


AONet: Attention network with optional activation for unsupervised video anomaly detection

Akhrorjon Akhmadjon Ugli Rakhmonov | Barathi Subramanian |
Bahar Amirian Varnousefaderani | Jeonghong Kim 

Department of Computer Science and Engineering, Kyungpook National University, Daegu, Republic of Korea

Correspondence

Jeonghong Kim, Department of Computer Science and Engineering, Kyungpook National University, Daegu, Republic of Korea.

Email: jhk@knu.ac.kr

Funding information

BK21 FOUR Project (AI-driven Convergence Software Education Research Program) Funded by the Ministry of Education, Department of Computer Science and Engineering, Kyungpook National University, Daegu, Republic of Korea, Grant/Award Number: 4120240214871; Basic Science Research Program of the National Research Foundation of Korea (NRF) Funded by the Ministry of Education, Republic of Korea, Grant/Award Number: 2021R1I1A3043970

Abstract

Anomaly detection in video surveillance is crucial but challenging due to the rarity of irregular events and ambiguity of defining anomalies. We propose a method called AONet that utilizes a spatiotemporal module to extract spatiotemporal features efficiently, as well as a residual autoencoder equipped with an attention network for effective future frame prediction in video anomaly detection. AONet utilizes a novel activation function called OptAF that combines the strengths of the ReLU, leaky ReLU, and sigmoid functions. Furthermore, the proposed method employs a combination of robust loss functions to address various aspects of prediction errors and enhance training effectiveness. The performance of the proposed method is evaluated on three widely used benchmark datasets. The results indicate that the proposed method outperforms existing state-of-the-art methods and demonstrates comparable performance, achieving area under the curve values of 97.0%, 86.9%, and 73.8% on the UCSD Ped2, CUHK Avenue, and ShanghaiTech Campus datasets, respectively. Additionally, the high speed of the proposed method enables its application to real-time tasks.

KEYWORDS

activation function, convolutional neural network, loss function, unsupervised learning, video anomaly detection

1 | INTRODUCTION

In computer vision, anomaly detection in the context of video surveillance has attracted considerable attention because it enables numerous applications, including monitoring traffic accidents and detecting unauthorized or criminal activities [1, 2]. However, this task is difficult for several reasons. First, gathering and annotating various abnormal events is challenging because of their rarity compared with normal events [3, 4]. Second, the

nature of what defines an “anomaly” is ambiguous. Activities can be deemed anomalous or normal depending on specific circumstances. For example, observing a person placing a package near someone’s front door in a uniform during the daytime may be considered a normal activity, as they are presumed to be making a regular delivery. However, if someone wearing casual clothing performs the same activity late at night, it may be perceived as anomalous. Therefore, recent research on video anomaly detection (VAD) has shifted toward

unsupervised methods that eliminate the need for human annotation [5, 6].

Some existing studies have employed dual-stream networks to detect activities that deviate from expected behaviors [7–9]. This approach divides video analysis into two distinct pathways: spatial and temporal. A spatial stream processes individual video frames to capture static visual content using convolutional neural networks (CNNs). The temporal stream focuses on the motion between frames, typically using optical flow or similar motion descriptors. This dual approach allows the resulting system to detect anomalies in both content and movement concurrently within videos. However, computing optical flow is resource intensive. An alternative method employs recurrent neural networks (RNNs) such as variational long short-term memory (LSTM) [1, 9–11] to capture temporal movement data. However, adding additional layers to a model significantly increases its complexity [12, 13].

Recent studies [14–16] have employed future frame prediction as a concept for anomaly detection. The objective of future frame models is to predict future frames based on several previous frames. An anomaly is identified when there is a substantial discrepancy between the predicted and actual future frames. Activation functions (AFs) play a pivotal role in future frame reconstruction models because they enable such models to capture and represent complex patterns and relationships within data. Therefore, researchers have focused on merging AFs to extract more insights from input data [17, 18].

Within the machine learning paradigm, attention mechanisms emulate human focus by emphasizing certain input sections such as specific objects while overlooking others [19]. In the scope of VAD, some researchers [20] have demonstrated that concentrating attention on the foreground, especially when dynamic objects are in motion, while ignoring the static background improves performance. Additionally, the use of a combination of loss functions in VAD models enhances model performance when detecting both spatial and temporal anomalies in video data. Various loss functions highlight the diverse aspects of prediction errors, improve model robustness, and allow for tailored solutions to complex detection tasks [21, 22].

To address the limitations described above, we propose an attention network with optional activation (AONet) that employs an autoencoder (AE) structure for efficient anomaly detection and combines the advantages of several robust AFs within its architecture for effective feature extraction. The adjustable ability of the optional AF (OptAF) in the proposed model architecture, which combines the benefits of the ReLU [23], leaky ReLU [24], and sigmoid [25] AFs, enables the proposed model to

learn the critical features of complex datasets utilizing parameterized adaptation. AONet seeks to harness spatial and temporal characteristics in an integrated manner. The features derived from a pre-trained CNN are directed into two branches to capture spatial patterns and movement attributes. Subsequently, these spatiotemporal attributes are fed into a decoder for future frame prediction. The decoder in the proposed model has a bottom-up channel attention module that enables it to leverage the interrelations between feature channels effectively. These design choices and novel components enable the proposed method to detect anomalies with high accuracy and speed. The major contributions of this study can be summarized as follows.

- An unsupervised learning method is employed by leveraging the future frame prediction approach, eliminating the need for labor-intensive annotation.
- We propose a novel OptAF that benefits from the advantages of existing robust AFs (ReLU, leaky ReLU, and sigmoid), allowing the proposed model to learn diverse deep features of complex datasets.
- The combination of several robust loss functions enables the proposed model to consider various aspects of prediction error for enhanced training.
- Experimental results demonstrate that the proposed method provides superior or competitive performance in terms of both accuracy and speed compared with existing state-of-the-art (SOTA) models.

The remainder of this paper is organized as follows. Section 2 discusses existing approaches to VAD and identifies disadvantages. A thorough explanation of the proposed model is provided in Section 3. Extensive experiments and their results are presented in Section 4. Section 5 presents the results of ablation studies. The implications of our findings are discussed in Section 6. Finally, Section 7 concludes this paper and outlines future research directions.

2 | RELATED WORK

Various approaches for VAD have been proposed over the past few years. These approaches can be roughly classified into two primary categories: reconstruction-based and prediction-based approaches.

2.1 | Reconstruction-based approach

In this approach, a model is taught to recreate inputs, and the AE architecture has emerged as the most

favorable model. This architecture consists of an encoder and decoder, where the encoder condenses inputs into a more compact form and the decoder reconstructs outputs from the condensed inputs, maintaining as much similarity as possible with the original inputs. Subsequently, normal and anomalous events are distinguished using reconstruction error, considering that normal events typically result in smaller errors than anomalous events.

Wei et al. [26] and Nawaratne et al. [27] independently developed models that extract both appearance and motion features from video inputs to learn normal events using an AE architecture. Their architectures employ stacked CNN and LSTM layers to capture spatial and temporal representations, respectively. However, the stacked convolutional LSTM layers in the model result in significant complexity.

