IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 March 2024, accepted 22 April 2024, date of publication 30 April 2024, date of current version 14 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3395329

RESEARCH ARTICLE

Anomaly Detection in Weakly Supervised Videos Using Multistage Graphs and General Deep Learning Based Spatial–Temporal Feature Enhancement

JUNGPIL SHIN, (Senior Member, IEEE), YUTA KANEKO,
ABU SALEH MUSA MIAH, (Member, IEEE),
NAJMUL HASSAN, (Graduate Student Member, IEEE),
AND SATOSHI NISHIMURA, (Member, IEEE)
School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu 965-8580, Japan

Corresponding author: Jungpil Shin (jpshin@u-aizu.ac.jp)

This work was supported by the Competitive Research Fund of The University of Aizu, Japan.

ABSTRACT Weakly supervised video anomaly detection (WS-VAD) is a crucial research domain in computer vision for the implementation of intelligent surveillance systems. Many researchers have been working to develop WS-VAD systems using various technologies by assessing anomaly scores. However, they are still facing challenges because of lacking effective feature extraction. To mitigate this limitation, we propose a multi-stage deep-learning model for separating abnormal events from normality to extract the hierarchical effective features. In the first stage, we extract two stream features using pre-trained techniques: the first stream employs a ViT-based CLIP module to select top-k features, while the second stream utilizes a CNN-based I3D module integrated into the Temporal Contextual Aggregation (TCA) mechanism. These features are concatenated and fed into the second-stage module, where an Uncertainty-regulated Dual Memory Units (UR-DMU) model is employed to learn representations of regular and abnormal data simultaneously. The UR-DMU integrates global and local structures, leveraging Graph Convolutional Networks (GCN) and Global and Local Multi-Head Self Attention (GL-MHSA) modules to capture video associations. Subsequently, feature reduction is achieved using the multilayer-perceptron (MLP) integration with the Prompt-Enhanced Learning (PEL) module via the knowledge-based prompt. Finally, we employed a classifier module to predict the snippet-level anomaly scores. In the training phase, the based function transfers the snippet-level scores into bag-level predictions for learning high activation in anomalous cases. Our approach integrates these cutting-edge technologies and methodologies, offering a comprehensive solution to video-based anomaly detection. Extensive experiments on ShanghaiTech, XD-Violence, and UCF-Crime datasets validate the superiority of our method over state-of-the-art approaches by a substantial margin. We believe that our model holds significant promise for real-world applications, demonstrating superior performance and efficacy in anomaly detection tasks.

INDEX TERMS Temporal contextual aggregation (TCA), uncertainty-regulated dual memory units (UR-DMU), graph convolutional networks and global/local multi-head self-attention (GL-MHSA), weakly supervised video anomaly detection (WS-VAD) anomaly detection.

I. INTRODUCTION

Fully supervised, unsupervised, and weakly supervised are the three prevailing paradigms in the field of video anomaly

