IEEE Open Journal of the Computer Society

Received 29 November 2024; accepted 9 December 2024. Date of publication 13 December 2024;
date of current version 9 January 2025. The review of this article was arranged by Associate Editor Peng Li.

Digital Object Identifier 10.1109/OJCS.2024.3517154

Multimodal Attention-Enhanced Feature Fusion-Based Weakly Supervised Anomaly Violence Detection

JUNGPIL SHIN [1] (Senior Member, IEEE), ABU SALEH MUSA MIAH [1] (Member, IEEE), YUTA KANEKO [1], NAJMUL HASSAN [1] (Graduate Student Member, IEEE), HYOUN-SUP LEE [2], AND SI-WOONG JANG [3]

[1]School of Computer Science and Engineering, University of Aizu, Aizuwakamatsu 965-0006, Japan
[2]Department of Game Engineering, Dongeui University, Busan 47340, Republic of Korea
[3]Department of Computer Engineering, Dongeui University, Busan 47340, Republic of Korea

CORRESPONDING AUTHORS: JUNGPIL SHIN; SI-WOONG JANG (e-mail: jpshin@u-aizu.ac.jp; swjang@deu.ac.kr).

This work was supported in part by the Ministry of SMEs and Startups (MSS, South Korea) through the Technological Innovation R&D Program under Grant RS-2024-00425148, in part by the Competitive Research Fund of The University of Aizu, Japan.

ABSTRACT Weakly supervised video anomaly detection (WS-VAD) plays a pivotal role in advancing intelligent surveillance systems within the field of computer vision. Despite significant research, WS-VAD continues to face challenges, particularly with unimodal approaches that struggle to extract meaningful features effectively. A few research studies have been done on the multimodal dataset fusion-based WS-VAD system, and their performance accuracy is unsatisfactory. In response, we propose a novel WS-VAD system leveraging multimodal datasets with an attention-enhanced feature fusion approach to address these challenges. Our system integrates three distinct data modalities—RGB video, optical flow, and audio signals—where each stream extracts complementary spatial and temporal features using an enhanced attention module to improve the detection accuracy and robustness. In the RGB video stream, we employ a multi-stage, attention-driven feature enhancement process to refine spatial and temporal features. This process begins with a ViT-based CLIP module, where the top k features are concatenated with I3D- and TCA-based spatiotemporal features. Temporal dependencies are then captured through uncertainty-regulated dual memory units (UR-DMUs), allowing the simultaneous learning of normal and anomalous patterns. The final stage selects the most relevant features, yielding a refined representation of RGB-based data. The second stream extracts enhanced spatiotemporal features from flow data using a deep learning and attention module. Lastly, the audio stream detects anomalies in sound patterns through an attention module integrated with the VGGish model, capturing auditory cues. The fusion of these three streams captures motion and audio signals often missed by visual analysis alone, significantly enhancing anomaly detection accuracy and robustness. Our multimodal fusion achieves an average precision (AP) of 88.28% on the XD-Violence dataset, outperforming prior models by nearly 2%, and attains AUCs of 98.71% on the ShanghaiTech dataset and 90.26% on the UCF-Crime dataset. These results underscore the effectiveness of our approach, consistently surpassing existing methods across three benchmark datasets and validating its robustness in WS-VAD applications.

INDEX TERMS Anomaly detection, Flow, RGB video, Audio signal, Multimodality fusion, Uncertainty-regulated dual memory unit (UR-DMU), Temporal contextual aggregation (TCA), Global/local multi-head self-attention (GL-MHSA), Weakly supervised video anomaly detection (WS-VAD), Magnitude contrastive (MC), Vision Transformer (ViT).

ABBREVIATIONS

AP	Average Precision
AUC	Area Under the Curve
MIL	Multiple Instances Learning
NVs	Normal Videos
AVs	Anomalous Videos
BCE	Binary Cross-Entropy

I. INTRODUCTION

Anomaly detection, specifically in video analysis, focuses on identifying unusual or abnormal events in video data. The field has three dominant approaches: supervised, unsupervised, and weakly supervised methods. Supervised methods, though effective, rely on fully annotated data, making them costly [1], [2]. Unsupervised methods, such as variational autoencoder decoders (VADs), struggle to accurately capture complex anomaly behaviours, limiting their effectiveness [3]. Weekly supervised video anomaly detection (WS-VAD) is crucial in various real behaviour applications, including security surveillance, traffic monitoring, and public safety. Identifying anomalies like accidents, thefts, or suspicious behaviours can prevent harm and enhance decision-making. For this reason, many researchers have been working on using various machine learning and deep learning technologies to automate the monitoring process, reduce human error, and promptly identify potential risks or threats. As video data becomes more prevalent in modern surveillance systems, the need for efficient, scalable, and accurate anomaly detection methods has become critical. WS-VAD has emerged as a promising approach, leveraging video-level labels rather than frame-by-frame annotations to detect anomalies more efficiently and cost-effectively [4], [5]. One popular method in WS-VAD is multiple instance learning (MIL), where videos are treated as "bags" containing snippets, and the model predicts whether the video contains abnormal events [6]. However, current WS-VAD methods face challenges in accurately distinguishing between normal and abnormal snippets due to the presence of normal instances within positive bags [7]. Moreover, unsupervised video anomaly detection (UVAD) methods are limited by the lack of prior knowledge about abnormal events, which can hinder the accuracy. Recent research has explored multimodal data fusion approaches, combining video and audio data to improve performance. These include unimodal methods like CNN-based I3D [8] and ViT-based CLIP [9], and multimodal systems that integrate modalities such as RGB, flow, and audio data [10], [11], [12], [13], [14], [15]. For example, temporal context aggregation (TCA) improves temporal feature extraction using self-attention mechanisms [16], [17], and uncertainty-regulated dual memory units (UR-DMUs) enhance temporal dependencies [18]. More recently, graph-based anomaly detection models have shown promise but lack temporal robustness [19], while many multimodal fusion systems underperform due to insufficient feature extraction [20]. Despite these advancements, challenges remain in handling complex, diverse anomaly types, including short-term, appearance-based, and audio-only anomalies. Furthermore, existing methods often fail to perform well due to ineffective feature representations and limited spatial and temporal enhancements extracted from unimodal datasets [4], [5], [6], [9], [17], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]. To overcome the problems caused by a lack of spatial-temporal features, recently some researchers combined RGB and audio datasets [10], [11], [12], [13], [14], [15], [32], [33], [34], [35]. However, this research work did not perform fusion at the feature level; instead, fusion was performed at the data level. Consequently, these model feature fusions may fail to boost performance due to their failure to implicitly align the multimodal features. They sometimes fail to take advantage of the multimodality of the data. We proposed a weakly supervised multimodal attention-enhanced feature fusion-based anomaly detection system to overcome these challenges. Our approach integrates CNN and ViT-based pre-trained features, attention-based spatial-temporal feature enhancement, and spatial feature enhancement to improve anomaly detection rates effectively. The main contributions of the proposed model are given below:

- *RGB Video Stream:*
  - *ViT-based CLIP Module:* This branch of the RGB stream utilizes a ViT-based CLIP module to select the top k features, capturing complex visual semantics and contextual information.
  - *CNN-based I3D Module with TCA:* The second branch leverages a CNN-based I3D module integrated with the TCA mechanism to extract rich spatiotemporal features.
  - *UR-DMU-based Feature Processing:* The combined features are processed through the UR-DMU model, which employs GCN and GL-MHSA modules to capture video associations. Feature reduction is achieved via a multilayer perceptron (MLP), producing the final feature representation for the RGB stream.
- *Flow Data Modality Stream:* In this stream, first, we computed the motion flow from the RGB consequence frames, and then we fed it into the I3D module to capture both spatial and temporal information that also highlights scene dynamics crucial for detecting anomalies related to unusual movements. The motion features are refined through an MLP and subsequently fed into a Transformer to capture long-range dependencies and temporal patterns, resulting in the final flow stream features.
- *Audio Stream:* The third stream extracts features from audio signals using a Transformer applied to VGGish-extracted features. This approach captures critical audio cues for anomaly detection that may not be visible in the visual data. The VGGish model processes audio inputs into detailed feature representations, which the Transformer further enhances by capturing temporal dependencies and contextual relationships, allowing the precise identification of subtle audio anomalies, complementing the visual streams.