Li et al. [9] proposed a model consisting of two streams to capture the spatial and temporal features of inputs. The inputs consist of 3D video cuboids constructed by combining multiple patches extracted from corresponding locations across consecutive frames. However, considering the computational complexity of optical flow, this method is resource intensive. Fang et al. [28] proposed a model consisting of multiple encoders and a solitary decoder to encode movement and appearance information. This model integrates a temporal encoder with two spatial encoders. The results of these encoders are combined and reconstructed using a single decoder. However, this model is impractical for efficient VAD because it employs several encoders for feature extraction. Hao et al. [29] proposed a method that uses a 3D CNN-based discriminator and 3D–2D U-shape structure to highlight disturbances in anomalous data, extract high-level spatiotemporal features, and enlarge the score gaps between normal and anomalous content to improve anomaly detection. However, this model struggles with real-time anomaly detection as a result of intensive resource requirements.

Abati and others [30] and Gong and others [31] independently employed deep AE models for image reconstruction. The former presented a probabilistic model that employs an autoregressive method to obtain density information. The latter proposed an AE equipped with a memory component, whose contents are learned by training with normal samples. However, the reconstruction of anomalous events using such a memory module can cause large errors during the testing phase. Li and others [8,32] introduced a multilayer reconstruction AE model and motion reconstruction loss to detect video anomalies. The main disadvantage of the mentioned models is their high computational costs.

Some scholars [11, 33] have employed the sparse coding technique to identify anomalies using a learned event

dictionary to reconstruct normal or anomalous events. Normal events result in small reconstruction errors, whereas anomalous events result in large reconstruction errors. However, these models require significant computational time because they used stacked RNN and LSTM layers.

2.2 | Prediction-based approach

Prediction-based models employ several preceding frames to predict whether a subsequent frame will be normal or anomalous, operating under the fundamental assumption that while normal events are predictable, anomalous events are not. Frame prediction models typically leverage both appearance and motion data from the provided video because inputs consist of several sequential frames that encompass motion features.

Li and others [15] first introduced the concept of future video frame prediction. Zhou and others [20] proposed an identical technique for predicting future frames. Imbalance between the background and foreground, which is common in VAD datasets, was addressed using attention-driven loss. Similarly, Li and others [34] employed a U-Net to predict future frames, and convolutional LSTM layers were incorporated between U-Net layers to extract temporal features. Lu and others [10] combined a variational AE and convolutional LSTM for future frame prediction to obtain temporal information from the frames of input videos. Li and others [32] proposed a method that incorporates convolutional LSTM, masked convolution, and attention modules to address the VAD problem using a future frame prediction technique. However, these models have high computational costs and require long training times. Chang and others [35] proposed an AE architecture that dissociates spatiotemporal representations and uses an efficient motion AE with a variance attention module and early fusion strategy to learn regularity in spatial and motion feature spaces. However, their method suffers from low accuracy on challenging datasets.

2.3 | Combined methods

Some researchers have combined the two approaches described above. For example, Tang and others [36] integrated a reconstruction approach with future frame prediction to exploit the advantages of both approaches. Specifically, two U-Net models were used to identify and reconstruct future frames. Chang and others [35] implemented a method that eliminates computationally intensive optical flow calculations. The first stream uses

an AE to encode spatial data, whereas the second stream uses a motion AE to predict the RGB differences between initial and final frames. Morais and others [37] proposed a VAD model that adopts dynamic skeleton features. This model separates skeletal motion into overall body motion and specific body positions. The decomposed features are then passed through two RNN branches to reconstruct initial inputs and predict future frames. However, these models are inefficient despite eliminating the need for optical flow computations to capture temporal information.

3 | PROPOSED METHODOLOGY

In this section, we describe the proposed AONet model. Figure 1 presents a comprehensive visual representation of the proposed model. Specifically, AONet contains three crucial modules: an encoder, spatiotemporal module, and decoder.

As shown in Figure 1, a sequence of frames is fed into the pre-trained CNN for feature extraction. The extracted graphical features are then passed through the spatiotemporal module to obtain the content and movement details within the frames. The outputs of the spatial and temporal branches are then summed before being fed into the decoder, which employs a bottom-up attention module to facilitate the more effective use of the interrelation between feature channels. Notably, OptAF, which combines the advantages of the leaky ReLU, ReLU, and sigmoid functions, is employed to learn the critical patterns of complex VAD datasets such as ShanghaiTech [38]. A more detailed description of OptAF is presented in Section 3.4. Furthermore, an objective function combining several robust loss functions supervises the model by considering various prediction error

characteristics. A detailed description of the objective function is presented in Section 3.5.

3.1 | Encoder and decoder

The model input, which is a sequence of frames, is passed through the encoder to extract visual features from the frames. WiderResNet38 [39] is used as the encoder in the proposed model because of its effectiveness compared with other deep CNNs. Instead of simply increasing depth, WiderResNet38 introduces a shallower and wider ResNet architecture with more filters per layer. The output of the final feature extraction layer of WiderResNet38 is inputted into the subsequent spatiotemporal module of the proposed model, while two other sets of high-level features from the intermediate layers of the encoder are used as skip connections to the corresponding decoder components.

Next, the final output of the encoder is fed into the spatiotemporal module to exploit the content and movement details within the frames. The results from both branches are combined by adding them element-wise prior to feeding them into the decoder to reconstruct future frames. The decoder consists of multiple layers, each including deconvolution, batch normalization, and OptAF, which enable the model to learn diverse features effectively. Additionally, an attention module is applied after each of the aforementioned layers to capture the channel interdependence of the features. Furthermore, the output features from the attention module are integrated with similar intermediate features obtained by the deep CNN in the encoder, which possesses an identical spatial resolution, ensuring that multiscale contextual information is preserved. Subsequently, the integrated features undergo deconvolution to increase their resolution to match that of the input frames. By selecting effective components such as WiderResNet38, employing innovative AFs, and integrating attention mechanisms, this model aims to advance the state of anomaly detection in video surveillance through precise and context-aware frame prediction and reconstruction.

3.2 | Spatiotemporal module

In this study, we used the temporal shift module (TSM) introduced by Lin and others [40] to obtain temporal information from videos. Unlike the 3D convolutional operation [41], which is computationally intensive, the temporal shift operation efficiently exploits temporal information in the input frames of a VAD dataset. Specifically, TSM moves the feature map along the temporal

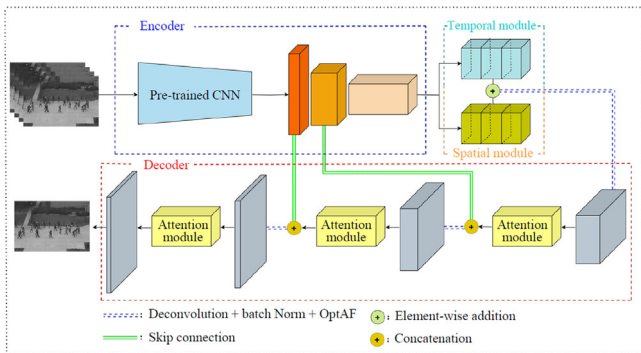


FIGURE 1 Overall architecture of the proposed AONet model.

dimension for efficient temporal modeling. Only a portion of the channels transition to the subsequent frame, leaving the remainder intact. Subsequently, the features of the current frame are merged with the features of the previous frame. Given input feature maps F_t with dimensions $\mathbb{R}^{N \times T \times C \times H \times W}$, where $N, T, C, H,$ and W denote the number of input samples, temporal dimensions, channels, height, and width, respectively, the output features are calculated as follows:

$$F'_t = \text{TS}(F_t), \quad (1)$$

where TS denotes a temporal shift operation. A graphical illustration of the temporal shift operation is presented in Figure 2.