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

event detection (VAED). The fully supervised paradigm is primarily characterized by its exceptional performance [1]. However, it is important to note that the training data for this paradigm necessitates the inclusion of frame-level normal or abnormal annotations, which in turn requires video annotators to identify and label abnormalities within the videos. Given that abnormalities can occur at any given moment, it becomes imperative for the annotators to spot nearly all frames. Regrettably, the process of accumulating a fully annotated large-scale dataset for supervised VAED can be both non-automated and time-consuming. On the other hand, the unsupervised paradigm involves training models exclusively on samples of normal events. It is based on the common assumption that unseen anomaly videos will exhibit high reconstruction errors [2], [3], [4]. Unfortunately, the performance of unsupervised Variational Autoencoder Decoder (VAED) tends to be substandard. This can be attributed to its limited comprehension of anomalies and its incapacity to encompass various forms of normality variants [5]. Consequently, weakly supervised approaches are widely regarded as the most viable paradigm. They outshine both unsupervised and supervised paradigms due to their competitive performance and cost-effectiveness regarding annotations. These approaches reduce cost by utilizing video-level labels instead of laborious fine-grained annotations [6], [7]. In recent time, WVAED has evolved into a well-established technical path of research for VAED [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. The WVAED issue is primarily perceived as a MIL (multiple instance learning) problem [8]. Generally speaking, WVAED models directly generate scores for anomalies by comparing the spatiotemporal features of normal and abnormal events through the MIL technique. The MIL approach deals with training data organized into sets known as positive and negative bags. In the context of MIL, a video is seen as a bag containing numerous instances, where each instance corresponds to a video snippet. A negative bag encompasses the entirety of normal snippets, whereas a positive bag encompasses both normal and abnormal snippets without any indication of the temporal boundaries of abnormal events. The conventional Multiple Instance Learning (MIL) framework assumes that all negative bags exclusively contain negative snippets and that positive bags contain at least one positive snippet. Supervision is solely provided for complete sets, and the individual labels of the snippets within the bags are not given [17]. The outputs of WVAED are inherently more reliable than those of unsupervised VAED due to its ability to comprehend the fundamental variability between normal and abnormal [18]. However, in the WVAED approach, the frames labelled as abnormal in the positive bag are often influenced by the frames labelled as normal in the negative bag, making it challenging to distinctly identify an abnormality in contrast to normality. Consequently, the detection of anomalous snippets can become problematic. Numerous researchers (e.g., [8], [9], [10], [19], [20]) have endeavoured to address this issue by employing multiple instance learning (MIL) frameworks. Many of the existing methodologies encode the extracted visual content by utilizing a backbone such as C3D [21] and I3D [22], which have been pre-trained on tasks related to action recognition. Nevertheless, Visual Activity Detection (VAD) requires representations that are able to clearly depict the events occurring in a given scene. Consequently, these current backbones are unsuitable for VAD due to the existence of a domain gap. In order to overcome this limitation, Joo et al. decided to draw inspiration from the recent achievements in vision-language research, specifically the works of [23], [24], and [25], which demonstrated the effectiveness of feature representation derived from contrastive language-image pretraining (CLIP). To achieve this, they employed the visual features encoded by the vision transformer (ViT) from CLIP. However, it is worth noting that the performance of WVAED methods based on MIL heavily relies on pre-trained feature extractors.

The drawback of the study is that they processed the video in individual frames or short clips to extract the long-range semantic contextual information. To overcome the problem, shao et al. proposed a Temporal Context Aggregation for Video Retrieval (TCA) framework for video representation learning. This innovative approach integrates long-range temporal context among frame-level features through the utilization of the self-attention mechanism [26], [27]. They used contrastive learning to reduce loss or error rate in the evaluation. To enhance the TCA features, Tean et al. employed Robust TCA features, including multiple instance learning (MIL) loss calculation approach [19]. They reported 84.30% and 97.21% AUC for the UCF-crime and Shanghai Tech datasets, respectively. To improve the AUC rate by increasing the feature effectiveness, PU et al. employed TCA to enhance the long-range dependency and PEL instead of contrastive learning to increase the correct prediction rate by reducing the error [28]. They employed MLP with PEL to reduce the features and casual convolution (CC) for the classification. PEL mainly integrates semantic priors utilizing knowledge-based prompts aiming to increase the recognition rate by boosting the discriminative capacity while ensuring high separability between subclass between the anomaly, and finally, they calculate the score and the error rate with the MIL loss function: the reported AUC rate 86.76%, 85.59% and 98.14% for UCF-crimes, XD-Violence and Shanghai Tech dataset respectively. To improve the recognition rate, Zhao et al. proposed a new temporal feature extraction using graph-based transformers, namely Uncertainty Regulated Dual Memory Units. (UR-DMU) through the I3D backbone pre-trained features [29]. They reported 86.97% and 94.02% for UCF-crime and XD-violence datasets, respectively. To improve the performance, more recently, sharif et al. proposed a two-stream pre-trained feature-based temporal feature enhancement module where they first extracted CNN-based I3D features in the first stream by selective top-k score and ViT-based Clip feature in the second stream [30]. Finally, they fused the two features and employed MLP and classification module for the classification. They reported 88.97% and 98.66% AUC for the UCF crime and the Shanghai tech dataset, respectively. The drawback of this model is that it did not achieve satisfactory performance for the real-time deployment due to a lack of feature effectiveness. Also, they utilized CNN and ViT-based pre-trained model features and temporal feature enhancement, but they did not consider the graph-based feature enhancement and spatial feature enhancement in the module. Also, the UR-DMU [29] utilized the graph-based feature enhancement, but they did not discuss time-varying enhancement and TCA [26], [28] reflected the vice versa problem. In addition, UR-DMU [29], TCA [26], [28] and I3D-CLIP [30] they are having lacking the extracting all possible kind of the feature. This research group inspired us to work here to extract all possible kinds of features to increase the anomaly detection rate. To overcome the challenges, we proposed here a multi-stage graph and general deep learning (DL) feature enhancement-based anomaly detection system. In the study, we proposed including CNN and ViT-based pre-trained features, temporal features, graph-based temporal features and spatial enhancement of the features.

