IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 October 2024, accepted 24 October 2024, date of publication 1 November 2024, date of current version 12 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3488797

RESEARCH ARTICLE

Enhanced Anomaly Detection in Pandemic Surveillance Videos: An Attention Approach With EfficientNet-B0 and CBAM Integration

SAREER UL AMIN[1], MUHAMMAD SIBTAIN ABBAS[2], BUMSOO KIM[3], YONGHOON JUNG[3], AND SANGHYUN SEO[4]

[1]Department of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea
[2]Department of Architectural Engineering, Chung-Ang University, Seoul 06974, South Korea
[3]Department of Applied Art and Technology, Chung-Ang University, Anseong-si 17546, South Korea
[4]College of Art and Technology, Chung-Ang University, Anseong-si 17546, South Korea

Corresponding author: Sanghyun Seo (sanghyun@cau.ac.kr)

This work was supported by the Ministry of Science and ICT (MSIT), South Korea, under the Information Technology Research Centre (ITRC) Support Program supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP) under Grant IITP-2024-RS-2024-00438056.

ABSTRACT We present a novel system for anomaly detection in surveillance videos, specifically focusing on identifying instances where individuals deviate from public health guidelines during the pandemic. These anomalies encompassed behaviours like the absence of face masks, incorrect mask usage, coughing, nose-picking, sneezing, spitting, and yawning. Monitoring such anomalies manually was challenging and prone to errors, necessitating automated solutions. To address this, a multi-attention-based deep learning system was employed, utilizing the EfficientNet-B0 architecture. EfficientNet-B0, featuring the Mobile Inverted Bottleneck Convolution (MBConv) block with Squeeze-and-Excitation (SE) modules, emphasizes informative channel characteristics while disregarding irrelevant ones. However, this approach neglected crucial spatial information necessary for visual recognition tasks. To improve this, the Convolutional Block Attention Module (CBAM) was integrated into EfficientNet-B0 to improve feature extraction. The baseline EfficientNet-B0 model's SE module was replaced with the CBAM module within each MBConv module to retain spatial information related to anomaly activities. Additionally, the CBAM module, when embedded after the second convolutional layer, was observed to significantly enhance the classification ability of the model across different anomaly classes, resulting in a significant accuracy boost from 87 to 96%. In conclusion, we demonstrated the efficacy of the CBAM module in refining feature extraction and improving the classification performance of the proposed method, showcasing its potential for robust anomaly detection in surveillance videos.

INDEX TERMS Anomaly detection, video surveillance, computer vision, attention method, intelligent surveillance system.

I. INTRODUCTION

The COVID-19 pandemic has had an immense effect on nations worldwide in an incredibly short period. The World Health Organisation (WHO) [1] reports that as of October 12, 2023, there were around 7 million COVID-19 fatalities globally and over 771 million confirmed cases. The virus can cause serious illnesses like Severe Acute Respiratory Syndrome (SARS) and Chronic Obstructive Pulmonary Disease (COPD), even though its initial symptoms are usually moderate ones like temperature and vomiting. Considering its first diagnosis in Wuhan, China, in December 2019, the virus has spread quickly worldwide, with the highly contagious Omicron form raising special concerns. COVID-19 is mostly transmitted by respiratory droplets produced when people cough, sneeze, talk, yell, or sing. These droplets have the

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

ability to propagate in the air, settle on people's lips or noses, enter their lungs, or be breathed.

To reduce the transmission of the virus, masks serve as a relatively simple barrier to prevent respiratory droplets from reaching other people. Studies have shown that covering the mouth and nose with masks can significantly stop droplets from spreading [2]. Howard et al. [3] claim that wearing masks outdoors can dramatically halt the spread of illnesses, especially in areas with high population density. In addition, the virus is highly contagious and easily transmitted through respiratory droplets when an infected person sneezes, coughs, spits, and yawns; it can also spread when people touch infected surfaces and then contact their face, including the nostrils, mouth, and eyes, because these tiny droplets fall promptly on floors or surfaces and are not sufficiently dense to travel in the air for extended periods [4]. Furthermore, as face masks successfully limit tiny respiratory droplets, wearing them has become vital to the COVID-19 protocol. According to Li et al. [5] and Longrich and Sheppard [6], surgical face masks and N95 have different blocking efficiencies of 91 and 68%, respectively, in preventing the spread of viruses by obstructing droplets from the airway. It is possible to avoid some pollutants from entering another individual's respiratory system by wearing face masks, which efficiently block the transmission of viruses and dust particles [7].

Existing studies have explored computer vision techniques for COVID-19 surveillance. Bhattacharya et al. [8] focused on deep learning for infection diagnosis, while Rehmani and Mirmahaleh [9] assessed preventive strategies. Shetty et al. [10] proposed monitoring methods for public safety, and Ge et al. [11] achieved 76.1% accuracy in face mask recognition using a locally linear embedding approach with convolutional neural networks (CNNs). Loey et al. [12] developed a ResNet50 and YOLO v2 combination for precise face mask detection. Rezaei and Azarmi [13] introduced an integrated YOLOv4-based system for crowd monitoring, emphasizing social distancing. However, these methods fail to address the complexities of anomaly detection and identifying instances of non-compliance with public health guidelines during the pandemic. Monitoring anomalies in public spots manually is challenging for authorities with limited resources and can result in human errors. Anomaly detection in videos has evolved into a crucial problem in computer vision due to its potential to support global efforts to combat the virus's spread.

To address this gap, the present study proposes a multi-attention-based deep learning system to detect anomalies in video. In this research, the proposed framework was improved by replacing the Squeeze-and-Excitation (SE) module with a Convolutional Block Attention Module (CBAM), enhancing spatial information preservation for anomaly detection. Integrating CBAM after the second convolutional layer significantly improved overall model performance, demonstrating the importance of attention mechanisms in boosting accuracy for anomaly detection and classification. This modification represents a notable advancement in addressing the complexities of anomaly detection while maintaining spatial context.

The following research questions serve as the research's main priority:

- How could EfficientNet-B0 be optimized to improve anomaly detection accuracy?
- ''What impact do attention mechanisms have on the performance of EfficientNet-B0?''
- How can the proposed framework be evaluated against contemporary techniques to demonstrate its effectiveness in accurately detecting and classifying anomalies in surveillance videos?
- What are the potential future directions and enhancements that can be explored to improve the performance of the proposed system further?

Our research work has contributed significantly, and its fundamental objectives can be summarised as follows:

- **Dataset Creation:** We employed the Canon EOS 600D camera to capture anomalous activities within indoor environments. A specialized anomaly detection dataset was compiled, encompassing eight distinct categories of anomaly activities: Coughing, face mask adherence, lack of mask, nose picking, sneezing, spitting, improper mask usage, and yawning. Each anomaly activity category comprises 18 videos, each approximately one minute in duration. Notably, within the context of this research, the ''face mask'' class was considered a regular activity, while the remaining classes were designated as anomalous activities. More facts about the suggested dataset are provided in Section IV-A.
- **Model Modification:** In this research, a significant modification was made to the original EfficientNet-B0 model [27], replacing the SE module present in every Mobile Inverted Bottleneck Convolution (MBConv) module with a CBAM [36]. This modification was instrumental in preserving crucial spatial information related to anomaly activities while efficiently gathering essential channel-specific features. The integration of the CBAM module led to a more comprehensive understanding of anomaly events, enhancing the model's ability to detect and classify various anomalous behaviors in surveillance videos accurately. This adjustment represents a noteworthy advancement in the model's overall performance, effectively addressing the complex challenges of anomaly detection while preserving spatial context.
- **Strategic Integration:** We strategically integrated the CBAM module into the EfficientNet-B0 architecture after the second convolutional layer. This strategic placement of the CBAM module played a pivotal role in enhancing the model's overall performance. By refining the extracted feature information and leveraging its attention-based mechanisms, the CBAM module significantly improved the model's capability to classify various anomaly classes in surveillance

videos. This enhancement not only streamlined the feature extraction process but boosted the model's classification capabilities, resulting in a substantial increase in accuracy. This demonstrates the significance of attention mechanisms in enhancing the efficiency of anomaly detection and classification in video data.

- **Empirical Validation:** The experiments were conducted using the anomaly detection datasets presented in this study. The outcomes of these experiments reveal that the proposed approach surpasses the performance of contemporary techniques across various metrics, including precision, recall, F1-Score, and accuracy. These promising findings emphasize the efficacy of the developed approach, indicating its potential as a reliable solution for enhancing anomaly detection capabilities in similar surveillance settings.

The following sections are structured in the following manner. Section II presents a concise summary of pertinent literature. Section III provides a comprehensive discussion of the suggested techniques. The results are presented in Section IV, along with thorough details. Our conclusion and future direction are presented in Section V.