As shown in Figure 2, the input features contain four frames $T = \{t_1, t_2, t_3, t_4\}$. After TSM is implemented, a segment of the channels from frame t_2 is replaced with a portion of a channel from frame t_1 .

The spatial module of the proposed model accumulates the features extracted from deep CNNs across frames. To reduce computational complexity, 1×1 convolution is applied to the aggregated features to reduce the channel count, given that the accumulated features include a substantial number of channels. The features from the temporal and spatial branches are combined as follows:

$$F = F_{\text{to}} + F_{\text{so}}, \quad (2)$$

where F_{to} and F_{so} represent temporal and spatial branch outputs, respectively.

3.3 | Attention module

Channel attention has been widely used in various fields to exploit the interdependence among feature channels

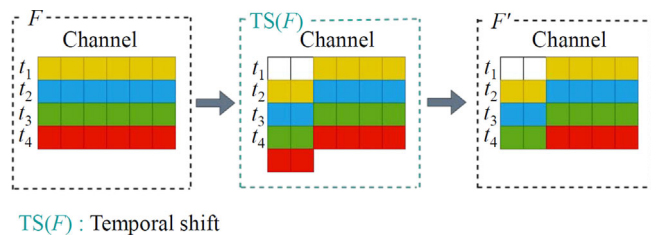


FIGURE 2 Graphical illustration of the temporal shift operation applied to feature map F to obtain output feature F' . Columns with corresponding colors represent features of different frames. One can see that the temporal shift operation is applied to a specific area in frame t_1 . As a result, that area is shifted to the next frame t_2 .

[42, 43]. In the proposed model, the output of each deconvolutional block in the decoder is passed through an attention module. A graphical representation of the attention module is presented in Figure 3.

The input for the attention module, which is also the output of the deconvolutional block, is $F \in \mathbb{R}^{C \times H \times W}$. First, a global average pooling (GAP) operation is applied to the feature map F to obtain the interrelation among channels. Next, 1×1 convolution is used to reduce the dimensions, followed by OptAF implementation and another 1×1 convolution that restores the channel dimensions. Therefore, channel attention can be calculated as follows:

$$A(F) = \sigma(W_2 \text{OptAF}(W_1 \text{GAP}(F))), \quad (3)$$

where σ and $A(F)$ denote the sigmoid AF and channel attention, respectively. Therefore, the overall output of the attention module is obtained as follows:

$$F' = F \otimes A(F), \quad (4)$$

where \otimes denotes the element-wise multiplication operation. Notably, for large and complex VAD datasets such as ShanghaiTech Campus and CUHK Avenue, the channel attention module is slightly modified to improve comprehension. Specifically, the input F is passed through a different convolutional block before being fed into the channel attention module described above. The convolutional block contains two 3×3 convolutional operations with OptAF implementation between them. Mathematically, this process is expressed as follows:

$$A_{\text{mod}}(F) = \text{Conv}(F) \otimes A(\text{Conv}(F)), \quad (5)$$

where A_{mod} and $\text{Conv}(F)$ denote the modified channel attention and convolutional blocks, respectively. Additionally, a residual connection is implemented to obtain the final output of the modified channel attention as follows:

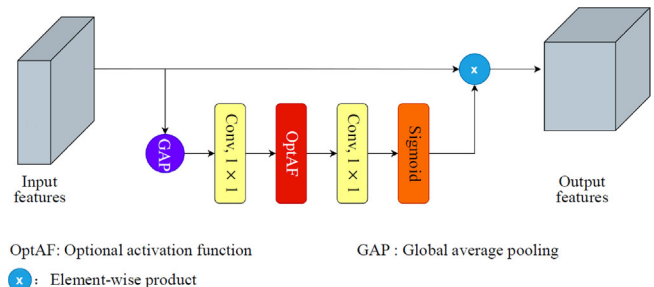


FIGURE 3 Graphical representation of the attention module.

$$F' = F \oplus A_{\text{mod}}(F), \quad (6)$$

where \oplus denotes element-wise addition.

3.4 | Optional AF

Learning even the smallest patterns in input frames is crucial for accurately detecting anomalous events considering their rare nature. The AF plays a significant role in the training dynamics and final performance of the model. Existing widely used AFs have disadvantages such as dead neurons or non-smooth transitions around zero. To address these issues, the proposed model leverages three well-known AFs (ReLU, leaky ReLU, and sigmoid) by combining the strengths of each function. Specifically, the positive region replicates the functionality of ReLU and can be expressed as follows:

$$\text{Positive} = \text{ReLU}(x). \quad (7)$$

ReLU provides impressive training speed but ignores negative values. In VAD, where every part of the input frame is equally important, all values must be considered. Therefore, OptAF also considers negative values. The negative region imitates the behavior of leaky ReLU as follows:

$$\text{Negative} = \alpha(x - \text{ReLU}(x)), \quad (8)$$

where α denotes a trainable parameter that controls the negative slope and varies across different datasets. Additionally, in the transition region, where the values are approximately zero, an enhanced sigmoid AF is employed for smoothing as follows:

$$\text{Transition} = \beta(x - \text{ReLU}(-\beta x))\sigma(x), \quad (9)$$

where β corresponds to another trainable value that controls the smoothing around zero and σ denotes the sigmoid AF. The proposed OptAF algorithm is defined as follows:

$$\text{OptAF}(x) = \text{positive} + \text{negative} + \text{transition}. \quad (10)$$

OptAF mitigates the disadvantages of each AF and benefits from their advantages through a comprehensive unification strategy. Additionally, the trainable parameters α and β allow for adaptation to any type of VAD dataset. A visualization of all the aforementioned AFs is presented in Figure 4.

As shown in Figure 4, OptAF incorporates all the benefits of these popular AFs while mitigating their disadvantages.