- *Gated Feature Fusion with Attention Module and Classification:* Features from all three streams are concatenated using a gated feature fusion mechanism with an attention module, producing a comprehensive final feature set for the classification module. The classifier then predicts snippet-level anomaly scores. During training, these scores are aggregated into bag-level predictions to identify high activations in anomalous cases.
- *Comprehensive Evaluation:* Extensive experiments on the XD-Violence, ShanghaiTech, and UCF-Crime datasets demonstrate the superior performance of our method in anomaly detection. On the XD-Violence dataset, our multimodal fusion achieves an average precision (AP) of 88.28%, surpassing prior models by nearly 2%. Additionally, it attains AUCs of 98.71% on ShanghaiTech and 90.26% on UCF-Crime. These results highlight the robustness and effectiveness of our approach, consistently outperforming state-of-the-art methods across multiple benchmarks in WS-VAD applications.

The article's remaining sections are organized as follows: Section II provides the literature review summary. The datasets used are described in Section III. The proposed model architecture and method description are included in Section IV. The performance analysis, including the ablation study, performance table, and state-of-the-art comparison, are described in Section V. Lastly, the conclusion is given in Section VI.

II. LITERATURE REVIEW

The WS-VAD-based approaches rely on video-level labels, adhering consistently to the MIL framework [6]. Many researchers have integrated pre-trained deep learning (DL) models into their experimental setups [21], [24], [36]. Sultani et al. [6] curated pre-annotated normal and abnormal video events at the video level on the UCF-Crime dataset then extracted C3D features [37] from video segments and subsequently employing a ranking loss function to train a fully connected neural network (FCNN) [38], [39], [40], [41]. The objective of this function was to calculate the loss between the highest-scoring ranked examples within the positive and negative bags. Tian et al. [24] introduced a model for WS-VAD, leveraging feature extractors such as C3D [37] and I3D [8]. Zang et al. [21] proposed a model using temporal convolution networks (TCNs) to extract C3D features from positive and negative video segments. They trained the network to discriminate between adjacent segments, employing inner and outer bag ranking losses to train their model with two branches of an FCNN [42], [43], [44], [45]. This approach focused on the highest- and lowest-scoring segments within positive and negative bags, respectively. Other approaches have also incorporated deep learning techniques, including self-attention and graph-based feature extraction [46], [47], [48], [49], [50]. Zhong et al. [5] developed a supervised model with noisy labels to generalize their WS-VAD system, integrating temporal segment networks [51] alongside C3D features. Zhu et al. [36] introduced an attention mechanism to improve temporal context capture in MIL, showcasing that C3D and I3D motion information can outperform single-frame models like VGG, long short term memory (LSTM), and Inception [52], [53]. More recently, vision transformer (ViT) based systems have been developed by some researchers to extract potential features by including CLIP to improve the detection performance [54]. To address the recent problem of the WS-VAD, Lv et al. [29] developed an unbiased MIL approach that trains a fair anomaly classifier alongside a tailored representation designed explicitly for WS-VAD. To integrate the strengths of both CNN- and ViT-based pre-trained models [42], [55], researchers have devised architectures like CNN-ViT-TSAN supported by MIL aiming to build a diverse array of models tailored to addressing challenges in WS-VAD. The main challenges of the above-mentioned work lie in their approach to processing videos frame by frame or in short clips, limiting their ability to capture long-range semantic contextual information effectively. To address this challenge, the authors of [16], [17] used a TCA framework for video representation learning aiming to integrate the long-range temporal context into frame-level features using self-attention mechanisms [16], [17]. They employed contrastive learning to lower loss or error rates during evaluation. To further enhance TCA features, they utilized robust TCA features alongside an MIL loss calculation approach [24]. Their approach achieved notable results, with reported AUC values of 84.30% for the UCF-Crime dataset and 97.21% for the ShanghaiTech dataset. Zhou et al. [56] introduced the BatchNorm-WVAD method, which enhances WS-VAD by incorporating BatchNorm and using the Divergence of Feature from the Mean vector (DFM) as an abnormality criterion. They improve anomaly recognition, reduce the impact of noisy labels, and achieve state-of-the-art performance on the UCF-Crime and XD-Violence datasets, with AUC and average precision (AP) scores of 87.24% and 84.93%, respectively. Pu et al. [20] employed TCA to enhance long-range dependencies and used PEL instead of contrastive learning to enhance correct prediction rates by reducing errors [20]. Their method incorporated an MLP with PEL for feature reduction and causal convolution (CC) for classification. PEL integrates semantic priors through knowledge-based prompts to enhance recognition rates and the discriminative capacity, ensuring high separability between anomaly subclasses. They reported impressive AUC rates of 86.76%, 85.59%, and 98.14% for the UCF-Crime, XD-Violence, and ShanghaiTech datasets, respectively. This underscores the effectiveness of their approach in improving the anomaly detection performance. Zhao et al. [18] introduced UR-DMUs to improve the performance accuracy rate, focusing on temporal feature extraction using graph-based Transformers via the I3D backbone. They achieved 86.97% and 94.02% accuracy on the UCF-Crime and XD-Violence datasets, respectively. Sharif et al. [9] later proposed a two-stream approach for temporal feature enhancement, combining CNN-based I3D and ViT-based CLIP features. They reported AUCs of 88.97% and 98.66% for the UCF-Crime and ShanghaiTech datasets but faced challenges

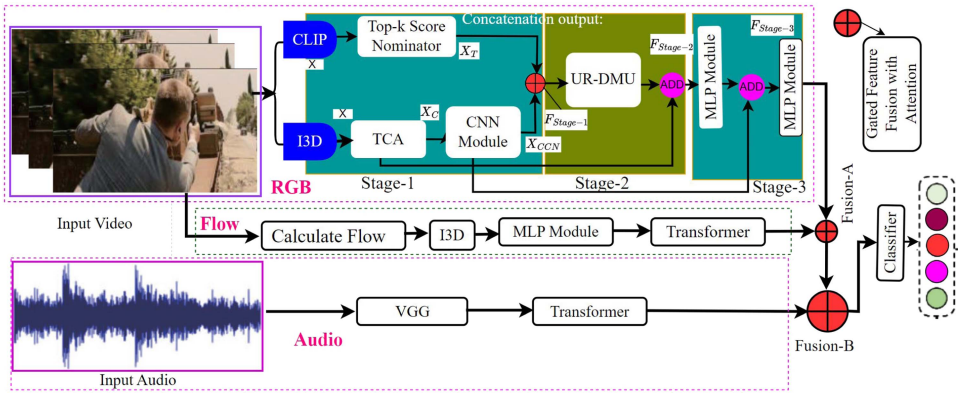


FIGURE 1. **Proposed model architecture.**

in real-time deployment due to feature effectiveness issues. These studies do not integrate graph-based and spatial feature enhancements. UR-DMUs [18] incorporated graph-based features but lacked time-varying enhancements, while TCA [16], [20] and I3D-CLIP [9] addressed temporal aspects but overlooked complementary features. However, the performance of UVAD methods are limited because they lack prior knowledge of abnormal events during training. Inspired by these gaps, we propose a novel anomaly detection system leveraging multi-stage graphs and deep learning feature enhancements. Our approach integrates CNN and ViT pre-trained features, temporal enhancements, graph-based temporal features, and spatial feature enhancements to optimize anomaly detection performance.

