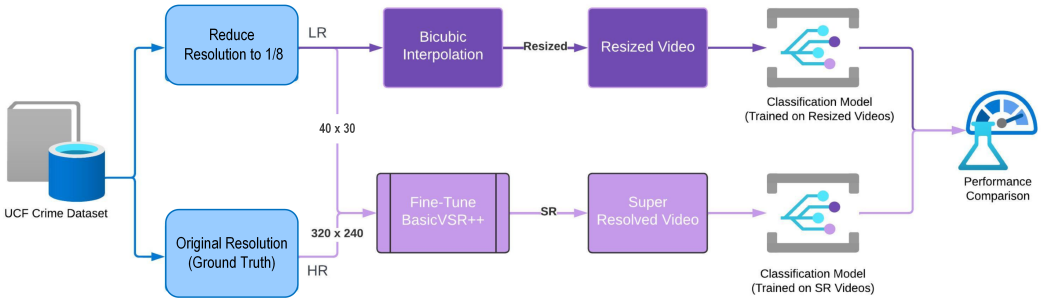


**Figure 1.** Proposed framework (TempoFuseNet) for anomaly classification for low-resolution videos.

### *3.1. Dataset*

In order to perform classification learning to classify surveillance videos into one of several classes of anomalies, a labeled dataset of videos is required. Various datasets are used by the research community to demonstrate anomaly detection in surveillance videos, and each of these datasets has its own characteristics [2,39,53,54]. This study is based on the UCF-Crime dataset [2], which is modified to make it more useful for the demonstration of anomaly classification for low-quality surveillance videos.

There are 128 hours of surveillance footage in the original UCF-Crime dataset. The dataset includes 1900 complete and unfiltered surveillance videos from the real world, along with thirteen actual anomalies such as assault, arrest, abuse, arson, burglary, fighting, shooting, explosion, road accident, vandalism, robbery, and shoplifting. These anomalies were included in the dataset due to their possible impact on the safety of the general public. We meticulously curated the dataset to address class imbalance by retaining a standardized set of 50 videos per class to ensure the relevance and practicality of our study. Because of this deliberate selection process, classes with insufficient representation were excluded, resulting in a focused dataset with eight distinct categories: assault, arrest, abuse, arson, burglary, fighting, explosion, and normal. This strategic enhancement to the UCF-Crime

dataset ensures a balanced and representative collection, which improves the precision and applicability of our experimental results. All videos in each class have the same spatial resolution of 320 × 240 pixels, which contributes to the consistency and reliability of our analytical framework.

Data Preparation

In order to perform learning on low-quality videos, the original videos are downsampled by eight times to obtain a low-resolution version of the original videos. The video resolution after downsampling is 40 × 30 pixels . Downsampling is performed by using bilinear interpolation (refer to Equation (1)), in which the target image pixels are obtained by performing linear interpolation in both the horizontal and vertical directions.

Let $LR(x', y')$ be the low-resolution pixel values at coordinates $(x', y')$, and let $HR(x, y)$ be the high-resolution pixel values at coordinates $(x, y)$. The downsampling operation can be expressed as:

$$LR(x', y') = f(HR(x, y)) \tag{1}$$

where

$$x' = [\, x/8], \quad y' = [\, y/8] \tag{2}$$

This stage results in two sets of data: one containing high-resolution (HR) videos that are the ground truth data, and the other has low-resolution (LR) videos, which are a downsampled version of the data and will be used for classification modeling.

### *3.2. Video Upscaling*

Video resizing is the most commonly used operation to change the resolution of a video to match the requirements of the input layer of a convolutional neural network. There are various algorithms that can be used to perform the operation of video upscaling, and the most common are nearest neighbor interpolation, bilinear interpolation, bicubic interpolation, and Lanczos interpolation [55]. Among these methods, nearest neighbor is the fastest, and Lanczos is the slowest and most complex. Their upscaling performance is similarly related, but we used bicubic interpolation in our implementation due to its acceptable performance in terms of speed and upscaling quality. In order to perform bicubic interpolation for video scaling, we the used Libswscale library from the FFmpeg 4.2.1 package. The Libswscale library, which is developed in C and is part of the FFmpeg multimedia framework, includes highly optimized functions for scaling, colorspace conversion, and pixel format transformations.

### *3.3. Video Super Resolution*

To obtain high-resolution videos, this study use a deep-learning-based video super-resolution approach as an effective strategy for overcoming technical limitations associated with low-quality videos: particularly, poor lighting and low spatial resolution. The proposed implementation employs GLEAN [56]: a framework that uses a Generative Adversarial Network (GAN) as a latent bank to extract rich and diverse priors from a pre-trained GAN model. Unlike traditional GAN-based methods, which involve adversarial loss and costly optimization through GAN inversion, our approach uses a single forward pass to generate high-resolution images.

To overcome poor lighting and low spatial resolution, a StyleGAN [57] is used in our implementation. The StyleGAN, fine-tuned on a dataset of surveillance videos with both low- and high-resolution representations of each frame, serves as a pre-trained latent feature bank. This latent feature bank functions similarly to a dictionary, but its distinct advantage is its nearly infinite feature bank, which provides superior priors for generating realistic high-resolution videos. Furthermore, our encoder–bank–decoder formulation, illustrated in Figures 2 and 3, is crucial for obtaining super-resolution images. Notably, the encoder accepts an input resolution of 40 × 30 pixels and outputs 320 × 240 pixels, demonstrating its ability to handle low-spatial-resolution scenarios. The latent feature bank,

which is powered by the pre-trained StyleGAN, ensures that the generated high-resolution videos retain realism and fidelity even in challenging lighting conditions.