3.5 | Loss function

The proposed model uses combined constraints on pixel intensity, potential blur, image quality at various resolutions, and gradients. The proposed model predicts the subsequent \hat{F}_{t+1} frame given the input frames $\{F_1, F_2, \dots, F_t\}$ and compares \hat{F}_{t+1} with the actual frame F_t . In general, the constraints related to pixel intensity and its gradient in the frames play a crucial role in reducing prediction error. Therefore, the proposed model employs intensity loss to ensure the similarity of pixels as follows:

$$L_{\text{int}}(F, \hat{F}) = \|F - \hat{F}\|_2^2. \quad (11)$$

The gradient constraint is incorporated to address possible blurriness, thereby achieving a more visually pleasing video frame. The proposed model employs a loss function that computes the discrepancy between the absolute gradients across two spatial dimensions as follows:

$$L_{\text{grad}}(F, \hat{F}) = \sum_{k,l} \left(\left| \hat{F}_{k,l} - \hat{F}_{k-1,l} \right| - \left| F_{k,l} - F_{k-1,l} \right| \right)_1 + \left(\left| \hat{F}_{k,l} - \hat{F}_{k,l-1} \right| - \left| F_{k,l} - F_{k,l-1} \right| \right)_1, \quad (12)$$

where k and l denote the two spatial dimensions. Additionally, the proposed model measures structural similarity using multiscale structural similarity (MS-SSIM). To ensure that large deviations in pixel values between the predicted and actual frames are minimized, the proposed model employs root mean squared (RMS) loss as follows:

$$\text{RMS}(F, \hat{F}) = \sqrt{\|F - \hat{F}\|_2^2 + \epsilon}. \quad (13)$$

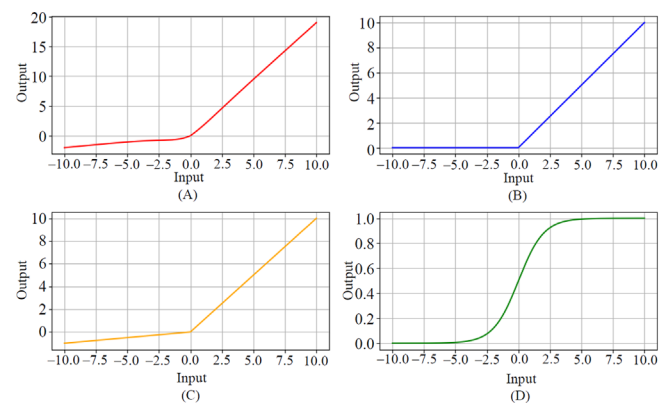


FIGURE 4 Graphical illustration of (A) OptAF, (B) ReLU, (C) leaky ReLU, and (D) sigmoid.

Consequently, the combined objective function for the proposed model, which includes the intensity, gradient, MS-SSIM, and RMS loss functions, is expressed as follows:

$$L(F, \hat{F}) = \alpha L_{\text{int}}(F, \hat{F}) + \beta L_{\text{grad}}(F, \hat{F}) + \gamma L_{\text{msssim}}(F, \hat{F}) + \lambda \text{RMS}(F, \hat{F}), \quad (14)$$

where α, β, γ , and λ are scalar values that balance the weights of the loss functions.

3.6 | Anomaly detection

Anomalies were detected using the anomaly score $AS(t)$. The peak signal-to-noise ratio (PSNR) was employed to assess the quality of predicted frames because it is a widely used metric in this context. The PSNR of a predicted frame is calculated as follows:

$$\text{PSNR}(F, \hat{F}) = 10 \log_{10} \frac{[\max_{\hat{F}}]^2}{\frac{1}{N} \sum_{k=1}^N (F_k - \hat{F}_k)^2}, \quad (15)$$

where N and $[\max_{\hat{F}}]$ denote the number of pixels and maximum value of \hat{F} , respectively. A higher PSNR suggests a higher likelihood of the frame being normal. After computing the PSNR for each frame in each test video, the proposed model follows the methodology described in [15] to normalize the PSNR values. Subsequently, the $AS(t)$ for each frame is calculated as follows:

$$AS(t) = \frac{\text{PSNR}_t - \min(\text{PSNR})}{\max(\text{PSNR}) - \min(\text{PSNR})}, \quad (16)$$

where $\min(\text{PSNR})$ and $\max(\text{PSNR})$ denote the lowest and highest PSNR values, respectively. Therefore, anomalous frames can be predicted based on $AS(t)$.

4 | EXPERIMENTS AND RESULTS

4.1 | Dataset description

To evaluate the performance of the proposed model, three benchmark datasets were selected, namely, the UCSD [44], CUHK Avenue [45], and ShanghaiTech Campus [38] datasets. The training sets of all datasets included only normal samples, whereas the testing sets contained both normal and anomalous samples.

4.1.1 | UCSD

This dataset is divided into two subsets (Ped1 and Ped2), each of which were captured at distinct outdoor locations. Ped1 has resolution of 158×238 , and Ped2 has a resolution of 240×360 . Typical recorded activities involve pedestrians walking through a camera's field of view, which are used during the training phase. The presence of cars, bikers, skaters, and wheelchairs is considered anomalous. Following the methodologies described in [46, 47], Ped1 was omitted from our experiments owing to its low resolution. The Ped2 dataset consists of 16 videos (2550 frames) for training and 12 videos (2010 frames) for testing.

4.1.2 | CUHK Avenue

This dataset consists of 16 training videos (15 328 frames) and 21 testing videos (15 324 frames). The resolution of the video frames is 360×640 pixels. There are 47 instances of anomalous events in the dataset, including object tossing, lingering, and rapid movements.

4.1.3 | ShanghaiTech Campus

This dataset is one of the most demanding datasets in VAD, containing 130 anomalous events. It consists of 330 training videos (274 515 frames) and 107 testing videos (42 883 frames) captured across 13 different scenes, each subject to different lighting environments and camera positions. The resolution of the video frames is 480×856 pixels.

4.2 | Training details

The proposed model was implemented using Python version 3.8 and Pytorch version 11.7. Experiments were conducted on a 32 GB NVIDIA Tesla V100-DGX5 GPU using CUDA 11.4. We resized the video frames of Ped2, CUHK Avenue, and ShanghaiTech Campus to 224×288 , 192×320 , and 192×288 pixels, respectively. Prior to being fed into the model, the intensity of every frame was adjusted to fall within a range of 1 to 1. The initial learning rate was established as $2e4$. The Adam optimizer was used to train the network. To control computational complexity, a pre-classified feature map was used instead of the final map extracted by WiderResNet38 during training on the CUHK Avenue and ShanghaiTech Campus datasets.

4.3 | Evaluation metric

In line with previous studies [15, 20], the performance of the proposed model was evaluated using the frame-level area under the curve (AUC). AUC was determined by calculating the area under the receiver operating characteristic curve, which varies based on different threshold values for anomaly scores. The AUC ranges from zero to one and superior anomaly detection performance is reflected by a higher AUC.

4.4 | Baseline models

We selected recent SOTA VAD models for comparison with the proposed model. These models can be divided into three categories, as discussed in Section 2: reconstruction-based models [8, 26–31, 33], prediction-based models [10, 15, 20, 35, 48], and combined models [36, 37].

4.5 | Experimental results on the UCSD Ped2 dataset

First, we conducted experiments using UCSD Ped2. Table 1 presents a performance comparison of the proposed model with SOTA VAD models on UCSD Ped2.

TABLE 1 Performance comparison on UCSD Ped2 in terms of AUC (%).