III. DATASETS

Anomaly detection datasets are crucial for developing and testing algorithms that identify unusual events. XD-Violence [18] is a key dataset covering diverse anomalies in terms of the scale, background, and type, providing a valuable resource for benchmarking and improving anomaly detection models. In this study, we used three datasets: one containing RGB, flow, and audio modalities, and two others—ShanghaiTech and UCF-Crime—with only RGB and flow data. These datasets enabled a comprehensive evaluation across different scenarios.

a) XD-Violence: The XD-Violence dataset contains a mix of video and audio media formats, covering diverse backgrounds like movies, games, and live scenes. It includes 4,754 videos in total, with 3,954 videos for training, each labeled at the video level. Additionally, there are 800 testing videos labeled frame by frame [18].

b) Other Datasets: We also used the ShanghaiTech and UCF-Crime datasets, which include RGB and flow data but no audio. The ShanghaiTech dataset contains 317,398 frames from various locations on the ShanghaiTech campus, with 307 normal and 130 anomaly videos across 13 scenes. Originally a benchmark for video anomaly detection, the dataset was reorganized by Zhong et al. to create a weakly supervised training set. We followed their approach for our experiments [5]. The UCF-Crime dataset consists of 1,900 untrimmed videos, totaling 128 hours, featuring 13 types of real-world anomalies like arson, burglary, and robbery. It offers more complex backgrounds than ShanghaiTech. The training set has 1,610 videos (800 normal, 810 anomalous), while the testing set includes 290 videos with frame-level labels [19].

IV. PROPOSED METHOD

In our proposed model for multimodal-based anomaly detection, we leverage a sophisticated combination of state-of-the-art technologies to enhance the accuracy and robustness of anomaly identification. Fig. 1 demonstrates the proposed model, where we used a multi-stage deep learning (DL) approach [18], [39], [57]. There are three modalities used in the dataset to extract the features from the three different streams, including RGB video, flow, and audio signal feature streams. In the RGB streams, we divided the untrimmed video into non-overlapping snippets using a 16-frame sliding window for the RGB video. Then, we applied a two-stage-based feature extraction approach. The first stage used a ViT-based CLIP module, and the top k features were concatenated with I3D- and TCA-based spatiotemporal features. The second stage captures temporal dependencies via the UR-DMUs, learning normal and abnormal data representations simultaneously. The third stage selects the most relevant spatiotemporal features using a two-layer MLP, optimizing it for further analysis or applications considered first-stream features. This section of the RGB video data stream is nearly identical to that of the previous model. It loads the weights from the previous model before beginning training [19]. In the flow data stream, motion from RGB frames is computed and fed into the I3D module to capture spatial and temporal information, highlighting scene dynamics critical for detecting anomalies. These motion features are then processed by an MLP for compact representation and a Transformer to capture long-range dependencies, producing the second-stream features. The audio stream utilizes a Transformer integrated with a VGGish model to extract nuanced features from audio signals, crucial for detecting anomalies through unusual audio patterns not evident in visual data. The VGGish model processes the audio

input to produce detailed feature representations, while the Transformer captures temporal dependencies and contextual relationships. This combination enhances the model's ability to detect subtle audio anomalies, complementing the visual streams for robust detection. Finally, we concatenated the features using the gated feature function with attention technique, and the output was fed into the classification module in two ways: the concatenated features of the RGB video and flow modalities and the concatenated features of the RGB video, flow, and audio signal-based features.

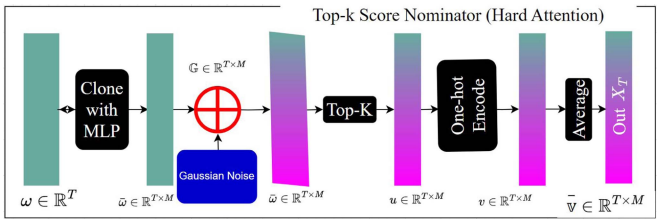


FIGURE 2. Internal structure of the top-k score selection module [19], [57].

A. PREPROCESSING

In WS-VAD, the training set consists solely of video-level labels. The set of training videos can be expressed as $W = \{(V_v, y_v)\}_{v=1}^{w}$, where each video $V_v = \{\text{Frame}_i\}_{i=1}^{N_v} \in \mathbb{R}^{N_v \times W \times H}$ represents a sequence of frames N_v, and each frame has a width W and a height H. The label of each video V_v, denoted as $y_v \in \{0, 1\}$, indicates the presence of an anomaly. We divided each video into a set of snippets, expressed as $\{\gamma_i\}_{i=1}^{\lfloor \frac{N_v}{\Delta} \rfloor}$, where each snippet contains an equal number of frames Δ. In the preprocessing step, we followed the existing methodology. First, we divided the untrimmed video into non-overlapping snippets using a 16-frame sliding window [18], [20], [57]. Then, we extracted features from each sample using 5-crop augmentation for the XD-Violence dataset, utilizing pre-trained models in the initial stage [18], [20], [57].

B. RGB DATA MODALITY STREAM

We considered the dynamic or RGB video as the data modality of the first stream, which was constructed with three stages: stage 1—initial feature, stage 2—feature enhancement, and stage 3—feature reduction.

1) STAGE 1: PRE-TRAINED MODEL-BASED FEATURE EXTRACTION

In the first stage, we introduce a multi-backbone framework, combining a CLIP model trained on Kinetics with an I3D model pre-trained on Kinetics. It is important to note that our I3D model extracted RGB and flow features. In this architecture, the I3D RGB model extracts features in 1024-dimensional space with 1024-dimensional flow features, while the CLIP model provides feature vectors in 512 dimensions. This dual-backbone approach leverages the strengths of both architectures to enhance the feature extraction process for video anomaly detection.

a) Top-k Score Nominator Selection from CLIP Transformer Features: This section of stage 1 integrates the CLIP pre-trained approach features with the top-k score nominator. Here, CLIP leverages ViTs to capture the correlation among the frames, mainly extracting the intricate internal relationship among the frames. The main concept of the CLIP model is that it is composed of a multi-backbone framework [58]. In the study, we considered $d_j = \lceil \frac{\Delta}{2} \rceil$ as the middle frame of the video snippet γ_j, which means that we did not consider all frames simultaneously. In our study, we applied the CLIP model to d_j of the snippet γ_j to represent its features as $\phi_{v_j} \in \mathbb{R}^{\aleph}$, where $\aleph$ represents the feature dimension. The final feature vector is constructed as $\phi_{v_{\text{vit}}} = \{\phi_j\}_{j=1}^{T_v} \in \mathbb{R}^{T\times\aleph}$ [9]. It produces a separate feature for each video, and each feature's dimension size is 512. The output of the CLIP [59] model is fed into the top-k score selection module, shown in Fig. 2. This module selects the most relevant video snippets based on the top-k score nominator, as described by Joo et al. [57]. The process involves cloning the CLIP model output, adding Gaussian noise, and calculating the magnitude. The top k scores are then selected to focus on the most significant parts of the video.

b) Spatial-Temporal Feature Enhancement with I3D and TCA with CNN: In this section of stage 1, we employed I3D [8] to enhance the frame-wise spatial features for the anomaly video data; then, we applied a TCN to improve the temporal features and generate spatial-temporal features. After that, the output is fed into the CNN module to improve the spatial features from the spatial-temporal features. I3D is a type of 3DCNN [60] used to extract 2D or 3D features from dynamic or video data to capture both spatial and temporal features from successive frames. In the study, the input comes from the RGB video, and the feature extraction process is expressed as (1):

$$\mathbf{F}^{I3D} = \text{I3D}(\mathbf{V}_{RGB}), \tag{1}$$

where $\mathbf{V}_{RGB}$ denotes the input RGB video frames and $\mathbf{F}^{I3D}$ represents the feature map generated by the I3D model. The TCA module took the output of the I3D feature $\mathbf{F}^{I3D}$ as the input to enhance the temporal features [20]. It mainly uses a self-attention mechanism to incorporate long-range temporal information between frame-level features by extracting strong relationships among the consecutive frames [16], [20]. Fig. 3 illustrates the TCA calculation procedure. The output of the I3D module $\mathbf{F}^{I3D}$ we considered here as X, which is projected in the latent space using linear layers to produce the similarity matrix M [19].

Then, we enhanced the similarity matrix using dynamic position encoding (DPE), and after calculating the local attention and context features T masked with similarity matrix [16], [19], [20]. The final feature X^o is obtained by combining global and local attention heads; then, after normalizing, we concatenate with a skip connection and use a linear layer to

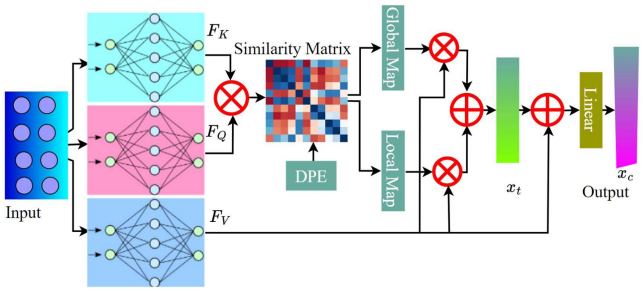


FIGURE 3. Working structure of the TCA module [19], [20].