II. LITERATURE REVIEW

Anomaly detection in COVID-19 surveillance videos has become an important area of research due to the pandemic. The use of surveillance videos to monitor and stop the virus's spread has gained widespread attention due to its potential to provide real-time insights into the behavior of individuals in public spaces, such as airports, hospitals, and public transportation [16]. In addition, anomaly detection in COVID-19 surveillance videos has used computer vision techniques to detect facial and body movements associated with symptoms of the virus. Several studies have explored computer vision techniques for anomaly detection in COVID-19 surveillance videos. According to [17], isolation, sanitation of the hands, and hiding one's nostrils and mouth with a piece of cloth are crucial tactics to combat this serious pandemic. They claim that the mask has emerged as an indispensable component of public health strategies to address the present pandemic. A suitable design for limiting airborne infections in the outdoor setting was presented by Leng et al. [18]. A Study on deep learning-based techniques for COVID-19 infection diagnosis in medical images was conducted in [8], [19], [20]. A comprehensive assessment of several COVID-19 disease preventive and treatment strategies was presented by Rehmani and Mirmahaleh [9]. Researchers offered effective social isolation monitoring methods by leveraging deep learning techniques. A system for detecting and tracking individuals using the YOLOv3 with the DeepSort approach was introduced by Punn et al. [21]; they used a publicly available image data repository. The authors looked at results using additional deep-learning models as well. A drone camera-based approach for tracking social distance was introduced by Ramadass and Arunachalam [22]. Additionally, they employed YOLO-V3, training it using their own dataset, including front and side perspectives. The study conducted by Shetty et al. [10] suggests a methodology that employs a technique to monitor an individual's behavior and ensure their safety in public places. The study aims to determine whether a person is putting on masks and to ensure social distancing following prescribed guidelines and standards. Ge et al. [11] achieved 76.1% accuracy in face mask recognition using a local linear embedding approach with CNNs. Teboulbi et al.'s [23] study describes the application of identifying face masks and social distance as an integrated system. This paper presents a comparative analysis of several face mask classification and detection methods. The system ensures social isolation by tracking individuals wearing or not wearing masks in a real-time environment. A cascade CNN scheme was presented by Bu et al. [24] to detect the covered faces in the MASKED FACE dataset. Ahmed et al.'s [25] study suggests a scheme for social distance as a preventative measure that is used to track, maintain, minimize, and manage real-time in-person interaction between individuals in top-view environments. Faster Recurrent Convolutional Neural Networks (FRCNN) were employed to detect people in real time. Limbasiya and Raut proposed a detector technique in their work [26]. This technique is a Single-Stage Detector (SSD) that employs an Artificial Neural Network (ANN) to integrate semantic information from multiple feature maps and a Machine Learning module that is specifically designed to identify instances of social distancing and mask-wearing. The method can ensure public safety and security in any environment. Loey et al. [12] recently published a paper describing a powerful deep-learning method that combines ResNet50 and YOLO-v2 for precise face mask detection. Their results showed that this model has an amazing 81% accuracy. Similarly, using a top-down perspective, Ahmed et al. [25] have presented a unique deep-learning approach for monitoring social distance. Their method exploits the YOLO-v3 detection algorithm to detect people in video sequences. Rezaei and Azarmi 's [13] study proposes an integrated YOLOv4-based system that employs surveillance cameras to detect individuals in crowds in indoor and outdoor environments automatically. Their suggested network combines an algorithm of sort tracking and an improved version of inverse perspective mapping (IPM) to detect resilient individuals and monitor social distance. In [14], the authors devised a space-embedding approach for multivariant time series anomaly detection (SES-AD). Being a hybrid approach, SES-AD shifted the raw series to low-dimensional space instead of directly searching the discords within the initial temporal series. This allowed the disparity vector to readily identify the important unexpected shift locations in the new dimension. Lastly, a statistical approach was used to identify the possible anomalies. Furthermore, Hu et al. [15] presented an innovative computational approach for detecting discrepancies in multivariate sequence data. The method

correctly finds the discrepancies by analyzing a recurrence plot generated from the initial temporal data.

III. METHODOLOGY

In this section, a comprehensive analysis of the suggested anomaly detection network and its fundamental constituent structure is provided. The visual illustration of the proposed framework is presented in Fig. 1 to facilitate a better understanding of its intricate design. In this study, the proposed strategy enhances the efficiency of the EfficientNet-B0 model using CBAM. The EfficientNet-B0 model is a lightweight CNN method developed for efficient training and deployment on edge devices. On the contrary, the CBAM modules are attention mechanisms that assist the proposed framework in concentrating on the most informative features and improving its generalization capability.

A. EFFICIENTNET-B0

EfficientNet-B0 is an approach for classification that was first developed by the Google Brain Researchers [27]. According to their research on network scaling, the network's depth, breadth, and resolution might be adjusted to improve performance. In addition, they scaled their neural network to generate deeper deep learning models, which outperform the earlier used CNNs in terms of efficacy and accuracy [28]. Apart from that, EfficientNet also carried out accurate and reliable large-scale image classification for ImageNet. Compared to the best current methods, such as VGGNets [29], GoogleNet [30], Xception [31], ResNets [32], and InceptionResNet [33], the entries in the in the EfficientNet family are roughly eight times smaller and six times faster for inference. Furthermore, the proposed method employed EfficientNet-B0 because of its higher efficiency and accuracy compared to existing CNNs. Moreover, this model employs a fixed set of scaled parameters to alter the system's breadth, depth, and resolution in an equitable manner. By leveraging a composite scaling method, EfficientNet-B0 produces special CNN models. The network depth is equal to the number of linked layers. The number of filters within a convolution layer determines its width. The given image's width and height define its resolution. The existing EfficientNet-B0 baseline architecture, which accepts an input picture size of $224 \times 224 \times 3$, is depicted in Fig. 1. Besides that, the model employs movable invert bottleneck Conv and several Conv layers featuring a 3×3 receptive space for capturing features at every single layer. Equations (1)–(5) show the depth, breadth, and resolution scalability that the authors suggest with respect to ϕ.

$$d = \alpha^{\phi} \tag{1}$$

$$w = \beta^{\phi} \tag{2}$$

$$r = \gamma^{\phi} \tag{3}$$

$$\text{s.t.} \quad \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \tag{4}$$

$$\alpha \geq 1, \quad \beta \geq 1, \quad \gamma \geq 1 \tag{5}$$

The constants α, β, and γ have been found by applying the grid search hyperparameter tuning technique. The network's depth, width, and resolution are represented by the variables d, w, and r. Additionally, the model's scalability components are managed by a preset variable called the coefficient. To optimize memory usage and network reliability, this method modifies the network's depth, breadth, and resolution according to its resources. In contrast to all previous deep CNNs, EfficientNet-B0 adjusted each dimension using a predetermined set of scalability coefficients, enabling it to surpass other contemporary algorithms trained on the ImageNet dataset.

B. CHANNEL AND SPATIAL ATTENTION MODULES

Attention mechanisms in computer vision refer to the ability to concentrate attention on significant regions of an image while ignoring others. Attention methods are applied in the human brain's visual cortex to swiftly and effectively interpret complicated visual content. Later, this approach was integrated into computer vision to enhance functionality. Attention methods can be considered dynamic processes that use flexible feature weighting to select salient details from a picture. The effectiveness of computer vision has been significantly enhanced by attention processes. Deep learning-based attention mechanisms are used in various tasks, including speech, image, and natural language processing. They filter a significant amount of deep learning samples, eliminating irrelevant details and choosing information that is more important to the task at hand [34].

The CBAM is a low-weight attention module employed in this work. CBAM is an innovative module that combines different attention methods, making it easy to integrate into CNN networks. This module enriches CNN's feature representation capabilities, making it a simple but effective improvement. Moreover, CBAM is a cost-effective solution that can be easily integrated into different network configurations without demanding a lot of processing costs [35]. The CBAM module can sequentially produce an attention pattern map within the channel and spatial boundaries. The final map is then produced by multiplying both sets of attention map details with the initial input map for adaptable pattern adjustments. Fig. 2 displays the CBAM method. As seen in Fig. 2(A), the method takes the intermediate activation map F as input and uses the Channel Attention Module (CAM) to construct the feature map F' and the Spatial Attention Module (SAM) to construct the feature map F''. Equation 6 could be used to explain the computation process.

$$\begin{cases} F' = M_C(F) \times F \\ F'' = M_S\left(F'\right) \times F' \end{cases} \tag{6}$$

Here, the multiplicative action within the matching components is represented by $\otimes$. The given input characteristic map is denoted as $F(\in R^{C \times H \times W})$. The CA reflects the resultant weight of F' as $M_C\left(\in R^{C \times 1 \times 1}\right)$. The resultant weight of