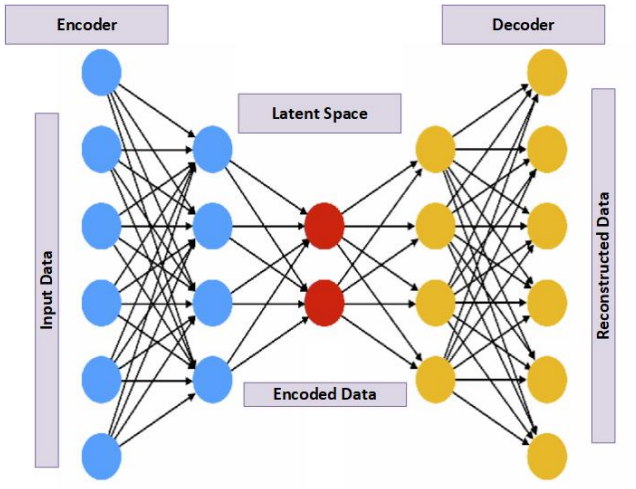


**Figure 2.** Encoder–bank–decoder representation [58].

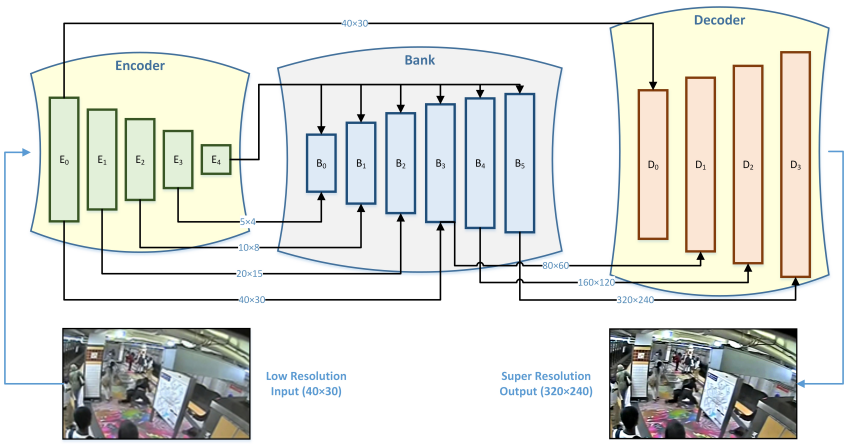


**Figure 3.** Video super-resolution framework (encoder–bank–decoder) based on pre-trained StyleGAN.

In order to obtain high-resolution videos apart from interpolation-based upscaling techniques, deep-learning-based video super resolution is an attractive approach. Generative Adversarial Networks (GANs) built using neural networks have shown excellent performance in video generation, enhancement, and super resolution, among other tasks. GLEAN [56] is an approach that uses a GAN-based model as a latent bank to obtain rich and diverse priors from pre-trained GAN. Unlike existing GAN-based approaches that generate realistic outputs through adversarial loss and the use of expensive optimization through GAN inversion, this approach uses a single forward pass to generate a high-resolution image. In this implementation, we used a StyleGAN [57] and fine-tuned it on a dataset of surveillance videos containing low-resolution and high-resolution representations of each frame.

Super-resolution images are obtained from low-resolution images by using an encoder–bank–decoder formulation. The latent features bank acts like a dictionary as in traditional approaches but differs in the sense that dictionaries contain a finite feature bank, whereas a GAN contains a practically infinite feature bank, making it a superior prior. The architecture of encoder–bank–decoder used in this implementation is provided in Figure 3. Note that the encoder accepts an input resolution of $40 \times 30$ pixels and provides an output of $320 \times 240$ pixels. The bank is a pre-trained StyleGAN that acts as a latent feature bank to provide realistic high-resolution videos.

### *3.4. Upscaling Performance*

As discussed, there are various interpolation-based approaches that can be used for frame-by-frame video upscaling. The performances of four interpolation-based upscaling approaches along with the ground truth and the super-resolution image obtained by our implementation of GLEAN are provided in Figure 4. The results are provided for a single frame from a surveillance video belonging to the "fighting" class. Nearest neighbor and bilinear interpolation are the simplest and fastest methods to upscale an image but create the lowest-quality results. The difference between them is that bilinear interpolation provides a smoother image by using blurring, whereas nearest neighbor provides a boxing effect, and the choice of which one to use mainly depends on the intended application. Lanczos and bicubic are the next level of quality for upscaling and involve greater computational complexity. Lanczos has better detail preservation and a sharper appearance, while bicubic interpolation provides a smoother appearance. Because the targeted scenario for super resolution involves videos with a large number of frames, we use bicubic interpolation (Equation (3)) to upscale the video, which allows for a thorough comparison to the super-resolution videos.

$$I'(x', y') = \sum_{i=-1}^{2} \sum_{j=-1}^{2} I(x+i, y+j) \cdot K(x'-x+i) \cdot K(y'-y+j) \tag{3}$$

A super-resolution image produced using the modified GLEAN model is of much higher quality compared to its counterparts in terms of preservation of details and reconstruction of the structure. The modified GLEAN model includes improved architectural features and training strategies that allow for more effective detail preservation during the upscaling process. This entails a more sophisticated latent space representation or a fine-tuned generator network, which allow the model to capture and reproduce intricate details in the low-resolution input. The modified GLEAN model's superior quality of super-resolution images results from its advanced architecture and training strategies, which enable effective preservation of details and accurate reconstruction of complex structures when compared to other methods. To extract $f_i$ features (Equations (4) and (5)) from a low-resolution image, we employed $E_i$ sequence operations followed by Convolutional layers and fully connected layers to generate a matrix C of representative features.

$$f_i = E_i(f_{i-1}), \text{ for } i \in \{1, \ldots, N\} \tag{4}$$

$$C = E_{N+1}(f_N) \tag{5}$$

Moreover, it is evident from Figure 4 that a super-resolution image has higher overall contrast as compared to the ground truth image, which is due to the use of the latent feature bank containing a pre-trained StyleGAN.