Model		AUC (%)
Reconstruction based	Abati et al. [30]	95.4
	Fang et al. [28]	95.6
	Gong et al. [31]	94.1
	Li and Chang [8]	91.6
	Luo et al. [33]	92.2
	Nawaratne et al. [27]	91.1
	Wei et al. [26]	89.5
	Hao et al. [29]	96.9
Prediction based	Liu et al. [15]	95.4
	Lu et al. [10]	96.0
	Yang et al. [48]	95.9
	Zhou et al. [20]	96.0
	Chang et al. [35]	96.7
Combined	Morais et al. [37]	96.3
	Tang et al. [36]	-
Ours		97.0

As indicated in Table 1, the proposed model achieved the highest accuracy. In particular, our model achieved an AUC of 97.0%, whereas the second-best model, which belonged to the reconstruction-based category [29], achieved an AUC of 96.9%. The highest AUCs among the combined [37] and prediction-based [35] models were 96.3% and 96.7%, respectively, which were 0.7% and 0.3% lower than that of the proposed model, respectively. Overall, the prediction-based models demonstrated more competitive results than the other two categories on this dataset, which can be considered as evidence of their effectiveness for VAD tasks.

4.6 | Experimental results on the CUHK Avenue dataset

The following experiments were performed using CUHK Avenue. Table 2 presents performance comparisons of our model with SOTA VAD models on the CUHK Avenue dataset.

As shown in Table 2, the proposed model achieved competitive performance with an AUC of 86.9%, whereas the highest AUC (87.1%) was achieved by the prediction-based model [35]. Regarding reconstruction-based models, our model outperformed the best-performing model proposed by Hao and others [29], which achieved an AUC of 86.6%, by 0.3%. Among the combined models,

TABLE 2 Performance comparison on CUHK Avenue in terms of AUC (%).

Model		AUC (%)
Reconstruction based	Abati et al. [30]	-
	Fang et al. [28]	86.3
	Gong et al. [31]	83.3
	Li and Chang [8]	84.2
	Luo et al. [33]	83.5
	Nawaratne et al. [27]	76.8
	Wei et al. [26]	79.7
	Hao et al. [29]	86.6
Prediction based	Liu et al. [15]	85.1
	Lu et al. [10]	85.7
	Yang et al. [48]	85.9
	Zhou et al. [20]	86.0
	Chang et al. [35]	87.1
Combined	Morais et al. [37]	86.3
	Tang et al. [36]	85.1
Ours		86.9

the model proposed by Morais and others [37] achieved the best AUC of 86.3%, which was 0.6 % lower than that of our model. In summary, the prediction-based category was more effective than the other two categories considering its higher overall results on the CUHK Avenue dataset.

4.7 | Experimental results on the ShanghaiTech Campus dataset

To verify the effectiveness of the proposed model on a large-scale challenging dataset, we conducted experiments on the ShanghaiTech Campus dataset. Table 3 presents performance comparisons of the proposed model with SOTA VAD models on the ShanghaiTech Campus dataset.

As shown in Table 3, the proposed model and the reconstruction-based model [29] achieved the highest accuracy with an AUC of 73.8%, whereas the second-best model proposed by Chang and others [35] achieved an AUC of 73.7%. Notably, this runner-up model is prediction based. The model proposed by Morais and others [37] achieved the highest AUC (73.4%) among the combined models, with a value 0.4% lower than that of our model.

4.8 | Speed comparison

We conducted experiments to compare the speed and AUC of the proposed method with two baseline SOTA

TABLE 3 Performance comparison on ShanghaiTech Campus in terms of AUC (%).

Model		AUC (%)
Reconstruction based	Abati et al. [30]	72.5
	Fang et al. [28]	73.2
	Gong et al. [31]	71.2
	Li and Chang [8]	-
	Luo et al. [33]	69.6
	Nawaratne et al. [27]	-
	Wei et al. [26]	67.2
	Hao et al. [29]	73.8
Prediction based	Liu et al. [15]	72.8
	Lu et al. [10]	-
	Yang et al. [48]	73.5
	Zhou et al. [20]	-
	Chang et al. [35]	73.7
Combined	Morais et al. [37]	73.4
	Tang et al. [36]	73.0
Ours		73.8

methods [15, 35] on the UCSD Ped2 dataset. Table 4 presents the results of the speed and AUC comparisons.

As shown in Table 4, when using the same GPU identified in Section 4.2, the proposed method detected anomalies in video frames at a speed of 29 FPS, matching or exceeding the performance of its SOTA counterparts. In particular, the methods from [15] and [35] achieved FPS values of 25 and 32, respectively. Additionally, the proposed method achieved an AUC of 97.0%, outperforming the other methods.

5 | ABLATION STUDY

5.1 | Performance comparison between OptAF and ReLU

To demonstrate the effectiveness of the OptAF used in the proposed model, we conducted experiments to compare its performance with that of the widely used ReLU AF. Table 5 presents performance comparisons between OptAF and ReLU in the proposed framework for all three datasets.

As shown in Table 5, OptAF outperformed ReLU on all datasets in terms of AUC. In particular, OptAF outperformed ReLU by 0.7% with an AUC of 97.0% on the UCSD Ped2 dataset. On the CUHK Avenue dataset, the AUC of ReLU was 86.1%, whereas that of OptAF was 86.9%, representing a 0.8% improvement. Similarly, the results for the most challenging dataset, ShanghaiTech Campus, proved the effectiveness of OptAF. ReLU yielded an AUC of 73.1%, whereas OptAF yielded an AUC of 73.8%. In summary, OptAF was more effective than ReLU at capturing critical patterns in large-scale challenging VAD datasets, as demonstrated by its superior results in these experiments.

TABLE 4 Speed and AUC comparisons on the UCSD Ped2 dataset.

Model	AUC (%)	FPS
Liu et al.	95.4	25
Chang et al.	96.7	32
Ours	97.0	29

TABLE 5 Performance comparison between OptAF and ReLU.

Activation function	AUC (%) on UCSD Ped2	AUC (%) on CUHK Avenue	AUC (%) on ShanghaiTech Campus
ReLU	96.3	86.1	73.1
OptAF	97.0	86.9	73.8

5.2 | Effectiveness of the combined loss function

Generally, the primary role of a loss function is to train a model by highlighting the different dimensions of prediction error. In many cases, the use of a combination of loss functions while training a model to optimize weight parameters can be beneficial. Typically, the use of a single loss function does not provide optimal results. Each loss function quantifies the model's performance from a unique perspective. Therefore, employing various robust loss functions contributes to determining optimal parameters. To verify the contribution of each loss function to the performance of the proposed model, we conducted experiments on the UCSD Ped2 dataset. Table 6 summarizes the effects of each loss function on the objective function of the proposed model.

As shown in Table 6, including all loss functions considered in the final objective function improves the performance of the proposed model. When using only L_{int} , the model achieved an AUC of 95.2%. When L_{grad} was added, the AUC increased to 95.9%, representing a 0.7% improvement. Similarly, including L_{msssim} on addition to the previous two loss functions increased the AUC to 96.4%, representing an improvement of 0.5%. Finally, incorporating RMS into the final objective function in addition to the three aforementioned loss functions

yielded the highest AUC of 97.0%, representing a 0.6% improvement. In summary, combining various robust loss functions in the final objective function enables the effective training of the proposed model by addressing different important aspects of prediction error.