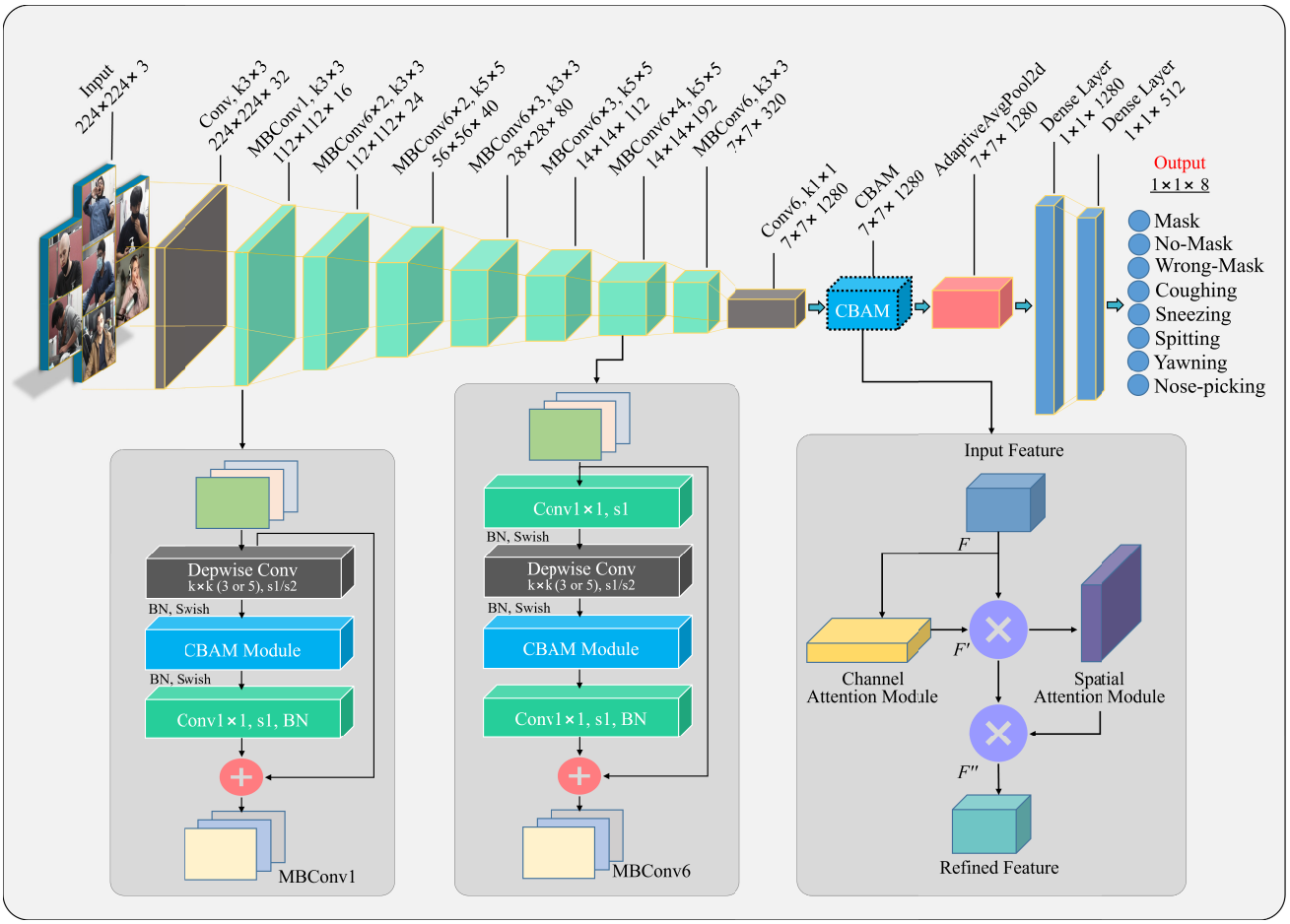

FIGURE 1. An overview of proposed ABADN framework. The baseline model was enhanced by substituting the SE modules with CBAM for each MBConv module. This modification preserved vital spatial information regarding anomaly activity while enhancing channel-specific data acquisition. The CBAM module was strategically incorporated after the second convolutional layer, significantly boosting the model's ability to classify diverse anomaly classes by refining feature information and improving its classification performance.

F'' by carrying out SA is represented by $M_S\left(\in R^{1\times H\times W}\right)$. Fig. 2(B) depicts the CAM's mechanism. In the first phase of the CAM, a couple of $C \times 1 \times 1$ feature maps are created by performing the average and max pooling operations depending on width and height. The ultimate CA weights, Mc, are then generated by feeding them to the provided MLP layer for summation, where they get activated through the activation function of the sigmoid. Equation 7 can be used to demonstrate the CA computation process.

$$M_C(F) = \sigma[\text{MLP}(\text{AvgPool}(F)] + \text{MLP}[\text{MaxPool}(F)] \quad (7)$$

MLP stands for a multilayer perception, and σ for a sigmoid function. In Fig. 2(C), the SAM's mechanism is displayed. F' is the feature map fed into the SA process. A couple of $1 \times H \times W$ activation map is created on the channel by executing the average and max pooling steps. This feature map is subsequently subjected to the concatenate operation. The SA weights M_C are generated by employing a 7×7 convolution filter to decrease the size of the map's features. Equation 8 represents the process of calculating SA.

$$M_S(F) = \sigma\left\{f^{7\times7}[\text{AvgPool}(F); \text{MaxPool}(F)]\right\} \quad (8)$$

where $f^{7\times7}$ represents the convolution process used to gather the spatial patterns of the target, with a convolution filter size of 7×7.

C. PROPOSED NETWORK

EfficientNet-B0 consists of the MBConv block, where a SE module is contained in each MBConv module. The gating or focusing of channel dimensions is managed by the SE module. The model may focus on the most relevant channel features while disregarding the less significant ones. This process of feature selection may result in the loss of essential spatial information, which is crucial for image recognition tasks. The model considers only the channel information; hence, spatial information is ignored. Therefore, in the current study, the CBAM module was added to Efficientnet-B0 to enhance its ability to extract significant patterns [35]. The CBAM module was specifically selected for its effectiveness in capturing both channel and spatial dependencies, which we believe are crucial for enhancing features in our anomaly detection model. It allows the network to focus on important regions and refine feature representations. Furthermore, the CBAM module aims to strengthen the attention mechanisms

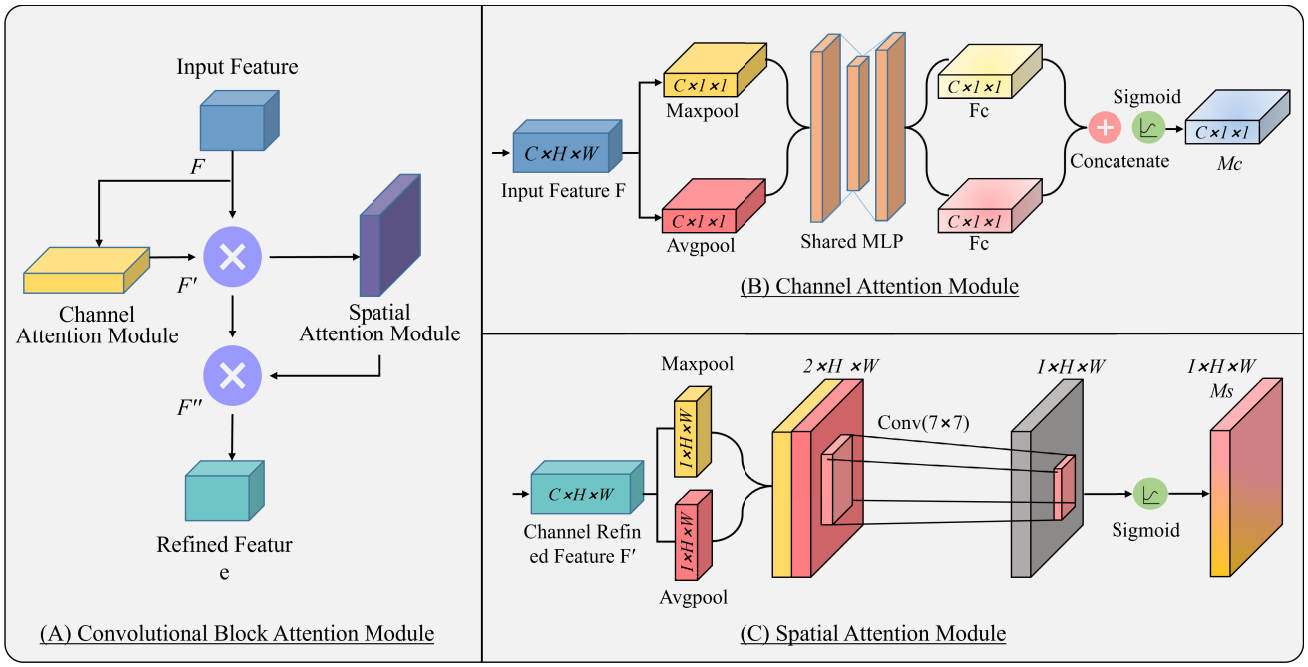

FIGURE 2. The intermediate feature map *F* is fed into the CBAM module, which creates the feature map *F′* using the CAM and the feature map *F″* using the SAM. Furthermore, (B) displays the CAM mechanism, while (C) displays the SAM mechanism.

of deep learning algorithms to increase their performance. Although CBAM seems to be a very simple module, its unique quality is that it conveniently captures both spatial and channel-wise attention [36]. Moreover, the CBAM module introduces two attention processes: the CAM and the SAM. The CAM focuses on channel-specific attention (CA), allowing the network to concentrate on significant channels while suppressing unnecessary ones. Conversely, the SAM concentrates on SA, which allows the model to pay attention to certain spatial regions in the feature maps [37]. Besides that, CBAM is an innovative method because it combines different attention approaches into a single module that can be easily integrated into CNN networks. In this way, CBAM can effectively improve the feature representation capabilities of CNNs. Due to its simplicity, it can be easily integrated into different network designs without requiring a lot of processing costs [36]. Fig. 1 depicts the improved proposed framework architecture. In comparison to the original Efficientnet-B0 method, the following improvements have been made:

- A CBAM module was inserted in place of the SE component within each MBConv component of the initial EfficientNet-B0 architecture. As a result, the network was able to gather channel characteristics without sacrificing important spatial characteristics regarding the anomaly activity.
- Following the second convolutional layer, the CBAM module was included in the EfficientNet-B0 architecture. It refined the retrieved feature information and strengthened the framework's classification capabilities, which increased the ability of the model to classify various anomaly types.