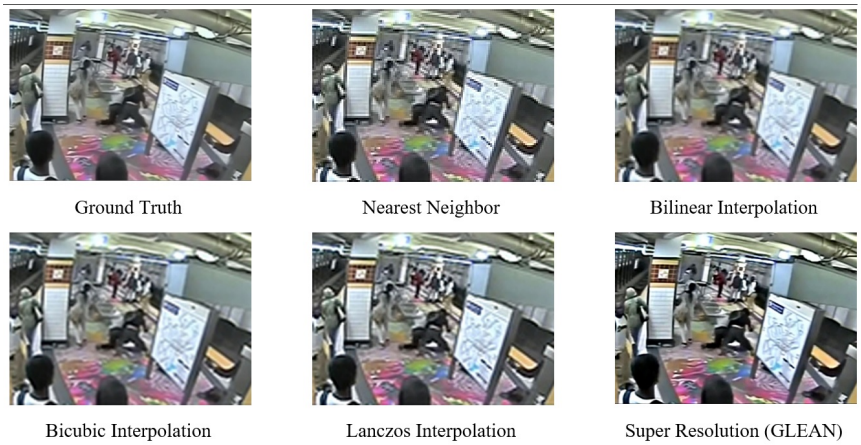


**Figure 4.** Video frame upscaling results for fighting scene.

Algorithms 1 and 2 are simplified pseudocode of the proposed "TempoFuseNet", with a focus on video super resolution using the modified GLEAN framework and anomaly classification using the two-stream architecture. This pseudocode is intended to provide an algorithmic and high-level representation of anomaly classification for real-world scenarios in 5G and IoT environments.

**Algorithm 1:** Video Super Resolution with GLEAN

```
Data: low_resolution_video
Result: super_resolved_video
GLEAN_model = initialize_GLEAN_with_StyleGAN();
foreach frame in low_resolution_video do
    super_resolved_frame = GLEAN_model.forward_pass(frame);
    Save super_resolved_frame;
return super_resolved_video;
```

**Algorithm 2:** Anomaly Classification with Two-Stream Architecture

```
Data: video_frames
Result: final_classification
for i to len(video_frames) 3 do
    spatial_features = extract_spatial_features(video_frames[i]);
    spatial_predictions = ResNet50_predict(spatial_features);
    Save spatial_predictions;
for frame to video_frames do
    SG3I_frame = convert_RGB_to_SG3I(frame);
    temporal_features = ResNet50_predict(SG3I_frame);
    Save temporal_features;
concatenated_features = concatenate(spatial_predictions, temporal_features);
temporal_model_output = apply_GRU(concatenated_features);
final_classification = Dense(temporal_model_output);
return final_classification;
```

### 3.5. Anomaly Classification

To perform anomaly classification, we used a two-stream architecture. Contrary to existing approaches that rely on the optical flow for one stream and the RGB image for the other stream, we used a simple but effective strategy that eliminates the need for expensive optical flow computation. The proposed two-stream architecture is depicted in Figure 5, whereas the details of both the spatial and temporal streams as well as late temporal modeling are provided later in this section.

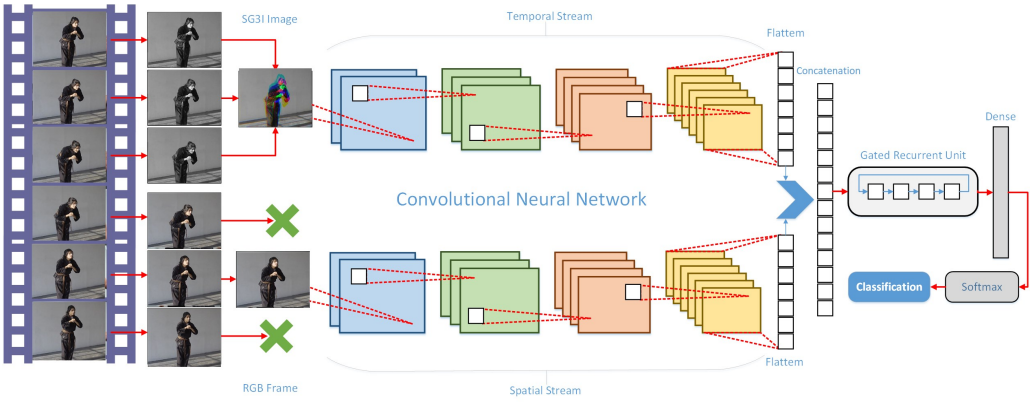


**Figure 5.** Proposed two-stream architecture for spatiotemporal learning for anomaly classification.

#### 3.5.1. Spatial Stream

The spatial stream consists of a pre-trained CNN model base with its classification and dense layers are removed. The network performs prediction on an individual frame basis, and every third video frame is provided to the CNN model to match the predicted computational performance of the temporal stream. The spatial stream uses a ResNet50 model that effectively acts as a feature extractor from the RGB images obtained from the video stream.

#### 3.5.2. Temporal Stream

In order to perform temporal learning without incurring a high computational load, we made use of Stacked Grayscale 3-channel Image (SG3I) (Equation (6)) [59].

$$SG3I(x, y) = R(x, y), G(x, y), B(x, y) \tag{6}$$

Here, SG3I(x, y) represents the function notation for the SG3I value at pixel coordinates $(x, y)$, and $R(x, y), G(x, y), B(x, y)$ represents the intensity value at the same pixel coordinates. SG3I relies on the simple idea of combining multiple frames of video into single frame. The objective is achieved by converting the RGB frames into grayscale images. These grayscale images are combined to form a single 3-channel RGB image, and then, combining the three grayscale images forms the SG3I image, which acts like a single RGB video frame. The frames are selected in sequential order, and each subsequent grayscale frame is fitted to the red, green, and blue channels to yield a single RGB image. This new image is fed to the same pre-trained CNN model as the spatial stream, which serve two purposes: the SG3I image preserves the short temporal characteristics, and the grayscale conversion lets the model focus more on motion-related features.