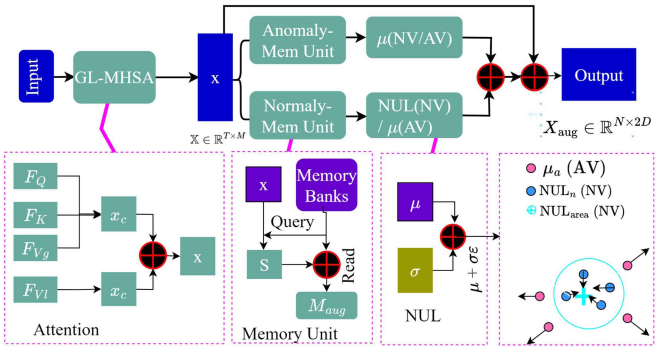


FIGURE 4. Working diagram of the UR-DMU module [18], [19] .

produce the TCA module output [19]:

$$X^c = \text{LN}(X + f_h(\text{Norm}(X^o))), \tag{2}$$

where LN represent the linear layer and Norm(·) denotes normalization.

Integrating the I3D module with the TCA module mainly extracts robust spatiotemporal contextual information from the video sequence that helps us capture the motion and appearance cues. TCA plays a pivotal role in integrating contextual information across multiple frames. TCA enhances the model's ability to discern anomalies by considering temporal dependencies within video sequences. This mechanism ensures that the model can effectively capture dynamic changes over time, improving the anomaly detection accuracy. Incorporating a 1D CNN, followed by ReLU activation and dropout regularization, contributes to feature dimensionality reduction while preserving essential information. This process ensures that the extracted features are concise yet informative, facilitating efficient anomaly detection without sacrificing discriminative power.

c) Feature Fusion: In the first stream, we employed the top-k score nominator [57] to select the top k segments based on their CLIP feature relevance, resulting in a refined set of 512-dimensional features denoted as X_T. In the second stream, we obtained the final feature from the FC module, denoted as X_{CCN}. These features were then concatenated, producing comprehensive 1024-dimensional features, denoted as $F_{stage-1}$, using the following equation:

$$F_{stage-1} = X_T \oplus X_{CCN}. \tag{3}$$

2) STAGE 2: UR-DMU-BASED FEATURES

We applied the UR-DMU module to enhance the fused features, which come from the attention-based temporal enhancement [18], [44], [46], [61], [62]. There are three main components in the UR-DMU shown in Fig. 4. Based on the GCN feature, it was constructed with global and local multi-head self-attention (GL-MHSA) to extract local and global dependencies as effective features. For training, videos with normal and abnormal footage are processed. The model generates a score for each snippet using the BCE loss and auxiliary losses. During testing, the mean-encoder network of the DUL module produces feature embeddings, which label video snippets to produce UR-DMU features F_{Urdmu} [19]. The final feature of stage 2, $X_{Stage-2}$, is obtained by adding F_{Urdmu} and TCA X^c:

$$X_{Stage-2} = F_{Urdmu} + X^c. \tag{4}$$

3) STAGE 3: FEATURE REDUCTION MLP MODULE

To select effective features from the graph-based UR-DMU $F_{stage-2}$ features, we employed a two-layer MLP for feature reduction. The MLP enables high-level semantic representations and non-linear feature transformation for anomaly detection. It includes two Conv1d layers, two GELU activations, and two dropout mechanisms [63]. Features from TCA are integrated before the first Conv1d layer, and a 512-dimensional feature from I3D is appended afterward. Each Conv1D layer is followed by GELU activation and dropout, as shown in (5).

$$\begin{aligned} F_{\text{MLP-1}} &= \text{Dropout}(\text{GELU}(\text{Conv1D}(F_{\text{stage-2}}))), \\ F_{\text{Stage-3}} &= F_{\text{MLP-1}} \oplus X_{\text{CNN}}, \\ F_{\text{Stage-3}} &= \text{Dropout}(\text{GELU}(\text{Conv1D}(F_{\text{Stage-3}}))). \end{aligned} \tag{5}$$

Finally, a causal convolution layer produces anomaly scores by integrating present and past observations, which is represented as follows:

$$RGB_{Feature}(F) = \sigma\left(f_t(F_{\text{Stage-3}})\right), \tag{6}$$

where $f_t(\cdot)$ denotes the causal convolution layer with a kernel size of Δt, and $\sigma(\cdot)$ is the sigmoid activation function. The output of the MLP is considered a final feature, denoted as $RGB_{Feature}$.

C. FLOW DATA MODALITY STREAM

The flow dataset stream took RGB video as input; this input was first processed to compute the optical flow, capturing motion dynamics between consecutive frames using TV-L_1 [64]. The optical flow between frames t and $t+1$ is given by

$$\mathbf{F}_{\text{flow}} = \text{TV-L}_1(\mathbf{V}_t^{RGB}, \mathbf{V}_{t+1}^{RGB}), \tag{7}$$

where $\mathbf{F}_{\text{flow}}$ represents the movement and temporal changes in the video sequence.

a) I3D Feature Extraction: After calculating the optical flow, we apply the inflated 3D ConvNet (I3D) model to extract

spatio-temporal features from the flow data. The I3D features are generated as follows:

$$\mathbf{F}_{\text{I3D}} = \text{I3D}(\mathbf{F}_{\text{flow}}), \tag{8}$$

where $\mathbf{F}_{\text{I3D}}$ captures the complex motion and spatial relationships in the flow information.

b) MLP Module: The features extracted by the I3D model are then refined through an MLP module to enhance their representation for subsequent processing. The MLP transforms the I3D features as follows:

$$\mathbf{F}_{\text{MLP}} = \text{ReLU}(\mathbf{W}_1 \cdot \mathbf{F}_{\text{I3D}} + \mathbf{b}_1), \tag{9}$$

where $\mathbf{W}_1$ and $\mathbf{b}_1$ are the weights and biases of the MLP, and ReLU$(\cdot)$ is the activation function. The output $\mathbf{F}_{\text{MLP}}$ is a refined representation of the I3D features.

c) Transformer Module: The refined features from the MLP are then fed into a Transformer model, which uses self-attention mechanisms to capture complex temporal dependencies within the feature information. The processing by the Transformer is expressed as follows:

$$\mathbf{F}_{\text{Trans}} = \text{Transformer}(\mathbf{F}_{\text{MLP}}), \tag{10}$$

where $\mathbf{F}_{\text{Trans}}$ represents the feature map after Transformer processing. The final feature representation, $\mathbf{F}_{\text{Trans}}$, referred to as the "Feature of Flow Modality," encapsulates the enriched temporal dynamics and motion information, making it suitable for downstream tasks such as classification, segmentation, or anomaly detection, providing a comprehensive analysis of the RGB video data. This approach is advantageous because it effectively highlights motion-related anomalies, providing complementary information to the RGB stream. The novelty of this stream lies in the integration of motion flow information with advanced Transformer-based processing, enhancing the model's ability to detect subtle and complex anomalies.