1) OPTIMIZATION

The EfficientNet-B0 model is optimized using the traditional Stochastic Gradient Descent (SGD) technique. It was challenging to determine an appropriate learning rate for the SGD algorithm since every parameter had an identical learning rate. Furthermore, the SGD optimization approach quickly converges to a local optima during model training, making it difficult for the model to provide a suitable training model for different kinds of anomaly detection tasks. This research used the Adam optimization technique to tackle the aforementioned problem. The Adam algorithm's parameters were independently changed, and each of them maintained a learning rate. To further improve the smoothness of the model convergence and lessen spikes in parameter updates, each learning rate tweaking has been bias-corrected. Momentum updates and learning rate modifications are coupled in the Adam optimization approach, and the initial and subsequent moments of the gradient constantly modify the learning rates of each parameter [38], [39], [40]. Equation 9 may be used to represent the computation procedure.

$$\begin{cases} \theta_t = \theta_{t-1} - \alpha \cdot \dfrac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \\ \hat{m}_t = \dfrac{m_t}{1 - \beta_1^t} \\ \hat{v}_t = \dfrac{v_t}{1 - \beta_2^t} \\ m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \\ v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \\ g_t = \nabla_\theta f_t \left(\theta_{t-1}\right) \end{cases} \tag{9}$$

where θ_t and θ_{t-1} reflect the respective parameters of the t and $t - 1$ updates. The gradient mean that has been

exponentially shifted is denoted by m_t. The squared gradient is expressed by v_t. The current value of m_t is represented as $\hat{m}_t$. The revised value for v_t is denoted by $\hat{v}_t$. The constants β_1 and β_2 are employed to regulate exponential decay. The symbol g_t represents the first-order derivative. $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ are the default values that correspond to each parameter.The first-order derivative is denoted by g_t. $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ are the default values that correspond to each parameter.

2) TRANSFER LEARNING

Transfer learning is a technique where a previously trained network is applied as an initial baseline for a new task. The pre-trained model is adjusted or employed as a feature extractor to fit the new goal, as opposed to starting from scratch and training a new model. This is particularly useful when the new task has limited data or computational resources, as transfer learning allows for the leveraging of knowledge learned from a large dataset and powerful computational resources during pre-training. By adopting the pre-trained algorithm, the new model can learn from a smaller set of data, generalize better, and converge faster [41]. Motivated by this, the proposed framework in this research is trained by employing a transfer learning approach. For initialization, only the pretrained weights were utilized. We leveraged our own custom anomaly dataset to train all of the models completely. The last dense layer in each method has been decreased from 1000 to 8 since there were only eight classes. In the last layer, the SoftMax activation function has been included. The models have been trained on 50 epochs utilizing the Adam optimization algorithm with categorical cross-entropy as a loss function. The parameters of the Adam optimisation method were $\epsilon = 10^{-8}$, $\alpha = 0.001$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The batch size was limited to 16 owing to hardware limitations, and the starting learning was set to 0.001. A dropout rate of 0.35 was employed prior to the last fully linked layer in the suggested model.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed ABADN framework is evaluated using the proposed datasets. A selection of normal and anomalous event frames taken from the presented dataset are shown in Fig. 3. Additionally, Fig. 6 displays the assessed dataset's Three-Dimensional Uniform Manifold Approximation and Projection (3D-UMAP) Projection. In order to assess how the CBAM module affects the model's performance in anomaly detection tasks, Fig. 4 illustrates the ablation study of the suggested design. The statistical information about the UCSDPed-1, UCSDPed-2, and Avenue datasets is provided in Table 2. In Tables 3 and 4, quantitative measures were used to evaluate the performance of the presented method. Fig. 5(A) and (B) depict the training and validation accuracy graphs of the proposed method employing our own Dataset, while Fig. 5(C) and (D) show the training and validation loss graph respectively. The presented method was executed in Keras with TensorFlow as its backend, using a Python version 3.9 programming environment. The results were conducted on a personal computer rigged with an NVidia GPU (3070 RTX) with 32GB of RAM running on the Windows 10 OS and the CUDA toolkit version 11.0 along with cuDNN v8.0. The quantitative results showed that our proposed framework was highly effective and outperformed the contemporary approaches by a substantial margin.

A. DATASET DESCRIPTION

The Canon EOS 600D is a versatile DSLR camera that offers a range of video features for research purposes. The camera provides Full high-dimensional (HD) 1080p video recording at 30 frames per second, which is ideal for capturing detailed footage. It also offers manual control over exposure, focus, and audio levels, which allows for greater control and precision in video recordings. Additionally, the camera has a built-in microphone and a jack for an external microphone, which is useful for capturing high-quality audio. One of the most significant video features of the Canon EOS 600D is its movie crop mode. This feature allows researchers to zoom in on subjects and capture footage with greater detail, even from a distance. The camera also has a video snapshot mode, which is useful for capturing short clips that can be combined into a longer video. Finally, the vari-angle LCD screen provides a clear view of the scene from different angles, making it easier to capture footage from unique perspectives. Overall, the Canon EOS 600D is an excellent camera for research purposes, offering a range of video features and manual controls for greater precision and control in video recordings. However, in this study, the Canon EOS 600D camera has been used for capturing anomaly activity in indoor environments. We made an anomaly detection dataset for eight different types of anomaly activities, including coughing, face mask, no mask, nose picking, sneezing, spitting, wrong mask, and yawning. Each anomaly activity consists of 18 videos of approximately one-minute duration. In this study, the face mask class is considered a normal activity, while the rest of the classes are considered anomalous activity. In Table 1, the statistical details of the presented dataset are given, while Fig. 3 shows a sample of normal and anomalous event frames taken from the dataset.

The CUHK-Avenue dataset was captured by an immovable surveillance camera with a 640 × 360 resolution [45]. The dataset includes 16 training video shorts that depict normal human behaviour and 21 test videos that highlight abnormal human behaviour. Normal actions include things such as strolling on the walkway and people congregating on the pavement. On the other hand, abnormal behaviour includes individuals wandering over the grassland, discarding stuff, lingering, and approaching the camera.

An article mentioned as [46] describes the UCSDPed1 dataset, which is made up of 14,000 images distributed throughout 70 video clips. It is separated into two sets: a test set with 36 video clips and a training set with 34. The dataset is a valuable baseline for unusual event identification

FIGURE 3. The dataset comprised a selection of frames representing both normal and abnormal events. These frames were carefully chosen to depict a range of activities captured, including instances of regular behavior, such as individuals correctly wearing masks, and instances of anomalies, such as individuals not adhering to mask-wearing protocols or engaging in activities like coughing, sneezing, and other prohibited actions. This selection of frames provides a comprehensive visual representation of the dataset, highlighting the diversity of behaviors and scenarios included in the study.

in surveillance footage since it contains 40 aberrant events, such as little carts, bicyclists, and micro trucks.

The UCSDPed2 dataset, which includes 4,560 images from 28 distinct videos, is available in [47]. It is divided into two sets: a training set of sixteen video footage and a test set comprising twelve video footage. The test set focuses on anomalous events, such as motorcycle riders. The dataset contains 12 examples of these specific types of events, making it a valuable resource for evaluating algorithms intended to detect anomalous bicycle-related incidents.

B. DATASET VISUALIZATION USING UMAP

UMAP is a recently developed manifold learning technique that seeks to efficiently integrate global structures and accurately depict local structures [42]. Compared with t-distributed Stochastic Neighbour Embedding (t-SNE), it provides plenty of features. It has been demonstrated