#### 3.5.3. Late Temporal Modeling

The features extracted from both the spatial and temporal streams are flattened and concatenated to perform feature fusion. In order to learn the long-term temporal characteristics of a video, late temporal modeling is performed from a concatenated feature set. Long Short-Term Memory (LSTM), bidirectional-LSTM, and Gated Recurrent Units (GRUs) are the three modeling approaches that are experimented with, and it is observed that GRUs provide the best temporal modeling characteristics, with a slight margin over LSTMs and bi-LSTMs. A possible explanation for the better performance of GRUs over LSTMs is the smaller size of the training dataset necessary to train a GRU. The GRU is followed by a dense layer and classification of the video into one of eight classes.

## 4. Experiments

### *4.1. Experimental Setup*

Our experimental setup is intended to address the challenges posed by poor lighting and low spatial resolution in order to comprehensively assess the performance of anomaly detection in low-quality surveillance videos. The trimmed UCF-Crime dataset, which includes eight anomaly identification classes, is used in two different types of experiments.

The Trimmed UCF-Crime dataset has original video dimensions of $320 \times 240$. To simulate real-world scenarios with poor lighting and low spatial resolution, we intentionally reduced the spatial resolution by a factor of eight, resulting in low-resolution videos with dimensions of $40 \times 30$. It is important to note that this intentional downsampling is only for experimental purposes and is not a component of the proposed anomaly detection system. Moreover, we recognize that the term "low quality" can be broad: our research focuses on a specific aspect, low spatial resolution, to evaluate the robustness of our proposed methodology under these conditions.

In the first experiment, the downscaled dataset is upscaled to its original spatial resolution using bicubic interpolation. This experiment allows us to assess the performance of our proposed anomaly classifier under standard upscaling conditions and serves as the

baseline for comparison. In the second experiment, we use an advanced GLEAN-based model for super resolution. This model upscales low-resolution videos, resulting in super-resolved video frames. These frames are then used for classification modeling with our proposed anomaly classifier. This experiment addresses the issue of low spatial resolution by employing advanced super-resolution techniques. The experiments are carried out in TensorFlow with the Keras 2.4.0 backend on a Windows 10 machine. Table 1 shows the detailed system setup used for the experiments. During the model's training phase, a random search is used to select specific hyperparameters. To ensure the best model performance, we use the Adam optimizer with a piecewise learning rate. If no progress is seen after minimizing the learning rate for three consecutive validation checks, the training is stopped. Table 2 summarizes the hyperparameter tuning process and offers insights into optimizing different parameters for the best results.

**Table 1.** System specifications.

| # | Type | Specifications |
|---|---|---|
| 1 | System | Dell Precision T5600 |
| 2 | CPU | 2× Intel® Xeon® Processor E5-2687W |
| 3 | RAM | 32GB DDR3 |
| 4 | GPU | GeForce RTX 2070 |
| 5 | GPU Memory | 8GB GDDR6 |
| 6 | CUDA Cores | 2304 |
| 7 | Storage | 512GB SSD |

This experimental setup was designed to simulate and effectively address the technical limitations associated with poor lighting and low spatial resolution in real-world surveillance scenarios. This was to ensure a thorough evaluation of TempoFuseNet, our proposed anomaly identification framework.

**Table 2.** Training parameters used to train the model.

| # | Training Parameter | Value |
|---|---|---|
| 1 | Optimizer | Adam |
| 2 | Initial Learning Rate | 0.003 |
| 3 | Learning Rate Schedule | Piecewise |
| 4 | Learning Rate Drop Factor | 0.5 |
| 5 | Gradient Decay Factor | 0.9 |
| 6 | L2 Regularization | 0.0001 |
| 7 | Max Epochs | 100 |
| 8 | Mini Batch Size | 32 |
| 9 | Loss Function | Categorical cross-entropy |
| 10 | Validation Check | Every epoch |

### *4.2. Evaluation Method and Metrics*

Model evaluation is a way to assess the skill of a prediction model, which is a classifier in our case. The model is trained and evaluated using holdout validation in which the data are partitioned into an 80:20 ratio with 80% of the data being used for model training and validation and 20% holdout data being used for model testing. The performance of both experiments is reported for the same train–test split of the data to make a fair comparison. Evaluation of the model's performance is made based on various performance metrics

obtained from the confusion matrix. The comparative performance for both models is also provided to assess the overall anomaly identification performance.

## 5. Results

Anomaly identification in surveillance videos is a difficult task, especially when using low-quality videos with poor spatial resolution and visual characteristics. Traditional methods, such as spatial interpolation, frequently result in limited improvement and can introduce undesirable artifacts. Alternatively, video super resolution, which improves spatial resolution, can be computationally expensive. This study addresses the challenges of low-quality videos using a video super-resolution approach based on StyleGAN priors. The StyleGAN improves not only the spatial resolution but also image sharpness and contrast. Unlike traditional video super-resolution methods, our approach selectively super-resolves frames that are relevant to anomaly identification, thereby improving computational efficiency. A two-stream architecture is used for classification modeling, which reduces the need for expensive optical flow computation. The RGB stream promotes spatial learning, whereas the SG3I stream emphasizes short-term temporal learning. Both streams use the same pre-trained CNN architecture, which has been fine-tuned for the dataset of interest. The learned features are concatenated and fed into a Gated Recurrent Unit (GRU) for long-term temporal modeling. The proposed approach effectively addresses the challenges posed by low-quality surveillance videos and delivers superior anomaly classification performance while minimizing the computational burden.