D. AUDIO DATA MODALITY STREAM

VGGish is a deep convolutional neural network that effectively extracts features from audio signals by treating the spectrogram of the audio signal as an image. VGGish is applied for feature extraction from audio signals as described below.

a) Audio Preprocessing: The audio signal is first converted into a spectrogram, which represents the frequency content of the signal over time. It is usually computed using the short-time Fourier transform (STFT). The spectrogram S of an audio signal $x(t)$ can be defined as

$$S(t, f) = \left| \sum_n x[n] \cdot w[n-t] \cdot e^{-j2\pi fn} \right|, \tag{11}$$

where $x[n]$ represent the input signal, $w[n]$ is a window function, t represents time, and f represents frequency.

b) Spectrogram as Input: The resulting spectrogram, a 2D representation of the audio signal, is fed into the VGGish network as if it were an image. This allows the network to leverage its pre-trained convolutional layers to extract relevant features.

c) Feature Extraction with VGGish: VGGish consists of 16 convolutional layers followed by 3 fully connected layers. For feature extraction, we typically use the outputs of one of the deeper convolutional layers or the fully connected layers. The convolutional layers perform a series of operations defined by

$$Conv(S(i, j)) = \text{ReLU}(W * S(i, j) + b), \tag{12}$$

where $*$ denotes the convolution operation, W is the filter (kernel), b is the bias, and ReLU is the activation function. $Conv(S(i, j))$ implies that the convolution operation is applied to the local region around S(i,j), where the feature map value at (i,j) is influenced by its neighbors through the kernel W. The pooling layers in VGGish help reduce the dimensionality of the feature maps while retaining important information. The equation below expresses the max-pooling operation, which produces the VGGish (S) features.

$$\text{VGGish}(S) = \max_{i,j \in R_k} S(i, j), \tag{13}$$

where R_k represents the pooling region (e.g., a 2×2 or 3×3 window) and $S(i, j)$ is the value at the position (i, j) in the feature map S.

d) Integrating VGGish and Transformer: The high-dimensional feature vector produced by VGGish encapsulates the temporal and spectral characteristics of the audio signal. This feature vector is then fed into the Transformer model to capture complex dependencies within the audio data, leveraging the Transformer's self-attention mechanism. The integration of VGGish and the Transformer enhances the ability to model both local and global temporal dependencies, making it particularly effective for anomaly detection tasks.

e) Audio Modality Feature Representation: The VGGish model processes the spectrogram through its layers, transforming it into a high-dimensional feature vector. This feature vector, denoted as $Audio_{Feature}(F)$, is further refined by the Transformer model. The final audio feature representation can be expressed as

$$Audio_{Feature}(F) = \text{Transformer}(\text{VGGish}(S)). \tag{14}$$

By integrating VGGish and the Transformer, we effectively transform raw audio data into a rich, high-dimensional feature space suitable for various audio processing tasks, especially discerning anomalies.

E. FEATURE CONCATENATION AND CLASSIFICATION

To leverage the combined information from different modalities, we perform feature concatenation followed by classification. Specifically, we first concatenate features from the RGB video modality with the flow modality and then apply a deep learning-based classification module. The process is described as follows.

a) Feature Concatenation: Let F_{rgb} represent the feature vector from the RGB video modality and F_{flow} represent the feature vector from the flow modality. The concatenated

feature vector F_{concat} is given by

$$F_{concat} = [F_{rgb}; F_{flow}], \tag{15}$$

where [;] denotes concatenation along the feature dimension.

b) Extended Feature Concatenation: For scenarios involving additional modalities, such as combining the RGB video, flow, and audio modalities, let F_{audio} represent the feature vector from the audio modality. The extended concatenated feature vector $F_{extended}$ is given by

$$F_{extended} = [F_{rgb}; F_{flow}; F_{audio}]. \tag{16}$$

F. CLASSIFICATION AND EXPERIMENT PROCEDURE

The concatenated feature vector F_{concat} is then passed through a deep learning-based classification module, which produces the final classification score. Let $\mathbf{W}_{clf}$ and $\mathbf{b}_{clf}$ be the weights and biases of the classification layer. The classification score S_{concat} is computed as follows in 17.

$$S_{concat} = \text{Softmax}(\mathbf{W}_{clf} \cdot F_{concat} + \mathbf{b}_{clf}). \tag{17}$$

Similarly, this extended feature vector $F_{extended}$ is used in the classification module to produce the performance accuracy. The classification score $S_{extended}$ is

$$S_{extended} = \text{Softmax}(\mathbf{W}_{clf} \cdot F_{extended} + \mathbf{b}_{clf}), \tag{18}$$

where Softmax converts the output logits into probabilities. To evaluate the performance, we use the area under the curve (AUC) metric with MIL and magnitude contrastive (MC) loss functions. During the training experiment, we optimize the objective function

$$L = L_{\text{ce}} + \lambda L_{\text{kd}}, \tag{19}$$

where L_{ce} represents the cross-entropy loss, L_{kd} denotes the knowledge distillation loss, and λ is a hyperparameter that balances these losses. This formulation enhances the model's ability to differentiate between positive and negative snippets by improving discriminative representations. During testing, we apply score smoothing (SS) to reduce transient noise and false alarms:

$$\tilde{s}_i = \frac{1}{\kappa} \sum_{j=i}^{i+\kappa-1} s_j. \tag{20}$$

Here, $\tilde{s}_i$ represents the smoothed score for the i-th snippet, and κ is the smoothing window size. This approach suppresses noise and biases, resulting in more stable prediction scores.

V. EXPERIMENTAL EVALUATION

We evaluated the proposed model using three anomaly datasets with various modalities. In the section below, we first describe the datasets and then include the performance accuracy for each dataset.

1) Environmental Setup and Evaluation Metrics: The system was built with a GeForce RTX 4090 24 GB GPU, CUDA version 11.7, NVIDIA driver 515, and 64 GB of RAM. The training utilized two learning rates of 0.00003 for the flow data flow and the audio data flow and 0.000001 for the RGB data flow and a batch size of 32, and it ran for two epochs using the Adam optimizer on the RTX 4090. For efficient graph convolution and attention with a low computational cost, the Python environment included OpenCV, Pickle, Pandas, and PyTorch for model development [65]. These packages, along with others [66], [67], facilitated the initial data processing and model development.

A. ABLATION STUDY

Table 1 presents the ablation study results, illustrating the impact of various components in our proposed model, including multi-backbone pre-trained models. The check marks indicate the inclusion of specific technologies in each experiment. The introduction of multi-head self-attention (MHSA) significantly boosts the AUC to 79.71%, underscoring the importance of attention mechanisms for focusing on critical information. Adding the dual memory unit (DMU) further improves the AUC to 79.74%, and incorporating memory combination (MC) raises it to 86.37%. The highest accuracy, 86.48%, is achieved by applying SS alongside these technologies. Extending the model with the RGB modality scores 86.37%, and adding the flow modality with a gated feature fusion with attention module increases the accuracy to 87.32%. We initialized the RGB weights from the model that achieved a score of 86.37%, minimizing the retraining needs. Similarly, incorporating the audio modality along with the flow and initializing RGB weights as before resulted in an even higher accuracy of 88.32%, highlighting the value of the audio modality. The ablation study demonstrates the effectiveness of combining these methodologies, showcasing the robustness and improvements in video anomaly detection provided through our multimodal approach.

B. PERFORMANCE RESULTS AND STATE-OF-THE-ART COMPARISON FOR XD-VIOLENCE DATASET

Table 2 demonstrates the performance of the proposed model. The proposed model demonstrates high performance on the XD-Violence dataset, with an AUC of 95.84%, an anomaly AUC of 86.92%, an AP of 88.28%, and an FAR of 0.0014%. These results highlight the model's accuracy and reliability in detecting anomalies. Table 3 demonstrates the performance and state-of-the-art comparison of the proposed model for video anomaly detection on the XD-Violence dataset, highlighting the performance of the proposed multimodal model compared to state-of-the-art models.Despite significant advancements in models for video anomaly detection, the previously mentioned existing studies have exhibited several limitations. For instance, HL-Net [10] achieved an AP of 78.64% using RGB and audio features, but it failed to leverage multiple data modalities like the flow modality, limiting its capacity to detect dynamic scene changes and subtle motion variations. Pang et al. [11] improved the performance to an AP of 81.69% with RGB and audio features yet did not incorporate the temporal dependencies essential for understanding complex interactions in videos. Similarly, ACF [12]

TABLE 1. Ablation Study Performance in Terms of AUC (%)

Data Modality	I3D	TCA	CLIP	Top-K	MHSA	DMU	PEL	MC	SS	XD (%)	UCF-Crimes (%)	Shanghai Tech (%)
RGB	✓	✓	✓	✓	✓					79.71		
RGB	✓	✓	✓	✓	✓	✓				79.74		
RGB	✓	✓	✓	✓	✓	✓	✓			83.68		
RGB	✓	✓	✓	✓	✓	✓	✓	✓		86.37		
RGB	✓	✓	✓	✓	✓	✓	✓	✓	✓	86.48	90.09	98.69
RGB+Flow	✓	✓	✓	✓	✓	✓	✓	✓	✓	87.32	90.26	98.71
RGB+Flow+Audio	✓	✓	✓	✓	✓	✓	✓	✓	✓	88.28		