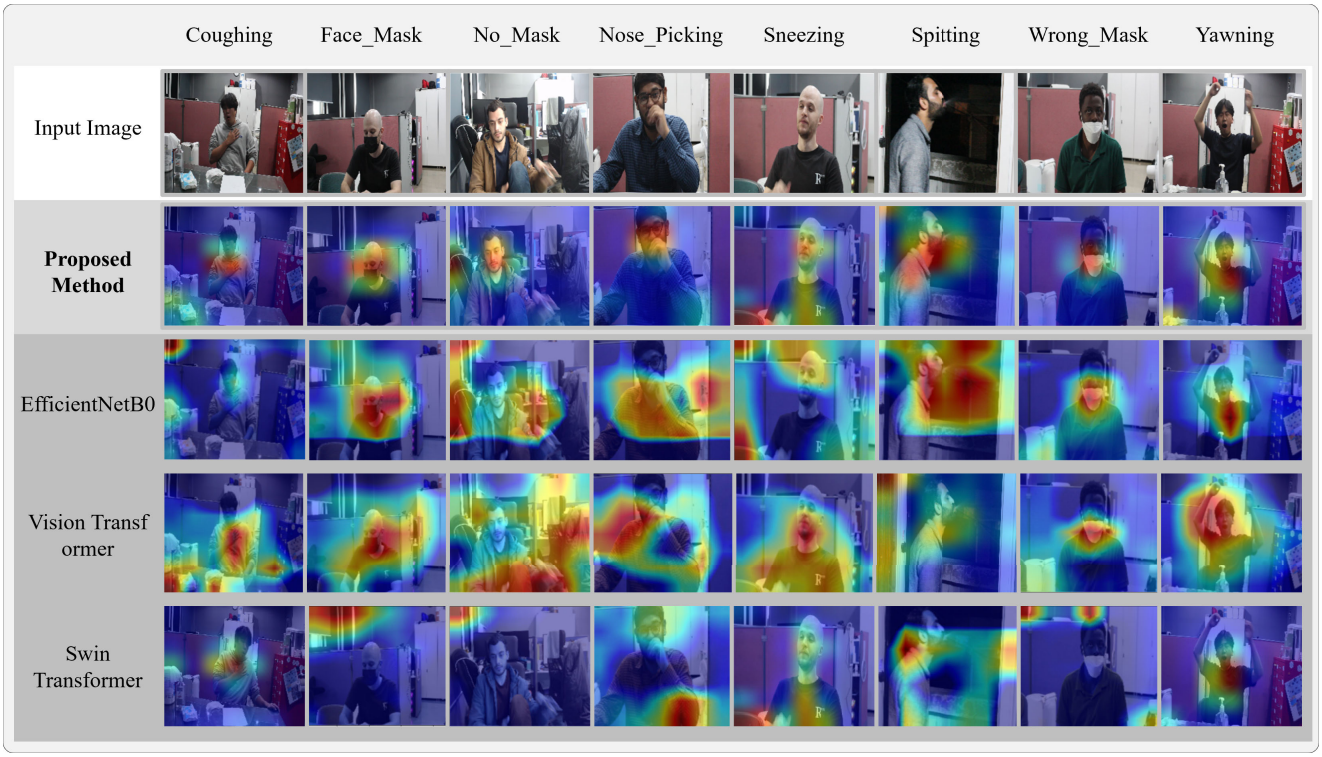

FIGURE 4. Visualization Results of Gradient-weighted Class Activation Mapping (Grad-CAM) [35]. We compare the proposed approach's (EfficientNet-B0 + CBAM) visual results with those of the baseline (EfficientNet-B0), Vision Transformer, and Swin Transformer. The last Conv layer outputs are used to compute the grad-CAM visualization. Every input image has the ground truth presented at the top.

TABLE 1. Statistical details of the proposed dataset.

Types of Anomalies	No. of videos	Training set	Testing set
Coughing	18	14	4
Face-Mask	18	14	4
No-Mask	18	14	4
Nose-picking	18	14	4
Sneezing	18	14	4
Spitting	18	14	4
Wrong-mask	18	14	4
Yawning	18	14	4
Total	144	112	32

TABLE 2. The statistical information about the datasets used for assessing video anomaly detection.

Dataset	# Videos	Train Set	Test Set	Length	Types of Anomalies
UCSDPed-1 [46]	70	34	36	$5min$	Bicycle riders, carts, small trucks, and etc.
UCSDPed-2 [47]	28	16	12	$5min$	Bicycle riders
Avenue [45]	38	16	21	$5min$	Run, throw, and find an unfamiliar object

that UMAP performs better with large datasets than t-SNE. Fig. 6 shows the outlined dataset's 3D-UMAP mapping. The underlying hypotheses of UMAP are as follows:

- The data is distributed uniformly over a Riemannian manifold.
- The Riemannian metrics do not change locally.
- There is local connectivity on the manifold.

The manifold may be represented as an HD fuzzy topology of samples under these assumptions. By seeking a fuzzy topology, an embedded manifold is found using the low-dimensional (LD) estimations. UMAP has generated the fuzzy topology pattern by displaying the sample points on the HD graph. An edge's weight indicates the likelihood that the two locations are connected in the resulting HD graph, thus forming a weighted graph. UMAP employed an exponential distribution of probabilities to compute the similarity between HD sample points.

$$P_{i/j} = \exp\left(-\frac{d(x_i, x_j) - \sigma_i}{\sigma_i}\right) \tag{10}$$

$d(x_i, x_j)$ represents the difference between the i^{th} and j^{th} data points, whereas σ represents the difference between the i^{th}

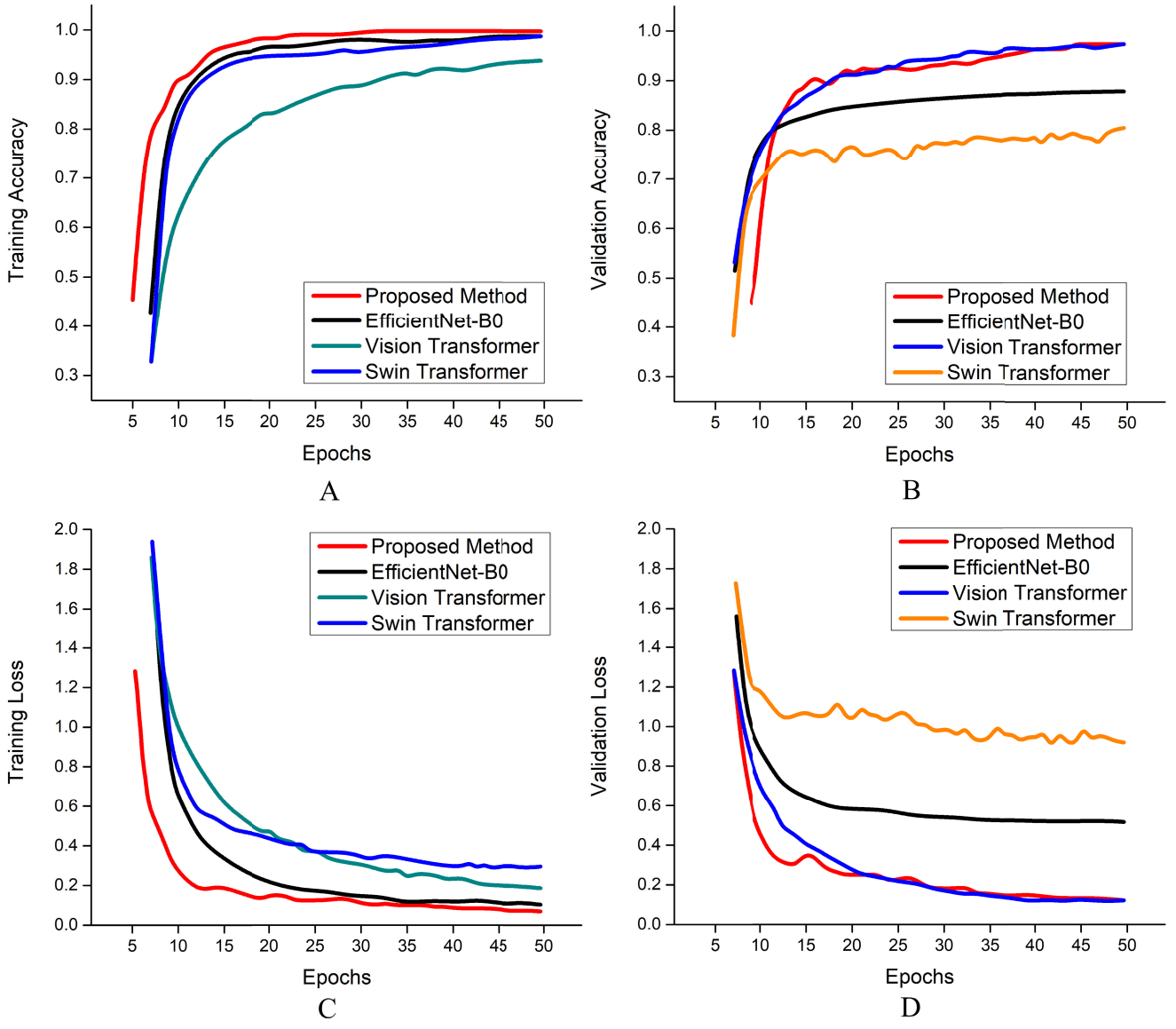

FIGURE 5. The training and validation accuracy graphs of the proposed strategy utilizing our own dataset can be seen in (A) and (B), while training and validation loss graphs are plotted in (C) and (D), respectively. The accuracy and loss graphs are plotted against the number of epochs.

data point and its initial nearest neighbour. In UMAP, the HD probability of the data is symmetric:

$$P_{i/j} = P_{(i/j)} + P_{(j/i)} - P_{(i/j)}P_{(j/i)} \tag{11}$$

As mentioned before, the resulting graph is a likelihood graph, and UMAP needs to find the k^{th} nearest neighbours:

$$k = 2^{\sum_i P_{ij}} \tag{12}$$

As close to the original HD graph as possible is produced by UMAP through the generation and optimization of an LD variant. For LD distance modelling, UMAP adopts a probability measure:

$$q_{ij} = \left(1 + \alpha(y_i - y_j)^{2b}\right)^{-1} \tag{13}$$

By default, $\alpha \approx 1.93$ and $b \approx 0.79$ for UMAP. UMAP employs Binary Cross-Entropy as a loss function due to its ability to obtain the global data pattern:

$$CE(P, Q) = \sum_i \sum_j P_{ij} \log\left(\frac{P_{ij}}{Q_{ij}}\right) + (1 - P_{ij}) \log\left(\frac{1 - P_{ij}}{1 - Q_{ij}}\right) \tag{14}$$

C. EVALUATION METRICS

Many metrics are frequently employed to assess the performance and efficacy of the suggested frameworks. These measures include the F1-score, recall (rec), precision (prec), and classification accuracy (acc). The four fundamental

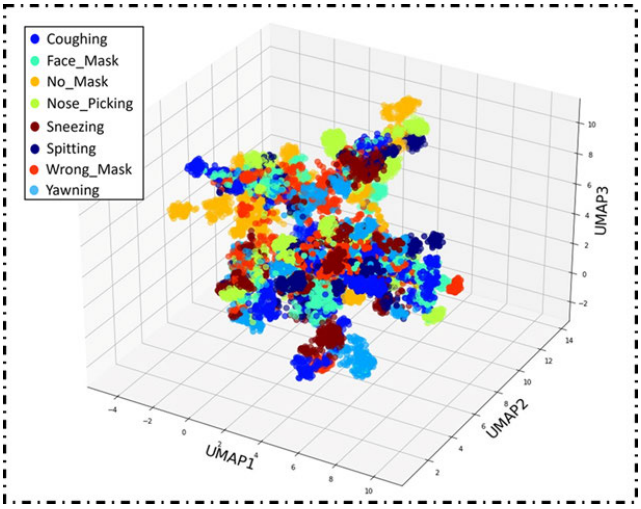

FIGURE 6. The 3D-UMAP Projection of the presented dataset.