5.3 | Performance comparison of pre-trained CNNs

To verify the effectiveness of the proposed model, we conducted an ablation study comparing its performance with that of various deep CNNs, where only the pre-trained CNN was changed in the network architecture. The results of this ablation study are presented in Table 7.

As shown in Table 7, the proposed method using WiderResNet38 as the pre-trained CNN achieved the best results on the ShanghaiTech Campus, CUHK Avenue, and UCSD Ped2 datasets with AUCs of 73.8.

5.4 | Performance comparison of the components of the proposed method

We conducted an ablation study to assess the performance of the three modules in the proposed method, namely, the spatial, temporal, and attention modules. Table 8 presents performance evaluations of the proposed method using various combinations of these modules.

As shown in Table 8, performance improved with the addition of the other two modules compared with using the spatial module alone. In particular, the temporal module significantly improved performance. The proposed method achieved the best results with all three modules, achieving AUCs of 97.0%, 86.9%, and 73.8% on

TABLE 6 Impact of loss functions.

Loss function	AUC (%) on UCSD Ped2
L_{int}	95.2
$L_{\text{int}} + L_{\text{grad}}$	95.9 (+0.7)
$L_{\text{int}} + L_{\text{grad}} + L_{\text{msssim}}$	96.4 (+0.5)
$L_{\text{int}} + L_{\text{grad}} + L_{\text{msssim}} + \text{RMS}$	97.0 (+0.6)

TABLE 7 Performance evaluation of pre-trained CNNs in the proposed method in terms of AUC (%).

Method	Pre-trained CNN	UCSD Ped2	CUHK Avenue	ShanghaiTech Campus
Proposed method	ResNet101 [49]	95.4	82.4	72.9
	SE-ResNext101 [50]	96.3	84.5	73.2
	WiderResNet38 [39]	97.0	86.9	73.8

TABLE 8 Performance evaluation of processing modules of the proposed method in terms of AUC (%).

Pre-trained CNN	Spatial	Temporal	Attention	UCSD Ped2	CUHK Avenue	ShanghaiTech Campus
WiderResNet38	Yes	No	No	96.8	84.8	71.8
	Yes	Yes	No	96.2	85.5	72.7
	Yes	No	Yes	96.5	85.9	72.6
	Yes	Yes	Yes	97.0	86.9	73.8

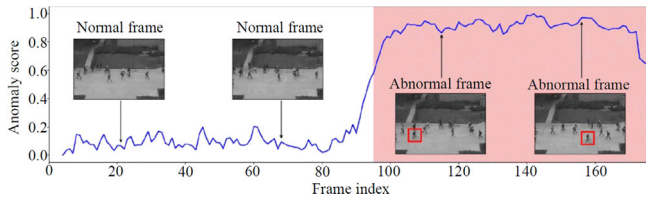


FIGURE 5 Visualization of frame-wise anomaly scores. The blue line represents the calculated anomaly scores, pink area indicates the actual occurrence of anomalous events, and red bounding box highlights the anomalous action object (a person riding a bicycle).

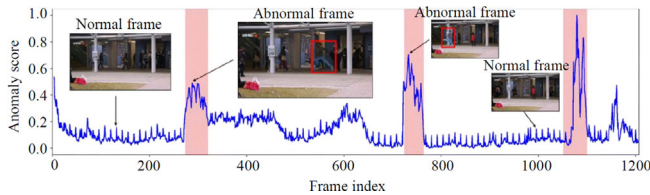


FIGURE 6 Visualization of frame-wise anomaly scores. The blue line represents the calculated anomaly scores, pink area indicates the actual occurrence of anomalous events, and red bounding box highlights the anomalous action object (a running person).

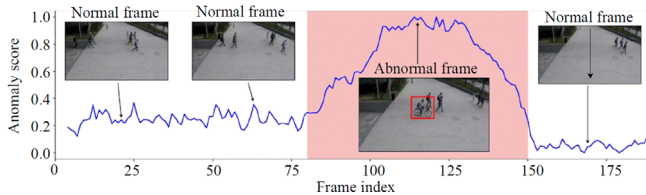


FIGURE 7 Visualization of frame-wise anomaly scores. The blue line represents the anomaly scores, pink area indicates the actual occurrence of anomalous events, and red bounding box highlights the anomalous action object (a person riding a bicycle).

the UCSD Ped2, CUHK Avenue, and ShanghaiTech Campus datasets, respectively.

6 | DISCUSSION

6.1 | Visualization results

Visualizations of the anomaly scores for the test video frames in the UCSD Ped2, CUHK Avenue, and ShanghaiTech Campus datasets are presented in the figures. Specifically, Figure 5 presents the frame-wise anomaly scores for the second video in the UCSD Ped2 test dataset.

As illustrated in Figure 5, the anomaly score rapidly increases when the anomalous event occurs but remains

low in the absence of such events. Similarly, Figure 6 presents the frame-wise anomaly scores for the second test video from the CUHK Avenue dataset.

As shown in Figure 6, the first two anomalous events (running people) cause the anomaly score to increase. However, a third increase in the anomaly score was triggered by camera disturbances. Figure 7 presents a visualization of the anomaly scores for the frames in the second test video from the ShanghaiTech Campus dataset.

As shown in Figure 7, the anomaly score increases significantly when a person riding a bicycle appears in the frames, demonstrating that the proposed method can effectively distinguish anomalous frames from normal frames.

6.2 | Limitations of the proposed method

After evaluating the performance of the proposed method on the three benchmark datasets, we identified several limitations. First, there are considerable performance differences between datasets, particularly between UCSD Ped2 and ShanghaiTech Campus. This discrepancy may have occurred because the ShanghaiTech Campus dataset is more challenging with 130 anomalous events and 13 diverse scenes, each featuring different lighting conditions and camera angles. Second, the proposed method is not sufficiently robust to common issues in video data such as camera shaking. As shown in Section 6.1, the proposed method incorrectly identified camera disturbances as anomalous events. Third, the anomaly detection speed of the proposed method can be further enhanced through more efficient design choices.

7 | CONCLUSION AND FUTURE WORKS

In this paper, we proposed an attention-based residual AE architecture using a novel OptAF for VAD. The proposed method detects anomalous events in an unsupervised manner by utilizing appearance and motion information from spatial and temporal modules, respectively. Temporal feature extraction was implemented using TSM. To enhance learning effectiveness, the proposed method incorporates attention modules and OptAF, which combines the advantages of ReLU, leaky ReLU, and sigmoid. Additionally, the combined loss function employed in the proposed method contributes to its superior performance. Extensive experimental results demonstrated that the proposed method outperformed baseline models on three benchmark VAD datasets. Ablation studies revealed that

the proposed OptAF-based model outperformed a model based on ReLU in terms of AUC. Additionally, the combined loss function improved performance, and the spatio-temporal and attention modules contributed to the performance boost of the proposed method. Furthermore, the proposed method demonstrated potential for real-world applications owing to its competitive speed. In the future, we plan to optimize the performance and efficiency of the proposed method further and apply it to real-world scenarios.