TABLE 2. Performance Results

Dataset Name	AUC (%)	Anomaly AUC (%)	AP (%)	FAR (%)
XD-Violence	95.84	86.92	88.28	0.0014

TABLE 3. State-of-The-Art Comparison of the Proposed Model for the XD-Violence Dataset

Method	Features	AP (%)
HL-Net [10]	RGB+Audio	78.64
Pang et al. [11]	RGB+Audio	81.69
ACF [12]	RGB+Audio	80.13
MSAF [32]	RGB+Audio	80.51
CUPL [13]	RGB+Audio	81.43
MACIL-SD (Light) [15]	RGB+Audio	82.17
CMA-LA [14]	RGB+Audio	83.54
MACIL-SD(Full) [15]	RGB+Audio	83.40
MACIL-SD (Light)* [15]	RGB+Flow	78.49
MACIL-SD(Full)* [15]	RGB+Flow	79.73
MACIL-SD (Light)* [15]	Audio+Flow	76.58
MACIL-SD(Full)* [15]	Audio+Flow	77.06
Wu et al. [33]	Audio+Flow	72.96
Wu et al. [33]	RGB+Audio+Flow	79.53
Xiao et al. [34]	RGB+Audio+Flow	83.09
MSBT et al. [35]	Audio+Flow	77.47
MSBT et al. [35]	RGB+Flow	80.68
MSBT et al. [35]	RGB+Audio	82.54
MSBT et al. [35]	RGB+Audio+Flow	84.32
Shin et al. [19]	RGB	86.26
Proposed Multimodal	RGB+Flow+Audio	88.28

and MSAF [32] achieved APs of 80.13% and 80.51%, respectively, but their models were constrained by the absence of temporal information processing, affecting their detection capabilities for movement patterns over time. Models like MACIL-SD (Light/Full) [15], though competitive, with APs up to 83.40%, relied on simplistic fusion strategies that may not fully exploit the complementary information across different modalities. Furthermore, studies such as Wu et al. [33] and Xiao et al. [34] showed limitations in integrating and processing complex multimodal data, as evidenced by their relatively lower performance. Even Shin et al. [19], whose model achieved an AP of 86.26% using RGB features alone, did not utilize the advantages of a multimodal approach, potentially missing out on the enhanced detection accuracy that could be gained from combining various data sources. In contrast, the proposed multimodal model addresses these limitations by integrating RGB, flow, and audio features, capturing both spatial and temporal information crucial for detecting anomalies in complex and dynamic video scenes. This model employs an advanced feature fusion strategy that enhances the detection capability by leveraging the strengths of each modality, surpassing earlier models that relied on less effective fusion methods. The result is a significant improvement in performance, with the proposed model achieving the highest AP of 88.28% on the XD-Violence dataset, demonstrating superior accuracy and reliability. Additionally, the model's ability to incorporate diverse features from multiple modalities enhances its robustness and adaptability to various real-world scenarios characterized by complex interactions and unpredictable motion patterns. By effectively handling temporal dependencies through the inclusion of flow features, the proposed model excels in identifying subtle anomalies that other models may overlook, providing a more comprehensive and accurate solution for video anomaly detection.

C. STATE-OF-THE-ART COMPARISON FOR OTHER DATASETS

We also evaluated the proposed model using the ShanghaiTech dataset and UCF-Crime dataset. Our model is designed to incorporate three modality features: RGB video, flow, and audio information. However, due to the limited availability of audio modality datasets, we focused our evaluation on the RGB video and flow modalities. Table 4 presents a state-of-the-art comparison of these datasets, demonstrating the effectiveness of our approach. The proposed multimodality model achieved an AUC of 98.71% on the ShanghaiTech dataset and an AUC of 90.26% on the UCF-Crime dataset, outperforming existing state-of-the-art models. This strong performance highlights the significant advantage of integrating multiple data modalities. By leveraging both RGB and flow information, our model can capture more complex patterns and nuances in the data, leading to more accurate anomaly detection. This approach is particularly

TABLE 4. State-of-The-Art Comparison of the Proposed Model for the UCF-Crime and ShanghaiTech Datasets

Model Name	Data Modalities		Year	Feature Extractor Name	ShanghaiTech Dataset	UCF-Crime Dataset
	Video	Flow			AUC (%)	AUC (%)
Yu et al. [17]	✓	-	2022	I3D	89.91	83.75
Gong et al. [26]	✓	-	2022	I3D	90.10	81.00
Majhi et al. [27]	✓	-	2023	13D-Res	96.22	85.30
Park et al. [28]	✓	-	2023	C3D	96.02	83.43
Park et al. [28]	✓	-	2023	I3D	97.43	85.63
Pu et al. [20]	✓	-	2023	I3D	98.14	86.76
Lv et al. [29]	✓	-	2023	X-CLIP	96.78	86.75
Sun et al. [30]	✓	-	2023	C3D	96.56	83.47
Sun et al. [30]	✓	-	2023	I3D	97.92	85.88
Wang et al. [31]	✓	-	2023	C3D	94.01	81.48
Sharif et al. [9]	✓	-	2023	I3D+CLIP	98.66	88.97
Shin et al. [19]	✓	-	2024	Hybrid model	98.69	90.00
Proposed Model	✓	✓	-	Multimodality	98.71	90.26

effective in diverse and dynamic environments, where relying on a single modality might result in missed or inaccurate detections. Our multimodal model's superior accuracy and robustness make it a strong candidate for real-world applications in surveillance and security, offering a more comprehensive solution than models that rely on single data types.

VI. CONCLUSION

This study presented a comprehensive multimodal deep learning model for weakly supervised video anomaly detection (WS-VAD), leveraging RGB video, optical flow, and audio data modalities. Our model integrates advanced feature extraction techniques, including a ViT-based CLIP module, CNN-based I3D with temporal context aggregation (TCA), and uncertainty-resilient dual memory units (UR-DMUs) with global/local multi-head self-attention (GL-MHSA) and a Transformer. These components, coupled with a multilayer perceptron (MLP) for feature refinement, significantly enhance the model's ability to distinguish between normal and abnormal behaviors. Each modality contributes uniquely: the RGB stream captures visual semantics, the flow stream emphasizes dynamic motion, and the audio stream detects anomalies through sound patterns. Integrating these streams via a gated feature fusion mechanism with an attention module creates a robust classifier that effectively predicts snippet-level anomaly scores and converts them into bag-level predictions during training. Extensive experiments on three benchmark datasets demonstrate that our model surpasses existing state-of-the-art approaches, delivering a high accuracy and robust performance. However, the model's dependence on high-quality, multimodal data introduces limitations, particularly in scenarios where one or more modalities may be unavailable or noisy. Future work will focus on addressing these limitations by exploring lightweight, real-time models and adaptive learning techniques, expanding the integration of modalities, and investigating transfer learning approaches to enhance robustness across diverse surveillance scenarios. Overall, our model shows great promise for real-world applications, offering a reliable and efficient solution for intelligent surveillance systems.

REFERENCES

[1] K. Liu and H. Ma, "Exploring background-bias for anomaly detection in surveillance videos," in *Proc. 27th ACM Int. Conf. Multimedia*, Nice, France, 2019, pp. 1490–1499.

[2] H. Md Sharif, L. Jiao, and C. W. Omlin, "Deep crowd anomaly detection by fusing reconstruction and prediction networks," *Electronics*, vol. 12, no. 7, 2023, Art. no. 1517.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.

[4] M. Z. Zaheer, A. Mahmood, M. Astrid, and S.-I. Lee, "CLAWS: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection," in *Proc. 16th Euro. Conf. Comput. Vis.–ECCV* 2020, Glasgow, U.K., Aug. 23–28, 2020, pp. 358–376.

[5] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 1237–1246.

[6] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 6479–6488.