outcomes-True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)-are the basis for all of these measures. These results serve as the foundation for calculating various evaluation measures. For instance, *prec* indicates the frequency of positive predictions amongst all positive outcomes, *rec* indicates the percentage of successfully predicted positive instances, and *acc* indicates the proportion of correctly predicted data samples to the total amount of data samples. Last, the harmonic mean of *prec* and *rec* yields the F1-score. The following provides the mathematical expressions for these measures:

$$prec = \frac{TP}{TP + FP} \tag{15}$$

$$rec = \frac{TP}{TP + FN} \tag{16}$$

$$F1 - Score = 2x\left(\frac{Prec \times Rec}{Prec + Rec}\right) \tag{17}$$

$$acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

D. ABLATION STUDY

An ablation study was conducted on the proposed architecture to evaluate the impact of the CBAM module on the model's performance in anomaly detection tasks. Specifically, we evaluated the impact of the CBAM module on the model's ability to extract features and classify different anomaly classes. The first modified version of the model was created by removing the SE module from the EfficientNet-B0 architecture. The results showed that the removal of the SE module decreased the accuracy of the model from 87 to 83%, indicating that the SE module was an essential component in the original EfficientNet-B0 architecture. The second modified version of the model was created by replacing the CBAM module with other attention mechanisms, such as the SE module and the non-local module. The results showed that the CBAM module outperformed the other attention mechanisms, indicating that the CBAM module was more effective in improving the model's ability to extract features and classify different anomaly classes. The third modified version of the model was created by embedding the CBAM module at different layers of the EfficientNet-B0 architecture. The results showed that the presented framework improved the accuracy of the model from 87 to 96%, indicating that the CBAM module was effective in sharpening the extracted features and optimizing the model's generalization.

The integration of the CBAM method improved accuracy as compared to the SE module because of the CBAM's unique focus on strengthening the spatial details of the data. While the SE module mainly concentrates on adjusting channel-wise characteristics by altering the weight of distinct channels, the CBAM module incorporates SA methods to enhance the process. SA enables the model to selectively focus on or ignore data from various spatial locations within each channel. CBAM makes it possible for the neural network to identify fine features, dense patterns, and spatial correlations in the data. This spatial knowledge is especially useful in image-related activities where comprehending the spatial context is critical for efficient detection and classification. The ability of the CBAM module to pay attention to these spatial features becomes extremely useful in circumstances where spatial characteristics are important for class differentiation or accurate prediction. The SAM in CBAM improves the model's capacity to identify and utilize spatial relationships, resulting in higher accuracy compared to the SE module.

E. NETWORK VISUALIZATION WITH GRAD-CAM

Grad-CAM is a technique used to visualize the regions of an input image that contributed most to the final prediction made by a deep learning model [35]. When combining CBAM with the EfficientNet-B0 architecture, Grad-CAM can provide valuable insights into the areas of an image that are most important for classification and decision-making. To generate a Grad-CAM visualization for the proposed model, the gradient of the predicted class score with respect to the feature maps is calculated. This gradient is then used to weight the feature maps, producing a heatmap that highlights the regions of the image that contributed most to the prediction. The resulting heatmap can be overlaid on the input image to provide a visual representation of the areas of the image that were most important for classification. This information can be used to identify important features or patterns in the data that may be relevant for further analysis. Using Grad-CAM in combination with CBAM-EfficientNet can provide valuable insights into the decision-making process of the model and the specific features or patterns that are most important for classification. This information can be useful in a variety of applications, from medical image analysis to object detection and classification in computer vision tasks. Fig. 4 represents the Grad-CAM Visualization of the proposed method with respect to other methods.

TABLE 3. In the context of accuracy, the suggested framework is compared to present fine-tuned approaches.

Methods	Classification Accuracy
EfficientNet-Bo	87%
Vision Transformer	91%
Swin Transformer	92%
Proposed Framework	**96%**

TABLE 4. Performance evaluation of the suggested approach with contemporary approaches using the proposed dataset for anomaly detection in surveillance videos.

Classifiers	Evaluation Metrics	Classes							
		Coughing	Face-Mask	No-Mask	Nose-Picking	Sneezing	Spitting	Wrong-Mask	Yawning
EfficientNet-Bo	Prec	0.84	0.92	0.80	0.80	0.84	1.00	0.85	0.89
	Rec	0.89	0.82	0.79	0.85	0.96	1.00	0.79	0.84
	F1-Score	0.86	0.87	0.79	0.82	0.90	1.00	0.82	0.87
Vision Transformer	Prec	0.91	0.97	0.91	0.83	0.85	1.00	0.91	0.86
	Rec	0.89	0.95	0.86	0.79	0.96	1.00	0.87	0.91
	F1-Score	0.90	0.96	0.89	0.81	0.90	1.00	0.89	0.89
Swin Transformer	Prec	0.94	0.96	0.93	0.82	0.89	1.00	0.96	0.82
	Rec	0.96	0.94	0.85	0.79	0.98	1.00	0.84	0.95
	F1-Score	0.95	0.95	0.88	0.80	0.93	1.00	0.89	0.88
Our Method	**Prec**	**0.96**	**0.96**	**0.98**	**0.97**	**0.92**	**1.00**	**0.91**	**0.97**
	Rec	**0.94**	**0.96**	**0.91**	**0.93**	**0.99**	**1.00**	**0.95**	**0.98**
	F1-Score	**0.95**	**0.96**	**0.94**	**0.95**	**0.95**	**1.00**	**0.93**	**0.97**

F. COMPARISON WITH CONTEMPORARY TECHNIQUES

In this study, three different methods were compared for anomaly detection in surveillance videos, as seen in Table 3. The first method used an EfficientNet-Bo model [27], which achieved an accuracy of 87% on the authors' own anomaly dataset. The second method employed a Vision Transformer [43] and achieved an accuracy of 91%. The third method employed a Swin Transformer [44] and achieved an accuracy of 92%. However, the proposed system outperformed all three methods, achieving an accuracy of 96%. This significant improvement in accuracy suggests that the proposed system is more effective at identifying anomalies than the other methods tested. It should be noted that while EfficientNet-Bo, Vision Transformer, and Swin Transformer have been shown to be effective in other image classification tasks, their performance on anomaly detection may vary based on the anomaly types being analyzed. Table 4 compares the classification performance of the presented framework with contemporary existing methods. The effectiveness of the proposed framework across all classes can be seen by looking at Table 4. However, the EfficientNet-B0 method reveals inadequate results, particularly in the class of No-Mask. The proposed technique performs exceptionally well, yielding high precision, recall, and F1-scores for all classes. Fig. 5(A) and (B) depict the training and validation accuracy graphs for the proposed model using our own Dataset, while Fig. 5(C) and (D) show the training and validation loss graph, respectively. Overall, these results suggest that the proposed system is a promising approach to anomaly detection, and further research is warranted to explore its potential for other types of datasets and anomalies.

The comparison study covering three standard datasets (UCSDPed-1, UCSDPed-2, and Avenue) shows that the technique we suggested outperforms various contemporary methods in pedestrian anomaly recognition, as shown in Table 5. Our technique outperforms Tang et al. (96.2%) and Yiwei et al. (96.0%) on UCSDPed-2, with a maximum accuracy of 97%, demonstrating its strength in identifying anomalies in surroundings with better clarity. Our solution obtains an impressive accuracy of 94% using the highly complicated Avenue dataset, exceeding approaches like Zhou et al. (86.1%) and Che et al. (89.6%) by a wide margin. Even though UCSDPed-1 proves more difficult, our approach still performs competitively with 85%, surpassing models such as Radu et al. (68.4%) and Weixin et al. (75.5%) but slightly lagging behind Lu et al. (91.8%). These findings show that our approach is accurate, scalable, and efficient on a variety of datasets. They also illustrate how effectively our approach generalizes to both basic and complicated pedestrian situations, which makes it an appealing choice for anomaly detection applications in real-life settings.

G. LIMITATIONS, ASSUMPTIONS, AND DEPENDENCIES

Hardware Limitations: The performance of the proposed framework could be affected by several hardware limitations. First, computational resources play an active role in ensuring that deep learning models perform effectively. However, EfficientNet-B0 is considered a lightweight model, but processing high-resolution videos or large datasets can need powerful GPUs. Next, the performance of the system is limited by the quality of the camera used during the data generation phase. Low-resolution cameras may struggle

TABLE 5. Our proposed method's accuracy contrasts with recent methods on the UCSDPed-1 [46], UCSDPed-2 [47], and Avenue [45] datasets.