### *5.1. Classification Performance*

#### 5.1.1. Bicubic Interpolation of Videos

The classification performance of the upscaled images using bicubic interpolation is provided in the confusion matrix in Figure 6. It is to be noted that out of 50 videos in each class, 40 videos are used for model training, and 10 videos are used for model testing. The confusion matrix provides the actual number of videos classified into each category. Table 3 provides the performance metrics for each class as well as the macro-averaged value for all classes. Classification accuracy is usually regarded as the most important performance metric for anomaly classification, followed by the FPR. Moreover, the values of precision, recall (sensitivity), F1-score, specificity, FPR, and FNR are also reported for each class and are averaged for all classes.

#### 5.1.2. Super-Resolution Videos

Like for the bicubicly interpolated videos, the classification performance for super-resolution videos is provided in the confusion matrix in Figure 7. The confusion matrix reports the classification performance based on 10 videos per anomaly class. Table 4 provides the performance metrics for each class as well as the macro-averaged value for all classes.

| | Abuse | Arrest | Arson | Assault | Burglary | Explosion | Fight | Normal |
|---|---|---|---|---|---|---|---|---|
| Abuse | 5 | 0 | 0 | 3 | 0 | 0 | 2 | 0 |
| Arrest | 0 | 5 | 0 | 0 | 0 | 0 | 3 | 2 |
| Arson | 0 | 0 | 5 | 0 | 0 | 4 | 0 | 2 |
| Assault | 3 | 0 | 0 | 5 | 0 | 0 | 2 | 0 |
| Burglary | 1 | 0 | 0 | 1 | 5 | 0 | 0 | 3 |
| Explosion | 0 | 0 | 3 | 0 | 0 | 6 | 0 | 1 |
| Fight | 2 | 0 | 0 | 2 | 0 | 0 | 6 | 0 |
| Normal | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 7 |

**Figure 6.** Confusion matrix for bicubic interpolation of videos.

**Table 3.** Classification performance for bicubic interpolation of videos.

| Class | Accuracy | Precision | Recall | F1-Score | Specificity | FPR | FNR |
|---|---|---|---|---|---|---|---|
| Abuse | 86.42% | 45.45% | 50.00% | 47.62% | 91.55% | 8.45% | 50.00% |
| Arrest | 93.83% | 100.00% | 50.00% | 66.67% | 100.00% | 0.00% | 50.00% |
| Arson | 87.65% | 55.56% | 45.45% | 50.00% | 94.29% | 5.71% | 54.55% |
| Assault | 86.42% | 45.45% | 50.00% | 47.62% | 91.55% | 8.45% | 50.00% |
| Burglary | 91.36% | 71.43% | 50.00% | 58.82% | 97.18% | 2.82% | 50.00% |
| Explosion | 90.12% | 60.00% | 60.00% | 60.00% | 94.37% | 5.63% | 40.00% |
| Fight | 86.42% | 46.15% | 60.00% | 52.17% | 90.14% | 9.86% | 40.00% |
| Normal | 86.42% | 46.67% | 70.00% | 56.00% | 88.73% | 11.27% | 30.00% |
| Macro-Average | 88.58% | 58.84% | 54.43% | 54.86% | 93.48% | 6.52% | 45.57% |

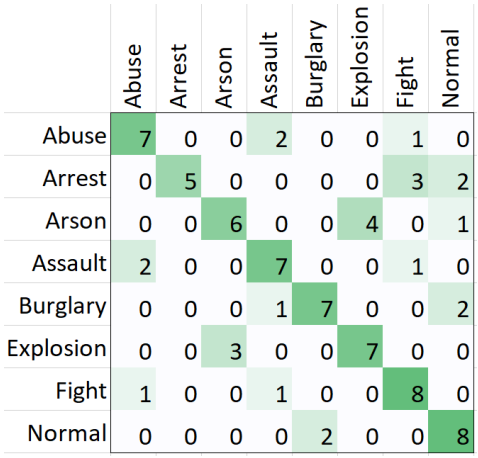


**Figure 7.** Confusion matrix for super-resolution videos.

**Table 4.** Classification performance for super-resolution videos.

| Class | Accuracy | Precision | Recall | F1-Score | Specificity | FPR | FNR |
|---|---|---|---|---|---|---|---|
| Abuse | 92.59% | 70.00% | 70.00% | 70.00% | 95.77% | 4.23% | 30.00% |
| Arrest | 95.06% | 100.00% | 60.00% | 75.00% | 100.00% | 0.00% | 40.00% |
| Arson | 90.12% | 66.67% | 54.55% | 60.00% | 95.71% | 4.29% | 45.45% |
| Assault | 91.36% | 63.64% | 70.00% | 66.67% | 94.37% | 5.63% | 30.00% |
| Burglary | 93.83% | 77.78% | 70.00% | 73.68% | 97.18% | 2.82% | 30.00% |
| Explosion | 91.36% | 63.64% | 70.00% | 66.67% | 94.37% | 5.63% | 30.00% |
| Fight | 92.59% | 66.67% | 80.00% | 72.73% | 94.37% | 5.63% | 20.00% |
| Normal | 91.36% | 61.54% | 80.00% | 69.57% | 92.96% | 7.04% | 20.00% |
| Macro-Average | 92.28% | 71.24% | 69.32% | 69.29% | 95.59% | 4.41% | 30.68% |

### *5.2. Comparison with Existing Approaches*

To validate the effectiveness of our proposed methodology, we conducted an extensive experimental variations. In addition to these experiments, we performed a comparative analysis between the TempoFuseNet framework and existing state-of-the-art approaches that center around multiclass anomaly classification, using the UCF-Crime dataset as our testing ground. In a similar context, Maqsood et al. [35] introduced a convolutional neural-network-based approach that initiates with video preprocessing to create 3D cubes through spatial augmentation. To streamline the analysis process, they employed a subset of the dataset: eliminating extraneous data and manually identifying atypical segments to ensure class balance. Subsequently, these 3D video cubes were fed into a convolutional neural network (CNN) to extract spatiotemporal features. Their analysis of the UCF-Crime dataset yielded a classification accuracy of 45% across fourteen distinct classes. In another study, Tiwari et al. [60] employed a fuzzy-rule-based approach for video summarization with the aim of addressing issues related to excessive data and high computational costs. Tiwari

et al. [60] achieved a classification accuracy of 53% in their classification experiment by leveraging a hybrid slow–fast neural network.