[7] Y. Liu et al., "Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models," *ACM Comput. Surv.*, vol. 56, no. 7, pp. 1–38, 2024.

[8] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 4724-4733.

[9] H. Md Sharif, L. Jiao, and C. W. Omlin, "CNN-ViT supported weakly-supervised video segment level anomaly detection," *Sensors*, vol. 23, no. 18, 2023, Art. no. 7734.

[10] P. Wu et al., "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Proc. 16th Euro. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 23–28, 2020, pp. 322–339.

[11] G. Pang, C. Yan, C. Shen, A. van den Hengel, and X. Bai, "Self-trained deep ordinal regression for end-to-end video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 12173–12182.

[12] D.-L. Wei, C.-G. Liu, Y. Liu, J. Liu, X.-G. Zhu, and X.-H. Zeng, "Look, listen and pay more attention: Fusing multi-modal information for video violence detection," in *Proc. 2022 IEEE Int. Conf. Acoust., Speech Signal Process.*, Singapore, 2022, pp. 1980–1984.

[13] C. Zhang et al., "Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, 2023, pp. 16271–16280.

[14] Y. Pu and X. Wu, "Audio-guided attention network for weakly supervised violence detection," in *Proc. 2nd Int. Conf. Consum. Electron. Comput. Eng.*, Guangzhou, China, 2022, pp. 219–223.

[15] J. Yu, J. Liu, Y. Cheng, R. Feng, and Y. Zhang, "Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection," in *Proc. 30th ACM Int. Conf. Multimedia*, Lisboa, Portugal, 2022, pp. 6278–6287.

[16] J. Shao, X. Wen, B. Zhao, and X. Xue, "Temporal context aggregation for video retrieval with contrastive learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3268–3278.

[17] S. Yu, C. Wang, L. Xiang, and J. Wu, "TCA-VAD: Temporal context alignment network for weakly supervised video anomly detection," in *Proc. 2022 IEEE Int. Conf. Multimedia Expo*, Taipei, Taiwan, 2022, pp. 1–6.

[18] H. Zhou, J. Yu, and W. Yang, "Dual memory units with uncertainty regulation for weakly supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, Washington, DC, USA, 2023, vol. 37, pp. 3769–3777.

[19] J. Shin, Y. Kaneko, A. S. M. Miah, N. Hassan, and S. Nishimura, "Anomaly detection in weakly supervised videos using multistage graphs and general deep learning based spatial-temporal feature enhancement," *IEEE Access*, vol. 12, pp. 65213–65227, 2024.

[20] Y. Pu, X. Wu, L. Yang, and S. Wang, "Learning prompt-enhanced context features for weakly-supervised video anomaly detection," in *Proc. IEEE Trans. Image Process.*, 2024.

[21] J. Zhang, L. Qing, and J. Miao, "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection," in *Proc. 2019 IEEE Int. Conf. Image Process.*, Taipei, Taiwan, 2019, pp. 4030–4034.

[22] B. Wan, Y. Fang, X. Xia, and J. Mei, "Weakly supervised video anomaly detection via center-guided discriminative learning," in *Proc. 2020 IEEE Int. Conf. Multimedia Expo*, London, U.K., 2020, pp. 1–6.

[23] D. Purwanto, Y.-T. Chen, and W.-H. Fang, "Dance with self-attention: A new look of conditional random fields on anomaly detection in videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, 2021, pp. 173–183.

[24] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, 2021, pp. 4955-4966.

[25] S. Majhi, S. Das, and F. Brémond, "DAM: Dissimilarity attention module for weakly-supervised video anomaly detection," in *Proc. 17th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, 2021, pp. 1–8.

[26] Y. Gong, C. Wang, X. Dai, S. Yu, L. Xiang, and J. Wu, "Multi-scale continuity-aware refinement network for weakly supervised video anomaly detection," in *Proc. 2022 IEEE Int. Conf. Multimedia Expo*, Taipei, Taiwan, 2022, pp. 1–6.

[27] S. Majhi, R. Dai, Q. Kong, L. Garattoni, G. Francesca, and F. Bremond, "Human-scene network: A novel baseline with self-rectifying loss for weakly supervised video anomaly detection," 2023, *arXiv:2301.07923*.

[28] S. Park, H. Kim, M. Kim, D. Kim, and K. Sohn, "Normality guided multiple instance learning for weakly supervised video anomaly detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Waikoloa, HI, USA, 2023, pp. 2664–2673.

[29] H. Lv, Z. Yue, Q. Sun, B. Luo, Z. Cui, and H. Zhang, "Unbiased multiple instance learning for weakly supervised video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, 2023, pp. 8022–8031.

[30] S. Sun and X. Gong, "Long-short temporal co-teaching for weakly supervised video anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2023, pp. 2711–2716.

[31] L. Wang, X. Wang, F. Liu, M. Li, X. Hao, and N. Zhao, "Attention-guided MIL weakly supervised visual anomaly detection," *Measurement*, vol. 209, 2023, Art. no. 112500.

[32] D. Wei, Y. Liu, X. Zhu, J. Liu, and X. Zeng, "MSAF: Multimodal supervise-attention enhanced fusion for video anomaly detection," *IEEE Signal Process. Lett.*, vol. 29, pp. 2178–2182, 2022.

[33] P. Wu, X. Liu, and J. Liu, "Weakly supervised audio-visual violence detection," *IEEE Trans. Multimedia*, vol. 25, pp. 1674–1685, 2023.

[34] Y. Xiao, G. Gao, L. Wang, and H. Lai, "Optical flow-aware-based multimodal fusion network for violence detection," *Entropy*, vol. 24, no. 7, 2022, Art. no. 939.

[35] S. Sun and X. Gong, "Multi-scale bottleneck transformer for weakly supervised multimodal violence detection," 2024, *arXiv:2405.05130*.

[36] Y. Zhu and S. Newsam, "Motion-aware feature for improved video anomaly detection," 2019, *arXiv:1907.10211*.

[37] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 4489–4497.

[38] A. S. M. Miah, J. Shin, A. M. Md Hasan, and Md A. Rahim, "BenSignNet: Bengali sign language alphabet recognition using concatenated segmentation and convolutional neural network," *Appl. Sci.*, vol. 12, no. 8, 2022, Art. no. 3933.

[39] A. S. M. Miah, A. M. Md, J. Hasan, Y. Shin Okuyama, and Y. Tomioka, "Multistage spatial attention-based neural network for hand gesture recognition," *Computers*, vol. 12, no. 1, 2023, Art. no. 13.

[40] A. S. M. Miah, Md. A. M. Hasan, S.-W. Jang, H.-S. Lee, and J. Shin, "Multi-stream general and graph-based deep neural networks for skeleton-based sign language recognition," *Electronics*, vol. 12, no. 13, 2023, Art. no. 2841.

[41] S. Ullah, N. Bhatti, T. Qasim, N. Hassan, and M. Zia, "Weakly-supervised action localization based on seed superpixels," *Multimedia Tools Appl.*, vol. 80, pp. 6203–6220, 2021.

[42] J. Shin et al., "Korean sign language recognition using transformer-based deep neural network," *Appl. Sci.*, vol. 13, no. 5, 2023, Art. no. 3029.

[43] N. Hassan, A. S. M. Miah, and J. Shin, "Enhancing human action recognition in videos through dense-level features extraction and optimized long short-term memory," in *Proc. 7th Int. Conf. Electron., Commun., Control Eng.*, 2024, pp. 19–23.

[44] J. Shin, A. S. M. Miah, K. Suzuki, K. Hirooka, and Md. A. M. Hasan, "Dynamic Korean sign language recognition using pose estimation based and attention-based neural network," *IEEE Access*, vol. 11, pp. 143501–143513, 2023.

[45] A. S. M. Miah, J. Shin, Md. A. M. Hasan, Y. Okuyama, and A. Nobuyoshi, "Dynamic hand gesture recognition using effective feature extraction and attention based deep neural network," in *Proc. IEEE 16th Int. Symp. Embedded Multicore/Many-Core Syst. Chip*, 2023, pp. 241–247.