Methods	UCSDPed-1 [46]	UCSDPed-2 [47]	Avenue [45]
Radu *et al.* [48]	68.4	82.2	80.6
Zhou *et al.* [49]	83.5	94.9	86.1
Che *et al.* [50]	—	—	89.6
Weixin *et al.* [51]	75.5	88.1	77.0
Lu *et al.* [45]	91.8	—	80.9
Tang *et al.* [52]	82.6	96.2	83.7
Qiang *et al.* [53]	85.2	—	85.8
Wen *et al.* [54]	83.1	95.4	85.1
Ramachandra *et al.* [55]	77.3	88.3	72.0
Yiwei *et al.* [56]	86.2	96.0	85.7
Proposed Method	**85**	**97**	**94**

to gather enough information to detect fine anomalies like coughing, sneezing, or complex behavioural features. Furthermore, memory constraint is an additional challenge. In this study, the proposed method integrates the CBAM attention module, which needed extra memory during the training and testing phases. As a result, this could limit the deployment of the proposed system on resource-constrained devices such as edge computing platforms. To ensure the scalability and reliability of the proposed system across different conditions these hardware limitations should be addressed.

Architectural Assumptions: Several assumptions have been made in the development and management of the suggested system. The first assumption is model generalization, which means that the dataset used to train the model is truly representative of real-life situations. If the data set does not cover a broad range of inputs, or if certain anomalous behaviours are omitted from the dataset, then the model is unlikely to generalize properly and perform well in other scenarios. Moreover, the system supposes that the attention mechanism, particularly the incorporation of the CBAM module, will improve the feature extraction of the EfficientNet-B0. However, this assumption may not be true in practice since the use of attention mechanisms may not be as effective as anticipated within a proposed system, especially depending on the type of anomalies being identified. For instance, the channel and SA may not be equally effective across various types of anomalies. Furthermore, the current system does not incorporate temporal relationships between frames and is mainly based on the spatial domain. This assumption may minimize the effectiveness of the proposed framework to detect anomalies in temporal sequences. An effective method may require a temporal features extractor to capture such dependencies.

Dependencies: The performance of the proposed framework depends upon several external factors. First, a balanced and diverse dataset is needed to improve the model's performance, while a limited and unbalanced dataset may cause overfitting issues or poor generalization. Next, external factors like illumination, occlusion, and background noise may affect the accuracy of the system. Variations in these factors may compromise the applicability of the proposed framework in a real-world environment. Furthermore, the pre-trained model should be accessible, specifically the pre-trained weight of the EfficientNet-B0. However, these weights enable the process of faster training; these weights may need specific licenses for commercial use, which slow down the deployment process. Finally, despite the automation of anomaly detection, there is still a need for human intervention. In some situations, there is a need for human intervention to confirm the type of detected anomalies in complex and challenging environments. The above dependencies should be addressed to ensure the success of proposed frameworks in practical applications.

H. ADVANTAGES AND DISADVANTAGES OF ADVANCED METHODS IN VIDEO ANOMALY DETECTION

The contemporary models in anomaly detection, including EfficienctNet-B0, Vision Transformer, and Swin Transformer, have some advantages and disadvantages. EfficienctNet-B0 is known for its scalability and lightweight architecture, making it suitable for training and deployment on resource-constrained devices. Moreover, it achieves good accuracy on the image classification tasks. In contrast, the Vision Transformer performs well on anomaly detection tasks and achieves an accuracy of 91% on the proposed dataset. Likewise, the Swin Transformer gains high accuracy (92%) compared to the Vision Transformer and shows its potential in anomaly detection tasks.

However, there are some drawbacks to these methods. It is not possible for initial EfficientNet-B0 to preserve salient spatial information. This spatial feature information plays an important role in precise video anomaly detection. Furthermore, the Vision Transformer and Swin Transformer needed more computational resources during the training and deployment compared to the lightweight EfficientNet-B0 model. This paper addresses these gaps by embedding the CBAM module at different layers of the EfficientNet-B0 architecture. The outcomes demonstrated that the suggested framework increased the model's accuracy from 87 to 96%, proving the CBAM module's potency in polishing the extracted traits while further improving the model's generalisability.

V. CONCLUSION AND FUTURE DIRECTION

In this paper, we introduced a novel approach for detecting anomalies in surveillance videos, with a focus on behaviours that deviate from public health guidelines during the

pandemic, including the absence of face masks, improper mask usage, coughing, nose-picking, sneezing, spitting, and yawning. The manual monitoring of such anomalies is challenging and error-prone, necessitating automated solutions. Our approach leverages a multi-attention-based deep learning system built upon the EfficientNet-B0 architecture. In addition, EfficientNet-B0, known for its MBConv block with SE modules, excels in highlighting informative channel characteristics. However, it falls short of retaining essential spatial information for visual recognition tasks. To address this limitation, we introduced the CBAM module into the original EfficientNet-B0 model, thereby enhancing its feature extraction capabilities. Furthermore, an ablation study was conducted to assess the impact of the CBAM module on the model's anomaly detection performance. We compared different versions of the model, including those with and without the SE module and alternative attention mechanisms like the non-local module. The results consistently demonstrated the superiority of the CBAM module in enhancing feature extraction and classifying various anomaly classes. In particular, embedding the CBAM module after the second convolutional layer significantly improved accuracy, from 87 to 96%, underscoring its effectiveness. We compared our proposed system with three alternative methods for anomaly detection in surveillance videos. While EfficientNet-B0, Vision Transformer, and Swin Transformer showed promise in other image classification tasks, the proposed system outperformed them all, achieving an accuracy of 96%. This substantial accuracy improvement suggests the efficacy of our approach in identifying anomalies. We also evaluated the proposed framework against existing methods, demonstrating its exceptional performance across all anomaly classes, with particularly noteworthy results in the ''No-Mask'' category. Even though our study has made significant strides in anomaly detection, we recognize the potential for further enhancements in real-time precision and efficacy. In our future work, we intend to implement a spatial-temporal attention block to refine our architecture's performance. Additionally, we plan to expand our experiments and gather more comprehensive data. By evaluating our proposed method on a wider range of datasets, we can assess its performance across different scenarios and validate its generalizability. Moreover, we will compare our approach to various existing methods, including both traditional and state-of-the-art techniques. This comparative analysis will help us understand the specific advantages and limitations of our method concerning others. Furthermore, we will conduct sensitivity analyses to investigate the impact of different parameter settings, dataset variations, and potential sources of noise. This analysis will provide insights into the stability and robustness of our approach, ensuring its reliability in real-world situations.

REFERENCES

[1] (2023). *Coronavirus (Covid-19) Dashboard, Report of World Health Organization*. [Online]. Available: https://covid19.who.int/

[2] S. Srinivasan, R. Rujula Singh, R. R. Biradar, and S. Revathi, ''COVID-19 monitoring system using social distancing and face mask detection on surveillance video datasets,'' in *Proc. Int. Conf. Emerg. Smart Comput. Informat. (ESCI)*, Mar. 2021, pp. 449–455.

[3] T. Kumar, ''An evidence review of face masks against COVID-19,'' *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 9, no. 12, pp. 919–923, Dec. 2021.

[4] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, ''Artificial intelligence (AI) applications for COVID-19 pandemic,'' *Diabetes Metabolic Syndrome, Clin. Res. Rev.*, vol. 14, no. 4, pp. 337–339, 2020.

[5] T. Li, Y. Liu, M. Li, X. Qian, and S. Y. Dai, ''Mask or no mask for COVID-19: A public health and market study,'' *PLoS ONE*, vol. 15, no. 8, Aug. 2020, Art. no. e0237691.

[6] N. R. Longrich and S. K. Sheppard. (2020). *Public Use of Masks to Control the Coronavirus Pandemic*. [Online]. Available: https://doi.org/10.20944/preprints202004.0021.v1

[7] B. Qin and D. Li, ''Identifying facemask-wearing condition using image super-resolution with classification network to prevent COVID-19,'' *Sensors*, vol. 20, no. 18, p. 5236, Sep. 2020.

[8] S. Bhattacharya, P. K. R. Maddikunta, Q.-V. Pham, T. R. Gadekallu, C. L. Chowdhary, M. Alazab, and M. J. Piran, ''Deep learning and medical image processing for coronavirus (COVID-19) pandemic: A survey,'' *Sustain. Cities Soc.*, vol. 65, Feb. 2021, Art. no. 102589.

[9] A. M. Rahmani and S. Y. H. Mirmahaleh, ''Coronavirus disease (COVID-19) prevention and treatment methods and effective parameters: A systematic literature review,'' *Sustain. Cities Soc.*, vol. 64, Jan. 2021, Art. no. 102568.

[10] S. V. Shetty, K. Anand, S. Pooja, K. A. Punnya, and M. Priyanka, ''Social distancing and face mask detection using deep learning models: A survey,'' in *Proc. Asian Conf. Innov. Technol. (ASIANCON)*, Aug. 2021, pp. 1–6.

[11] S. Ge, J. Li, Q. Ye, and Z. Luo, ''Detecting masked faces in the wild with LLE-CNNs,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 426–434.

[12] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, ''Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection,'' *Sustain. Cities Soc.*, vol. 65, Feb. 2021, Art. no. 102600.

[13] M. Rezaei and M. Azarmi, ''DeepSOCIAL: Social distancing monitoring and infection risk assessment in COVID-19 pandemic,'' *Appl. Sci.*, vol. 10, no. 21, p. 7514, Oct. 2020.

[14] Z. Ji, Y. Wang, K. Yan, X. Xie, Y. Xiang, and J. Huang, ''A space-embedding strategy for anomaly detection in multivariate time series,'' *Expert Syst. Appl.*, vol. 206, Nov. 2022, Art. no. 117892.

[15] M. Hu, X. Feng, Z. Ji, K. Yan, and S. Zhou, ''A novel computational approach for discord search with local recurrence rates in multivariate time series,'' *Inf. Sci.*, vol. 477, pp. 220–233, Mar. 2019.