On the other hand, our study utilized a trimmed UCF-Crime dataset comprising eight classes and fifty videos. For the anomaly classification task, we applied a two-step approach: First, we upscaled the low-resolution (LR) videos using bicubic interpolation and an encoder–bank–decoder configuration for super resolution. The encoder and decoder played pivotal roles in downscaling and upscaling, while the bank was a pre-trained StyleGAN acting as a latent feature repository to enhance super-resolution performance based on feature priors. Our experiments encompassed both types of upscaled images, and the results were systematically compared in order to highlight the effectiveness of our super-resolution approach. Anomaly recognition was executed through a two-stream architecture wherein a pre-trained CNN model extracted features from RGB images in the spatial stream, and Stacked Grayscale 3-channel Images (SG3I) were used in the temporal stream, substantially reducing the computational load of optical flow computation while capturing short-term temporal characteristics. The features from both streams were concatenated and passed through a Gated Recurrent Unit (GRU) layer to capture long-term temporal characteristics. The output of the GRU layer was then processed through dense and softmax layers before reaching the final classification layer. Our proposed methodology, coupled with the encoder–bank–decoder super-resolution model, yielded remarkable results, achieving a classification accuracy of 92.28%, an F1-score of 69.29%, and a false positive rate of just 4.41%.

### *5.3. Comparison of Bicubic Interpolation and Super-Resolution Approaches*

In order to perform a comparison of both approaches, a bar-chart is plotted, as shown in Figure 8, for seven classification evaluation metrics; the chart clearly shows the superiority of super resolution over bicubic interpolation to perform anomaly identification. It is to be noted that the reported scores for accuracy, precision, recall, F1-score, and specificity are higher for super-resolution videos in comparison to bicubic interpolation videos, which is desirous, as higher values for these metrics indicate good classification performance. On the other hand, FPR and FNR should be lower for a good classification system, and therefore, their values are lower for video super-resolution scenarios. A clear performance gap indicates that super-resolution-based anomaly detection models are very effective when the video stream is of low spatial resolution.

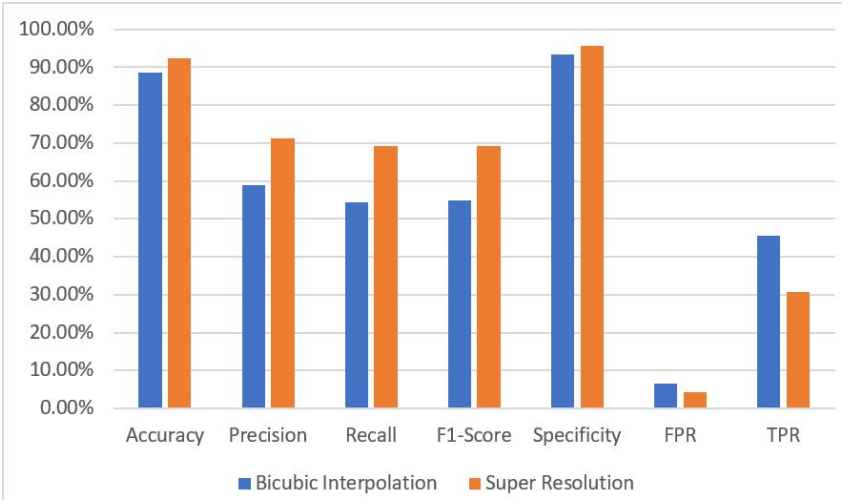


**Figure 8.** Comparison of bicubic interpolation and super-resolution approaches.

## 6. Conclusions

This study addressed the challenge of multi-class anomaly identification using low-quality surveillance videos within 5G and IoT environments. By conducting experiments on the trimmed UCF-crime dataset, the videos were downscaled to 1/8 resolution and then

upscaled using bicubic interpolation and super-resolution techniques. The TempoFuseNet framework employed a two-stream architecture that was followed by GRU for long-term temporal modeling. The experimental results showcased remarkable performance, with a classification accuracy of 92.28%, F1-score of 69.29%, and false positive rate of 4.41%. Moreover, the integration of super resolution in the anomaly classifier yielded substantial enhancements over the videos upscaled using bicubic interpolation. Specifically, the super-resolution-based approach achieved a 3.7% improvement in accuracy, a significant 14.34% boost in the F1-score, and a commendable 2.11% reduction in the false positive rate. Hence, TempoFuseNet outperforms existing state-of-the-art methods in multiclass classification performance and effectively addresses the technical limitations caused by low-quality videos, making it a robust solution for real-world surveillance scenarios, particularly in 5G and IoT environments.

*Future Work*

This study makes significant progress in improving video quality and anomaly detection in surveillance scenarios. However, future research could focus on integrating real-time processing capabilities and on investigating methods for automatically fine-tuning the model in response to changes to the lighting, spatial characteristics, or other dynamic factors. Moreover, integrating multi-modal data sources, such as contextual information, could improve anomaly detection accuracy and broaden the system's applicability in a variety of surveillance scenarios.