AUTHOR CONTRIBUTIONS

Akhrorjon Akhmadjon Ugli Rakhmonov: Conceptualization; methodology; software; writing—original draft. **Barathi Subramanian:** Data curation; writing—review and editing. **Bahar Amirian Varnousefaderani:** Visualization; investigation. **Jeonghong Kim:** Supervision.

ACKNOWLEDGMENTS

This study was supported by the BK21 FOUR project (AI-driven Convergence Software Education Research Program) funded by the Ministry of Education, Department of Computer Science and Engineering, Kyungpook National University, Daegu, Republic of Korea (4120240214871), and the Basic Science Research Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Republic of Korea (2021R1I1A3043970).

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflicts of interest.

ORCID

Jeonghong Kim  <https://orcid.org/0000-0002-7466-1376>

REFERENCES

1. E. M. Imah and R. D. I. Puspitasari, *Violent crowd flow detection from surveillance cameras using deep transfer learning-gated recurrent unit*, ETRI J. **46** (2024), no. 4, 671–682. <https://doi.org/10.4218/etrij.2023-0222>
2. A. Mukherjee, V. Hassija, and V. Chamola, *QuARCS: quantum anomaly recognition and caption scoring framework for surveillance videos*, IEEE Trans. Consumer Electron. (2024). <https://doi.org/10.1109/TCE.2024.3440520>
3. X. Wei, Y. Zhang, X. Zhang, Q. Ge, and B. Yin, *Real-time passenger flow anomaly detection in metro system*, IET Intell. Transport Syst. **17** (2023), no. 10, 2020–2033.
4. H. Zhu, P. Wei, and Z. Xu, *A spatio-temporal enhanced graph-transformer autoencoder embedded pose for anomaly detection*, IET Comput. Vis. **18** (2023), 405–419.
5. A. A. U. Rakhmonov, B. Subramanian, B. Olimov, and J. Kim, *Extensive knowledge distillation model: an end-to-end effective anomaly detection model for real-time industrial applications*, IEEE Access **11** (2023), 69750–69761. <https://doi.org/10.1109/ACCESS.2023.3293108>
6. B. A. Ugli Olimov, K. C. Veluvolu, A. Paul, and J. Kim, *UzADL: anomaly detection and localization using graph Laplacian matrix-based unsupervised learning method*, Comput. Ind. Eng. **171** (2022), 108313. <https://doi.org/10.1016/j.cie.2022.108313>
7. A. Hussain, W. Ullah, N. Khan, Z. A. Khan, M. J. Kim, and S. W. Baik, *TDS-Net: transformer enhanced dual-stream network for video anomaly detection*, Expert Syst. Appl. **2024** (2024), 124846.
8. N. Li and F. Chang, *Video anomaly detection and localization via multivariate gaussian fully convolution adversarial autoencoder*, Neurocomputing **369** (2019), 92–105.
9. N. Li, F. Chang, and C. Liu, *Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes*, IEEE Trans. Multim. **23** (2020), 203–215.
10. Y. Lu, K. M. Kumar, S. Shahabeddin Nabavi, and Y. Wang, *Future frame prediction using convolutional VRNN for anomaly detection* (16th IEEE Int. Conf. Adv. Video Signal Based Surveillance (AVSS), Taipei, Taiwan), 2019, pp. 1–8.
11. 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 Sec. **14** (2019), no. 10, 2537–2550.
12. X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W. Woo, *Convolutional LSTM network: a machine learning approach for precipitation nowcasting*, Adv. Neural Inf. Process. Syst. **28** (2015), 802.
13. Y. Wu, F. He, D. Zhang, and X. Li, *Service-oriented feature-based data exchange for cloud-based design and manufacturing*, IEEE Trans. Services Comput. **11** (2015), no. 2, 341–353.
14. M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, *Learning temporal regularity in video sequences* (Proc. IEEE Conf. Comput. Vision Pattern Recognit., Las Vegas, NV, USA) IEEE, Piscataway, New Jersey, USA 2016, pp. 733–742.
15. W. Liu, W. Luo, D. Lian, and S. Gao, *Future frame prediction for anomaly detection—a new baseline* (Proc. IEEE Conf. Comput. Vision Pattern Recognit., Salt Lake City, UT, USA), IEEE, Piscataway, New Jersey, USA 2018, pp. 6536–6545.
16. W. Luo, W. Liu, D. Lian, and S. Gao, *Future frame prediction network for video anomaly detection*, IEEE Trans. Pattern Anal. Mach. Intell. **44** (2021), no. 11, 7505–7520.
17. W. W. Y. Ng, G. Zeng, J. Zhang, D. S. Yeung, and W. Pedrycz, *Dual autoencoders features for imbalance classification problem*, Pattern Recognit. **60** (2016), 875–889.
18. L. Vu and Q. U. Nguyen, *An ensemble of activation functions in autoencoder applied to IoT anomaly detection* (6th NAFOSTED Conf. Inf. Comput. Sci., Hanoi, Vietnam), IEEE, Piscataway, New Jersey, USA 2019. <https://doi.org/10.1109/NICS48868.2019.9023860>
19. Y. Liu, H. Shen, T. Wang, and G. Bai, *Vehicle counting in drone images: an adaptive method with spatial attention and multi-scale receptive fields*, ETRI J. (2024). <https://doi.org/10.4218/etrij.2023-0426>
20. J. T. Zhou, L. Zhang, Z. Fang, J. Du, X. Peng, and Y. Xiao, *Attention-driven loss for anomaly detection in video surveillance*, IEEE Trans. Circuits Syst. Video Technol. **30** (2019), no. 12, 4639–4647.
21. Y. Liu, J. Liu, K. Yang, B. Ju, S. Liu, Y. Wang, D. Yang, P. Sun, and L. Song, *AMP-Net: appearance-motion prototype network*