[16] S. U. Amin, Y. Kim, I. Sami, S. Park, and S. Seo, ''An efficient attention-based strategy for anomaly detection in surveillance video,'' *Comput. Syst. Sci. Eng.*, vol. 46, no. 3, pp. 3939–3958, 2023.

[17] K. O'Dowd, K. M. Nair, P. Forouzandeh, S. Mathew, J. Grant, R. Moran, J. Bartlett, J. Bird, and S. C. Pillai, ''Face masks and respirators in the fight against the COVID-19 pandemic: A review of current materials, advances and future perspectives,'' *Materials*, vol. 13, no. 15, p. 3363, Jul. 2020.

[18] J. Leng, Q. Wang, and K. Liu, ''Sustainable design of courtyard environment: From the perspectives of airborne diseases control and human health,'' *Sustain. Cities Soc.*, vol. 62, Nov. 2020, Art. no. 102405.

[19] S. U. Amin, S. Taj, A. Hussain, and S. Seo, ''An automated chest X-ray analysis for COVID-19, tuberculosis, and pneumonia employing ensemble learning approach,'' *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105408.

[20] A. Hussain, S. U. Amin, H. Lee, A. Khan, N. F. Khan, and S. Seo, ''An automated chest X-ray image analysis for COVID-19 and pneumonia diagnosis using deep ensemble strategy,'' *IEEE Access*, vol. 11, pp. 97207–97220, 2023.

[21] N. S. Punn, S. K. Sonbhadra, S. Agarwal, and G. Rai, ''Monitoring COVID-19 social distancing with person detection and tracking via fine-tuned YOLO v3 and deepsort techniques,'' 2020, *arXiv:2005.01385*.

[22] L. Ramadass and S. Arunachalam, ''Applying deep learning algorithm to maintain social distance in public place through drone technology,'' *Int. J. Pervasive Comput. Commun.*, vol. 16, no. 3, pp. 223–234, Jun. 2020.

[23] S. Teboulbi, S. Messaoud, M. A. Hajjaji, and A. Mtibaa, ''Real-time implementation of AI-based face mask detection and social distancing measuring system for COVID-19 prevention,'' *Sci. Program.*, vol. 2021, pp. 1–21, Sep. 2021.

[24] W. Bu, J. Xiao, C. Zhou, M. Yang, and C. Peng, "A cascade framework for masked face detection," in *Proc. IEEE Int. Conf. Cybern. Intell. Syst. (CIS), IEEE Conf. Robot., Autom. Mechatronics (RAM)*, Nov. 2017, pp. 458–462.
[25] I. Ahmed, M. Ahmad, and G. Jeon, "Social distance monitoring framework using deep learning architecture to control infection transmission of COVID-19 pandemic," *Sustain. Cities Soc.*, vol. 69, Jun. 2021, Art. no. 102777.
[26] B. Limbasiya and C. Raut, "COVID-19 face mask and social distancing detector using machine learning," *Int. Res. J. Eng. Technol.*, vol. 8, no. 5, pp. 2056–2061, 2021.
[27] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.
[28] H. Khan, M. Ullah, F. Al-Machot, F. A. Cheikh, and M. Sajjad, "Deep learning based speech emotion recognition for Parkinson patient," *Electron. Imag.*, vol. 35, no. 9, pp. 1–6, Jan. 2023.
[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[31] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.
[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[33] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 4278–4284.
[34] H. Khan, T. Hussain, S. Ullah Khan, Z. Ahmad Khan, and S. W. Baik, "Deep multi-scale pyramidal features network for supervised video summarization," *Expert Syst. Appl.*, vol. 237, Mar. 2024, Art. no. 121288.
[35] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, vol. 11211, Sep. 2018, pp. 3–19.
[36] M. Munsif, S. U. Khan, N. Khan, and S. W. Baik, "Attention-based deep learning framework for action recognition in a dark environment," *Hum.-Centric Comput. Inf. Sci.*, vol. 14, pp. 1–21, Jan. 2024.
[37] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.
[38] Y. Yu and F. Liu, "Effective neural network training with a new weighting mechanism-based optimization algorithm," *IEEE Access*, vol. 7, pp. 72403–72410, 2019.
[39] W. E. L. Ilboudo, T. Kobayashi, and K. Sugimoto, "Robust stochastic gradient descent with student-t distribution based first-order momentum," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 3, pp. 1324–1337, Mar. 2022.
[40] D. Cheng, S. Li, H. Zhang, F. Xia, and Y. Zhang, "Why dataset properties bound the scalability of parallel machine learning training algorithms," *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 7, pp. 1702–1712, Jul. 2021.
[41] S. Feng and M. F. Duarte, "Few-shot learning-based human activity recognition," *Expert Syst. Appl.*, vol. 138, Dec. 2019, Art. no. 112782.
[42] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.
[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[44] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[45] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in Matlab," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2720–2727.
[46] A. B. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 909–926, May 2008.
[47] A. B. Chan and N. Vasconcelos, "Counting people with low-level features and Bayesian regression," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2160–2177, Apr. 2011.
[48] R. T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2914–2922.
[49] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "AnomalyNet: An anomaly detection network for video surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2537–2550, Oct. 2019.
[50] C. Sun, Y. Jia, Y. Hu, and Y. Wu, "Scene-aware context reasoning for unsupervised abnormal event detection in videos," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 184–192.
[51] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 439–444.
[52] Y. Tang, L. Zhao, S. Zhang, C. Gong, G. Li, and J. Yang, "Integrating prediction and reconstruction for anomaly detection," *Pattern Recognit. Lett.*, vol. 129, pp. 123–130, Jan. 2020.
[53] Y. Qiang, S. Fei, and Y. Jiao, "Anomaly detection based on latent feature training in surveillance scenarios," *IEEE Access*, vol. 9, pp. 68108–68117, 2021.
[54] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.
[55] B. Ramachandra and M. J. Jones, "Street scene: A new dataset and evaluation protocol for video anomaly detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2558–2567.
[56] Y. Lu, K. M. Kumar, S. S. Nabavi, and Y. Wang, "Future frame prediction using convolutional VRNN for anomaly detection," in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.

SAREER UL AMIN received the bachelor's degree in computer science from the Islamia College University Peshawar, Pakistan, and the master's degree in computer science and engineering from Chung-Ang University, Seoul, South Korea. He is currently a Research Associate with Chung-Ang University, where he is actively engaged in cutting-edge research projects in the field of computer science. During his undergraduate studies, he was a Research Assistant with the Digital Image Processing Laboratory (DIP), Islamia College University Peshawar, where he gained valuable experience in research methodologies and techniques. His research interests include medical image analysis, anomaly detection in surveillance videos, human actions and activity recognition, sequence learning, and image and video analytics. He is particularly passionate about leveraging deep learning methodologies for multimedia understanding and interpretation. With a strong academic background and a diverse research portfolio, he is dedicated to advancing the frontiers of computer science and engineering. He strives to make meaningful contributions to the field through his research endeavors and looks forward to furthering his career in academia and industry.

MUHAMMAD SIBTAIN ABBAS received the bachelor's degree in computer science from the University Institute of Information Technology (UIIT), Arid Agriculture University, Pakistan. He is currently pursuing the master's degree with Chung Ang University. His keen research interests include machine/deep learning, anomaly detection, activity recognition, the IoT, and smart applications for construction management and worker safety.

BUMSOO KIM received the B.S. degree in art and technology from Chung-Ang University, South Korea, in 2023, where he is currently pursuing the M.S. degree in applied art and technology. He was an Artificial Intelligence Researcher with Vive Studios, South Korea. His research interests include style transfer, face stylization/cartoonization, face re-aging, and face swapping.

YONGHOON JUNG received the B.S. degree in computer engineering from Sungkyul University, in 2022. He is currently pursuing the M.S. degree with the Graduate School of Advanced Imaging Science, Multimedia & Film, Chung-Ang University, specializing in entertainment technology. His research interests include artificial intelligence and computer vision, with an emphasis on synthetic data generation and domain adaptation techniques. He aims to apply these cutting-edge technologies to solve complex issues in the real world. His dedication to his field is evident as he continues to explore innovative solutions to enhance technological applications.

SANGHYUN SEO received the B.S. degree in computer science and engineering from Chung-Ang University, Seoul, South Korea, in 1998, and the M.S. and Ph.D. degrees from the GSAIM Department, Chung-Ang University, in 2000 and 2010, respectively. He was a Senior Researcher with G-Inno Systems, from 2002 to 2005. He was a Postdoctoral Researcher with Chung-Ang University, in 2010, and the LIRIS Laboratory, Lyon 1 University, from February 2011 to February 2013. He was with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, from May 2013 to February 2016. He was also with Sungkyul University, from March 2016 to February 2019. He is currently a Faculty Member with the College of Art and Technology, Chung-Ang University. His research interests include computer graphics, non-photorealistic rendering and animation, real-time rendering using GPU, VR/AR, and game technology. He has been a program committee member of many international conferences and workshops. He has been a Reviewer of *Multimedia Tools and Applications* (MTAP), *Computers & Graphics* (Elsevier), U.K., the *Journal of Supercomputing* (JOS), and *The Visual Computer* (Springer). He has edited a number of international journal special issues as a Guest Editor, such as the *Journal of Real-Time Image Processing*, the *Journal of Internet Technology*, and *Multimedia Tools and Applications*. He has been an Associate Editor of the *Journal of Real-Time Image Processing*, since 2017.