[46] A. S. M. Miah, A. M. Md Hasan, and J. Shin, "Dynamic hand gesture recognition using multi-branch attention based graph and general deep learning model," *IEEE Access*, vol. 11, pp. 4703–4716, 2023.

[47] C. Zhu, X. Li, C. Wang, B. Zhang, and B. Li, "Deep learning-based coseismic deformation estimation from InSAR interferograms," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5203610.

[48] A. S. M. Miah, J. Shin, A. M. Md, Y. Hasan Fujimoto, and A. Nobuyoshi, "Skeleton-based hand gesture recognition using geometric features and spatio-temporal deep learning approach," in *Proc. 11th Eur. Workshop Vis. Inf. Process.*, 2023, pp. 1–6.

[49] J. Lu, C. Li, X. Huang, C. Cui, and M. Emam, "Source camera identification algorithm based on multi-scale feature fusion," *Comput., Materials Continua*, vol. 80, no. 2, pp. 3047–3065, 2024.

[50] L. Kang, B. Tang, J. Huang, and J. Li, "3D-MRI super-resolution reconstruction using multi-modality based on multi-resolution CNN," *Comput. Methods Programs Biomed.*, vol. 248, 2024, Art. no. 108110.

[51] L. Wang et al., "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, Nov. 2019.

[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[53] N. Hassan, A. S. M. Miah, and J. Shin, "A deep bidirectional LSTM model enhanced by transfer-learning-based feature extraction for dynamic human activity recognition," *Appl. Sci.*, vol. 14, no. 2, 2024, Art. no. 603.

[54] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[55] A. S. M. Miah, A. M. Md, Y. Hasan Tomioka, and J. Shin, "Hand gesture recognition for multi-culture sign language using graph and general deep learning network," *IEEE Open J. Comput. Soc.*, vol. 5, pp. 144–155, 2024.

[56] Y. Zhou, Y. Qu, X. Xu, F. Shen, J. Song, and H. T. Shen, "Batchnorm-based weakly supervised video anomaly detection," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Aug. 2024, doi: 10.1109/TCSVT.2024.3450734.

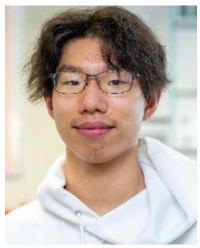
[57] H. K. Joo, K. Vo, K. Yamazaki, and N. Le, "CLIP-TSA: CLIP-assisted temporal self-attention for weakly-supervised video anomaly detection," in *Proc. IEEE Int. Conf. Image Process.*, Kuala Lumpur, Malaysia, 2023, pp. 3230–3234.

[58] H. K. Joo, K. Vo, K. Yamazaki, and N. Le, "CLIP-TSA: CLIP-assisted temporal self-attention for weakly-supervised video anomaly detection," in *Proc. IEEE Int. Conf. Image Process.*, 2023, pp. 3230–3234.

[59] Y. Qiao et al., "Robust domain generalization for multi-modal object recognition," in *Proc. 5th Int. Conf. Artif. Intell. Electromechanical Automat.*, 2024, pp. 392–397.
[60] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2012.
[61] A. S. M. Miah, A. M. Md, Y. Hasan, Y. Okuyama Tomioka, and J. Shin, "Spatial-temporal attention with graph and general neural network-based sign language recognition," *Pattern Anal. Appl.*, vol. 27, no. 2, 2024, Art. no. 37.
[62] J. Shin, A. S. M. Miah, Y. Akiba, K. Hirooka, N. Hassan, and Y. S. Hwang, "Korean sign language alphabet recognition through the integration of handcrafted and deep learning-based two-stream feature extraction approach," *IEEE Access*, vol. 12, pp. 68303–68318, 2024.
[63] W. Zhu, P. Qiu, O. M. Dumitrascu, and Y. Wang, "PDL: Regularizing multiple instance learning with progressive dropout layers," 2023, *arXiv:2308.10112*.
[64] C. Zach, T. Pock, and H. Bischof, "A duality based approach for real-time TV-L^1 optical flow," in *Proc. 29th DAGM Conf. Pattern Recognit.*, Berlin, Heidelberg, 2007, pp. 214–223.
[65] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32.
[66] S. Gollapudi, *Learn Computer Vision Using OpenCV*. Berlin, Germany: Springer, 2019.
[67] T. Dozat, "Incorporating Nesterov momentum into adam," in *Proc. 4th Int. Conf. Learn. Representations*, 2016, pp. 1–4.

JUNGPIL SHIN (Senior Member, IEEE) received the B.Sc. degree in computer science and statistics and the M.Sc. degree in computer science from Pusan National University, Busan, South Korea, in 1990 and 1994, respectively, and the Ph.D. degree in computer science and communication engineering from Kyushu University, Fukuoka, Japan, in 1999, under a scholarship from the Japanese Government (MEXT). He was an Associate Professor, a Senior Associate Professor, and a Full Professor with the School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Japan, in 1999, 2004, and 2019, respectively. He has coauthored more than 420 published papers for widely cited journals and conferences. His research interests include pattern recognition, image processing, computer vision, machine learning, human–computer interaction, non-touch interfaces, human gesture recognition, automatic control, Parkinson's disease diagnosis, ADHD diagnosis, user authentication, machine intelligence, bioinformatics, and handwriting analysis, recognition, and synthesis. He is a member of ACM, IEICE, IPSJ, KISS, and KIPS. He is also an Editorial Board Member for Scientific Reports. He was included among the top 2% of scientists worldwide in the 2024 edition of the Stanford University/Elsevier.

ABU SALEH MUSA MIAH (Member, IEEE) received the B.Sc. Eng. and M.Sc. Eng. degrees in computer science and engineering from the Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh, in 2014 and 2015, respectively, achieving the first merit position, and the Ph.D. degree in computer science and engineering from The University of Aizu, Aizuwakamatsu, Japan, in 2024, under a scholarship from the Japanese government (MEXT). He was a Lecturer and Assistant Professor with the Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology, Saidpur, Bangladesh, in 2018 and 2021, respectively. Since 2024, he has been a Visiting Researcher (Postdoc) with the University of Aizu, since 2024. He has authored and coauthored more than 50 publications in widely cited journals and conferences. His research interests include AI, ML, DL, human activity recognition, hand gesture recognition, movement disorder detection, Parkinson's disease, HCI, BCI, and neurological disorder detection.



YUTA KANEKO is currently working toward the bachelor's degree in computer science and engineering with The University of Aizu (UoA), Aizuwakamatsu, Japan. He was with the Pattern Processing Laboratory, UoA, in 2023, under the direct supervision of Dr. Jungpil Shin. He is currently working on human activity recognition. His research interests include computer vision, pattern recognition, and deep learning.

NAJMUL HASSAN (Graduate Student Member, IEEE) received the M.Sc. degree in electronics from the University of Peshawar, Peshawar, Pakistan, in 2016, and the M.Phil. degree in electronics from Quaid e Azam University Islamabad, Islamabad, Pakistan in 2018. He is currently working toward the Ph.D. degree with the School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Japan, under a scholarship from the Japanese Government (MEXT) (since fall 2023). He was a Visiting Researcher with the Department of Electronics, Quaid e Azam University Islamabad, in 2022. He has authored and coauthored more than seven publications published in widely cited journals and conferences. His research interests include human action recognition, human gesture recognition, Alzheimer's disease diagnosis, and image processing algorithms dealing with special images like underwater images, nighttime images, and foggy images.

HYOUN-SUP LEE received the B.S., M.S., and Ph.D. degrees from Dong-Eui University, Pusan, South Korea, in 2004, 2006, and 2017, respectively. From 2012 to 2015, he was a CTO with Albam Company Ltd. Since 2014, he has been a Professor with the Department of Game Engineering, Dong-Eui University. His research interests include Big Data, data analysis, and artificial intelligence.

SI-WOONG JANG received the B.S., M.S., and Ph.D. degrees from Pusan National University, Pusan, South Korea, in 1984, 1993, and 1996, respectively. From 1986 to 1993, he was a Research Worker with Daewoo Telecom Company Ltd. Since 1996, he has been a Professor with the Department of Computer Engineering, Dong Eui University, Busan, South Korea. His research interests include image processing, artificial intelligence, and the Internet of Things.