- assisted automatic video anomaly detection system, *IEEE Trans. Ind. Inf.* **20** (2024), no. 2, 2843–2855. <https://doi.org/10.1109/TII.2023.3298476>
22. Y. Lu, F. Yu, M. K. K. Reddy, and Y. Wang, *Few-shot scene-adaptive anomaly detection* (Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK), 2020, pp. 125–141.
 23. V. Nair and G. E. Hinton, *Rectified linear units improve restricted boltzmann machines* (Proc. 27th Int. Conf. Mach. Learn. (ICML-10), Haifa, Israel), Omnipress, Madison, Wisconsin, USA 2010, pp. 807–814.
 24. A. L. Maas, A. Y. Hannun, and A. Y. Ng, *Rectifier nonlinearities improve neural network acoustic models* (Proc. ICML, Vol. 30, Atlanta, GA, USA), JMLR.org, New York, NY 2013, pp. 3.
 25. C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, Vol. 4, Springer, Berlin, Germany 2006.
 26. H. Wei, K. Li, H. Li, Y. Lyu, and X. Hu, *Detecting video anomaly with a stacked convolutional lstm framework* (Int. Conf. Comput. Vision Syst., Thessaloniki, Greece), Berlin, Germany, 2019, pp. 330–342.
 27. R. Nawaratne, D. Alahakoon, D. De Silva, and X. Yu, *Spatio-temporal anomaly detection using deep learning for real-time video surveillance*, *IEEE Trans. Ind. Inf.* **16** (2019), no. 1, 393–402.
 28. Z. Fang, J. T. Zhou, Y. Xiao, Y. Li, and F. Yang, *Multi-encoder towards effective anomaly detection in videos*, *IEEE Trans. Multimedia* **23** (2020), 4106–4116.
 29. Y. Hao, J. Li, N. Wang, X. Wang, and X. Gao, *Spatiotemporal consistency-enhanced network for video anomaly detection*, *Pattern Recognit.* **121** (2022), 108232.
 30. D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, *Latent space autoregression for novelty detection* (Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit., Long Beach, CA, USA), IEEE, Piscataway, New Jersey, USA, 2019, pp. 481–490.
 31. D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. Hengel, *Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection* (Proc. IEEE/CVF Int. Conf. Comput. Vision, Seoul, Rep. of Korea), 2019, pp. 1705–1714.
 32. B. Li, Z. Li, and Z. Yin, *Video anomaly detection via improved future frame prediction* (Fourth Int. Conf. Comput. Vision Data Mining (ICCVDM 2023), Vol. 13063, Changchun, China), SPIE, Bellingham, Washington, USA 2024, pp. 121–129.
 33. W. Luo, W. Liu, D. Lian, J. Tang, L. Duan, X. Peng, and S. Gao, *Video anomaly detection with sparse coding inspired deep neural networks*, *IEEE Trans. Pattern Anal. Mach. Intell.* **43** (2019), no. 3, 1070–1084.
 34. Y. Li, Y. Cai, J. Liu, S. Lang, and X. Zhang, *Spatio-temporal unity networking for video anomaly detection*, *IEEE Access* **7** (2019), 172425–172432.
 35. Y. Chang, Z. Tu, W. Xie, B. Luo, S. Zhang, H. Sui, and J. Yuan, *Video anomaly detection with spatio-temporal dissociation*, *Pattern Recognit.* **122** (2022), 108213.
 36. Y. Tang, L. Zhao, S. Zhang, C. Gong, G. Li, and J. Yang, *Integrating prediction and reconstruction for anomaly detection*, *Pattern Recognit. Lett.* **129** (2020), 123–130.
 37. R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, *Learning regularity in skeleton trajectories for anomaly detection in videos* (Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit., Long Beach, CA, USA), IEEE, Piscataway, New Jersey, USA 2019, pp. 11996–12004.
 38. W. Luo, W. Liu, and S. Gao, *A revisit of sparse coding based anomaly detection in stacked RNN framework* (Proc. IEEE Int. Conf. Comput. Vision., Venice, Italy), 2017, pp. 341–349.
 39. Z. Wu, C. Shen, and A. Van Den Hengel, *Wider or deeper: Revisiting the ResNet model for visual recognition*, *Pattern Recognit.* **90** (2019), 119–133.
 40. J. Lin, C. Gan, and S. Han, *TSM: temporal shift module for efficient video understanding* (Proc. IEEE/CVF Int. Conf. Comput. Vision, Seoul, Rep. of Korea), IEEE, Piscataway, New Jersey, USA 2019, pp. 7083–7093.
 41. D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, *Learning spatiotemporal features with 3D convolutional networks* (Proc. IEEE Int. Conf. Comput. Vision., Santiago, Chile), IEEE, Piscataway, New Jersey, USA 2015, pp. 4489–4497.
 42. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, *CBAM: convolutional block attention module* (Proc. Eur. Conf. Comput. Vision (ECCV), Munich, Germany), 2018, pp. 3–19.
 43. Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, *Image super-resolution using very deep residual channel attention networks* (Proc. Eur. Conf. Comput. Vision (ECCV), Munich, Germany), Springer, Berlin, Germany 2018, pp. 286–301.
 44. V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, *Anomaly detection in crowded scenes* (IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Beijing, China), 2010, pp. 1975–1981.
 45. C. Lu, J. Shi, and J. Jia, *Abnormal event detection at 150 FPS in MATLAB* (IEEE Int. Conf. Comput. Vis., Sydney, Australia), 2013, pp. 2720–2727.
 46. K. Doshi and Y. Yilmaz, *Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate*, *Pattern Recognit.* **114** (2021), 107865.
 47. V.-T. Le and Y.-G. Kim, *Attention-based residual autoencoder for video anomaly detection*, *Appl. Intell.* **53** (2023), no. 3, 3240–3254.
 48. Y. Yang, D. Zhan, F. Yang, X.-D. Zhou, Y. Yan, and Y. Wang, *Improving video anomaly detection performance with patch-level loss and segmentation map* (IEEE 6th Int. Conf. Comput. Commun. (ICCC), Chengdu, China), 2020, pp. 1832–1839.
 49. K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition* (Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Las Vegas, NV, USA), 2016, pp. 770–778.
 50. J. Hu, L. Shen, and G. Sun, *Squeeze-and-excitation networks*, (Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA), 2018, pp. 7132–7141.

AUTHOR BIOGRAPHIES



Akhrorjon Akhmadjon Ugli Rakhmonov received his BS degree from Ferghana Polytechnic Institute, Uzbekistan, in 2014 and his MS degree from Keimyung University, Rep. of Korea, in 2017. He is currently pursuing a PhD degree at Kyungpook National University, Rep. of Korea. His research interests include solving computer vision problems using machine and deep learning techniques.



Barathi Subramanian received her BS degree from Nirmala College for Women affiliated with Bharathiar University, India, in 2013; her MS degree from Bharathiar University, in 2015; and her PhD degree from Kyungpook National University, Daegu, Rep. of Korea, in 2023. She is currently a post-doctoral scholar at the Stanford University, CA, USA. Her research interests include solving computer vision problems using machine and deep learning techniques.



Bahar Amirian Varnousefaderani received her BS degree from Shahid Beheshti University, Iran, in 2021. She is currently pursuing an MS degree at Kyungpook National University, Daegu, Rep. of Korea. Her research interests include solving computer vision problems using machine and deep learning techniques.



Jeonghong Kim received his BS degree from Kyungpook National University, Daegu, Rep. of Korea, in 1984; his MS degree from Kyungpook National University in 1986; and his PhD degree from Chungnam National University, Daejeon, Rep. of Korea, in 2001. He is currently a professor at Kyungpook National University. His research interests include solving computer vision problems using machine and deep learning techniques.

How to cite this article: A. A. U. Rakhmonov, B. Subramanian, B. Amirian Varnousefaderani, and J. Kim, *AONet: Attention network with optional activation for unsupervised video anomaly detection*, ETRI Journal **46** (2024), 890–903, DOI [10.4218/etrij.2024-0115](https://doi.org/10.4218/etrij.2024-0115)