The main contributions of the proposed model are given below:

- **Stage 1: General Deep Learning Model Based Dual-Stream Feature Extraction:**
  The first stage of our methodology is characterized by the innovative use of two streams, each contributing distinct yet complementary features to the anomaly detection process. Leveraging CLIP and I3D, we extract rich semantic information and spatiotemporal features, respectively, setting a solid foundation for subsequent analysis. Building upon the extracted features, we seamlessly integrate them into the Temporal Contextual Aggregation (TCA) mechanism. This module helps to capture comprehensive contextual information by reusing the similarity matrix and implementing adaptive fusion. This integration facilitates the effective capture of temporal dependencies across video frames, enhancing the model's ability to discern anomalous patterns amidst dynamic scenes.
- **Stage 2: Graph-Based UR-DMU Model Integration and Refinement:**
  In the second stage, we introduce the Uncertainty-regulated Dual Memory Units (UR-DMU) model, renowned for its ability to simultaneously learn representations of regular and abnormal data. By incorporating global and local structures through GCN and Global/Local Multi-Head Self Attention (GL-MHSA) modules, our model captures intricate associations within video data. Additionally, refinement through a Multi-Layer Perceptron (MLP) enables non-linear mapping, further enhancing the model's discriminatory capabilities.
- **Stage-3 Feature Reduction Classification and Evaluation with Impact and Promise:**
  In the third stage, we used a feature reduction module using two-layer multilayer perception (MLP) integrated with PEL to refine and learn discriminative features through knowledge-based prompts. This integration of non-linear mapping further enhances the model's ability to differentiate between normal and anomalous behaviour.
- **Classification and Evaluation** Finally, we employed a classifier module to predict the snippet-level anomaly scores. In the training phase, the based function transfers the snippet-level scores into bag-level predictions for learning high activation in anomalous cases. We evaluate the proposed model with three benchmark datasets, namely UCF-Crime Dataset, ShanghaiTech Dataset, and XD-Violence. The extensive performance result proves the superiority of the proposed model. Through the integration of these cutting-edge technologies and methodologies, our approach offers a comprehensive solution to video-based anomaly detection. We believe that our model holds significant promise for real-world applications, demonstrating superior performance and efficacy in anomaly detection tasks.

II. LITERATURE REVIEW

The methodologies utilized in WVAED rely on labels at the video level, which consistently adhere to the MIL ranking framework [8]. According to the MIL approach, a regression model is trained using the WVAED method with the assumption that the maximum score of the positive bag is greater than that of the negative bag in order to assign scores for video snippets. These [8], [9], [19], and [6], [11] all incorporated pre-trained models based on convolutional neural networks into their experimental procedure setups. In addition, Sultani et al. [8] meticulously curated pre-annotated normal and abnormal video events at the video level to construct the widely recognized UCF-Crime dataset. The dataset was employed for anomaly detection by utilizing a weakly supervised framework. Within the confines of this framework, C3D features [31] were extracted for video segments, and then a ranking loss function was used to train a fully connected neural network (FCNN). The purpose of this function was to compute the loss between the most highly scored rank examples in the positive bag and the negative bag. Tian et al. [19] presented a model and utilized the C3D [31] and I3D [22] models for the aiming of feature extractors in their WVAED model. They contended that by selecting the top 3 features based on their magnitude, a more pronounced differentiation can be achieved between normal and abnormal videos (AVs). Specifically, in cases where multiple abnormal snippets exist within an anomalous video, the average snippet feature magnitude of the anomalous video surpasses that of normal videos (NVs). Hang et al. [9] presented a model to extract positive and negative video-segmented C3D features by using a temporal convolution network [31]. Specifically, they trained the network between the previous adjacent segment and the current segment. Further, they used inner and outer bag ranking losses to train the model based on two branches of an FCNN. This loss accounted for the greatest and lowest-scoring parts in terms of the positive bags and negative bags.