**Author Contributions:** Conceptualization, G.S. and U.I.B.; methodology, G.S.; software, G.S.; validation, G.S., U.I.B., R.H.R. and F.Z.; formal analysis, G.S. and U.I.B.; investigation, G.S., U.I.B., R.H.R. and F.Z.; resources, G.S.; data curation, G.S., U.I.B. and R.H.R.; writing—original draft preparation, G.S., U.I.B., R.H.R. and F.Z.; writing—review and editing, G.S., U.I.B., R.H.R. and F.Z.; visualization, G.S.; supervision, U.I.B. and R.H.R.; project administration, G.S., U.I.B. and R.H.R.; funding acquisition, F.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported in part by the National Natural Science Foundation of China with award number 62372409, and by the Ministry of Science and Technology with award number DL2023147001L.

**Data Availability Statement:** The dataset used in this study for conducting experiments for Anomaly Recognition is publicly available at: UCF-CRIME. Retrieved September 2023, https://www.crcv.ucf.edu/projects/real-world/ (accessed on 8 January 2024).

**Acknowledgments:** This study acknowledges partial support from the National Center of Big Data and Cloud Computing (NCBC) and HEC of Pakistan.

**Conflicts of Interest:** The authors declare that they have no known competing financial interests or personal relationships that could appear to have influence the work reported in this paper.

## References

1. 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, 26 June–1 July 2016; pp. 733–742.
2. 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.
3. Zhong, J.X.; Li, N.; Kong, W.; Liu, S.; Li, T.H.; Li, G. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 1237–1246.
4. Akhtar, M. Automated analysis of visual leaf shape features for plant classification. *Comput. Electron. Agric.* **2019**, *157*, 270–280.
5. Ahmad, N.; Asif, H.M.S.; Saleem, G.; Younus, M.U.; Anwar, S.; Anjum, M.R. Leaf image-based plant disease identification using color and texture features. *Wirel. Pers. Commun.* **2021**, *121*, 1139–1168. [CrossRef]
6. Aslam, M.A.; Wei, X.; Ahmed, N.; Saleem, G.; Amin, T.; Caixue, H. Vrl-iqa: Visual representation learning for image quality assessment. *IEEE Access* **2024**, *12*, 2458–2473. [CrossRef]
7. Ahmed, N.; Asif, H.M.S. Perceptual quality assessment of digital images using deep features. *Comput. Inform.* **2020**, *39*, 385–409. [CrossRef]

8. Ahmed, N.; Asif, H.M.S. Ensembling convolutional neural networks for perceptual image quality assessment. In Proceedings of the 2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS), Karachi, Pakistan, 14–15 December 2019; pp. 1–5.
9. Saleem, G.; Bajwa, U.I.; Raza, R.H.; Alqahtani, F.H.; Tolba, A.; Xia, F. Efficient anomaly recognition using surveillance videos. *PeerJ Comput. Sci.* **2022**, *8*, e1117. [CrossRef] [PubMed]
10. Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S. A review of video surveillance systems. *J. Vis. Commun. Image Represent.* **2021**, *77*, 103116. [CrossRef]
11. Duong, H.T.; Le, V.T.; Hoang, V.T. Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey. *Sensors* **2023**, *23*, 5024. [CrossRef]
12. Ahmed, N.; Asif, H.M.S.; Khalid, H. Image quality assessment using a combination of hand-crafted and deep features. In *Proceedings of the Intelligent Technologies and Applications: Second International Conference, INTAP 2019, Bahawalpur, Pakistan, 6–8 November 2019*; Revised Selected Papers 2; Springer: Singapore, 2020; pp. 593–605.
13. Khalid, H.; Ali, M.; Ahmed, N. Gaussian process-based feature-enriched blind image quality assessment. *J. Vis. Commun. Image Represent.* **2021**, *77*, 103092. [CrossRef]
14. Ahmed, N.; Asif, S. BIQ2021: A large-scale blind image quality assessment database. *J. Electron. Imaging* **2022**, *31*, 053010. [CrossRef]
15. Ahmed, N.; Asif, H.M.S.; Saleem, G.; Younus, M.U. Image quality assessment for foliar disease identification (agropath). *J. Agric. Res.* **2021**, *59*, 177–186.
16. Ahmed, N.; Shahzad Asif, H.; Bhatti, A.R.; Khan, A. Deep ensembling for perceptual image quality assessment. *Soft Comput.* **2022**, *26*, 7601–7622. [CrossRef]
17. Ahmed, N.; Asif, H.M.S.; Khalid, H. PIQI: Perceptual image quality index based on ensemble of Gaussian process regression. *Multimed. Tools Appl.* **2021**, *80*, 15677–15700. [CrossRef]
18. Ullah, W.; Ullah, A.; Hussain, T.; Khan, Z.A.; Baik, S.W. An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos. *Sensors* **2021**, *21*, 2811. [CrossRef]
19. Zhou, Y.; Du, X.; Wang, M.; Huo, S.; Zhang, Y.; Kung, S.Y. Cross-scale residual network: A general framework for image super-resolution, denoising, and deblocking. *IEEE Trans. Cybern.* **2021**, *52*, 5855–5867. [CrossRef]
20. Kwan, C.; Zhou, J.; Wang, Z.; Li, B. Efficient anomaly detection algorithms for summarizing low quality videos. In Proceedings of the Pattern Recognition and Tracking XXIX, Orlando, FL, USA, 15–19 April 2018; SPIE: Bellingham, WA, USA; Volume 10649, pp. 45–55.
21. Zhou, J.; Kwan, C. Anomaly detection in low quality traffic monitoring videos using optical flow. In Proceedings of the Pattern Recognition and Tracking XXIX, Orlando, FL, USA, 15–19 April 2018; SPIE: Bellingham, WA, USA, 2018; Volume 10649, pp. 122–132.
22. Lv, Z.; Wu, J.; Xie, S.; Gander, A.J. Video enhancement and super-resolution. In *Digital Image Enhancement and Reconstruction*; Elsevier: Amsterdam, The Netherlands, 2023; pp. 1–28.
23. Wu, J. Introduction to convolutional neural networks. *Natl. Key Lab Nov. Softw. Technol. Nanjing Univ. China* **2017**, *5*, 495.
24. Nguyen, T.N.; Meunier, J. Hybrid deep network for anomaly detection. *arXiv* **2019**, arXiv:1908.06347.
25. Ryoo, M.; Kim, K.; Yang, H. Extreme low resolution activity recognition with multi-siamese embedding learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; PKP Publishing Services Network; Volume 32.
26. Saleem, G.; Bajwa, U.I.; Raza, R.H. Toward human activity recognition: A survey. *Neural Comput. Appl.* **2023**, *35*, 4145–4182. [CrossRef]
27. 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]
28. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3551–3558.
29. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
31. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 568–576.
32. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
33. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2740–2755. [CrossRef]
34. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
35. Maqsood, R.; Bajwa, U.I.; Saleem, G.; Raza, R.H.; Anwar, M.W. Anomaly recognition from surveillance videos using 3D convolution neural network. *Multimed. Tools Appl.* **2021**, *80*, 18693–18716. [CrossRef]

36. Luo, W.; Liu, W.; Gao, S. Remembering history with convolutional lstm for anomaly detection. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 439–444.
37. Ullah, M.; Mudassar Yamin, M.; Mohammed, A.; Daud Khan, S.; Ullah, H.; Alaya Cheikh, F. Attention-based LSTM network for action recognition in sports. *Electron. Imaging* **2021**, *2021*, 302–311. [CrossRef]
38. Riaz, H.; Uzair, M.; Ullah, H.; Ullah, M. Anomalous human action detection using a cascade of deep learning models. In Proceedings of the 2021 9th European Workshop on Visual Information Processing (EUVIP), Paris, France, 23–25 June 2021; pp. 1–5.
39. Liu, W.; Luo, W.; Lian, D.; Gao, S. Future frame prediction for anomaly detection—A new baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6536–6545.
40. Landi, F.; Snoek, C.G.; Cucchiara, R. Anomaly locality in video surveillance. *arXiv* **2019**, arXiv:1901.10364.
41. Sabokrou, M.; Khalooei, M.; Fathy, M.; Adeli, E. Adversarially learned one-class classifier for novelty detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3379–3388.
42. Zhou, Y.; Qu, Y.; Xu, X.; Shen, F.; Song, J.; Shen, H. BatchNorm-based Weakly Supervised Video Anomaly Detection. *arXiv* **2023**, arXiv:2311.15367.
43. Pu, Y.; Wu, X.; Wang, S. Learning Prompt-Enhanced Context Features for Weakly-Supervised Video Anomaly Detection. *arXiv* **2023**, arXiv:2306.14451.
44. Gan, K.Y.; Cheng, Y.T.; Tan, H.K.; Ng, H.F.; Leung, M.K.; Chuah, J.H. Contrastive-regularized U-Net for Video Anomaly Detection. *IEEE Access* **2023**, *11*, 36658–36671. [CrossRef]
45. Ryoo, M.; Rothrock, B.; Fleming, C.; Yang, H.J. Privacy-preserving human activity recognition from extreme low resolution. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
46. Chen, J.; Wu, J.; Konrad, J.; Ishwar, P. Semi-coupled two-stream fusion convnets for action recognition at extremely low resolutions. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 139–147.
47. Xu, M.; Sharghi, A.; Chen, X.; Crandall, D.J. Fully-coupled two-stream spatiotemporal networks for extremely low resolution action recognition. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1607–1615.
48. Demir, U.; Rawat, Y.S.; Shah, M. Tinyvirat: Low-resolution video action recognition. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 7387–7394.
49. Hou, W.; Wang, X.; Chouinard, J.Y.; Refaey, A. Physical layer authentication for mobile systems with time-varying carrier frequency offsets. *IEEE Trans. Commun.* **2014**, *62*, 1658–1667. [CrossRef]
50. Ataer-Cansizoglu, E.; Jones, M.; Zhang, Z.; Sullivan, A. Verification of very low-resolution faces using an identity-preserving deep face super-resolution network. *arXiv* **2019**, arXiv:1903.10974.
51. Bai, Y.; Zhang, Y.; Ding, M.; Ghanem, B. Sod-mtgan: Small object detection via multi-task generative adversarial network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 206–221.
52. Wang, Z.; Ye, M.; Yang, F.; Bai, X.; Satoh, S. Cascaded SR-GAN for scale-adaptive low resolution person re-identification. In Proceedings of the IJCAI, Stockholm, Sweden, 13–19 July 2018; Volume 1, p. 4.
53. Lu, C.; Shi, J.; Jia, J. Abnormal event detection at 150 fps in matlab. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2720–2727.
54. 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, Beijing, China, 24–28 October 2010; pp. 1975–1981.
55. Han, D. Comparison of commonly used image interpolation methods. In Proceedings of the Conference of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013), Los Angeles, CA, USA, 1–2 July 2013; Atlantis Press: Amsterdam, The Netherlands, 2013; pp. 1556–1559.
56. Chan, K.C.; Xu, X.; Wang, X.; Gu, J.; Loy, C.C. GLEAN: Generative latent bank for image super-resolution and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 3154–3168. [CrossRef]
57. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8110–8119.
58. Barkhordar, E.; Shirali-Shahreza, M.H.; Sadeghi, H.R. Clustering of Bank Customers using LSTM-based encoder-decoder and Dynamic Time Warping. *arXiv* **2021**, arXiv:2110.11769.
59. Kim, J.H.; Won, C.S. Action recognition in videos using pre-trained 2D convolutional neural networks. *IEEE Access* **2020**, *8*, 60179–60188. [CrossRef]
60. Tiwari, A.; Chaudhury, S.; Singh, S.; Saurav, S. Video Classification using SlowFast Network via Fuzzy rule. In Proceedings of the 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Luxembourg, 11–14 June 2021; pp. 1–6.