Similarly, Zhong et al. [6] and Zhu et al. [11] implemented models that trained a feature-based encoder and classifier simultaneously. Zhong et al. [6] analyzed WVAED and performed it as a supervised learning problem using noise labels. Extensive experiments were undertaken to evaluate the universal applicability of their model, using both temporal segment network [32] and C3D [31]. Zhu and Newsam [11] integrated an attention block to their MIL ranking model to account for temporal context. They claimed that motion information features extracted by C3D [31] and I3D [22] outperformed features obtained from individual images using pre-train model VGG16 [33] and Inception [34]. ViT-based pre-trained models can be classified into two types: single-stream and dual-stream. In the single-stream approach, text and picture (or video) representations are modeled using a single transformer in a single framework, while the dual-stream model uses a decoupled encoder to encode text and image (or video) separately. Among the most notable ViT feature extractors are CLIP [35], ViLBERT [36], Visual-BERT [37], and data-efficient CLIP [38]. For the WVAED problem, Joo et al. [20] recently presented a temporal self-attention framework for CLIP-assisted [35]. They implemented the experiments on open accessible datasets to perform their end-to-end WVAED model. Li et al. [39] presented a multi-instance learning network based on transformers to get anomaly scores for both videos and snippets. They used the video-level anomaly probability in the inference stage to lessen the snippet-level anomaly score's volatility. Lv et al. [40] introduced an unbiased MIL scheme that trained an unbiased anomaly classifier and a tailored representation for WVAED. In view of the available solutions, it has been observed that, in general, CNN and ViT are typically utilized in isolation. To leverage the benefits offered by both CNN- and ViT-based pre-trained models, an architecture known as CNN-ViT-TSAN, which is supported by Multiple Instance Learning (MIL), has been devised. This architecture aims to establish a range of models for addressing the problem of Weakly Supervised Variational Autoencoder Design (WVAED). The drawback of the study is that they processed the video in individual frames or short clips to extract the long-range semantic contextual information. To overcome the problem, shao et al. proposed a TCA framework for video representation learning. This innovative approach integrates long-range temporal context among frame-level features through the utilization of the self-attention mechanism [26], [27]. They used contrastive learning to reduce loss or error rate in the evaluation. To enhance the TCA features, Tean et al. employed Robust TCA features, including multiple instance learning (MIL) loss calculation approach [19]. They reported 84.30% and 97.21% AUC for the UCF-crime and Shanghai Tech datasets, respectively. To improve the AUC rate by increasing the feature effectiveness, PU et al. employed TCA to enhance the long-range dependency and PEL instead of contrastive learning to increase the correct prediction rate by reducing the error [28]. They employed MLP with PEL to reduce the features and casual convolution (CC) for the classification. PEL mainly integrates semantic priors utilizing knowledge-based prompts aiming to increase the recognition rate by boosting the discriminative capacity while ensuring high separability between subclass between the anomaly, and finally, they calculate the score and the error rate with the MIL loss function: the reported AUC rate 86.76%, 85.59%, and 98.14% for UCF-crimes, XD-Violence and Shanghai Tech dataset respectively. To improve the recognition rate, Zhao et al. proposed a new temporal feature extraction using graph-based transformers, namely Uncertainty Regulated Dual Memory Units. (UR-DMU) through the I3D backbone pre-trained features [29]. They reported 86.97% and 94.02% for UCF-crime and XD-violence datasets, respectively. To improve the performance, more recently, sharif et al. proposed a two-stream pre-trained feature-based temporal feature enhancement module where they first extracted CNN-based I3D features in the first stream by selective top-k score and ViT-based Clip feature in the second stream [30]. Finally, the fused the two features and employed MLP and classification module for the classification. They reported 88.97% and 98.66% AUC for the UCF crime and the Shanghai tech dataset, respectively. The drawback of this model is that it did not achieve satisfactory performance for the real-time deployment due to a lack of feature effectiveness. Also, they utilized CNN and ViT-based pre-trained model features and temporal feature enhancement, but they did not consider the graph-based feature enhancement and spatial feature enhancement in the module. Also, the UR-DMU [29] utilized the graph-based feature enhancement, but they did not discuss time-varying enhancement, and TCA [26], [28] reflected the vice versa problem. In addition, UR-DMU [29], TCA [26], [28] and I3D-CLIP [30] they are having lacking the extracting all possible kind of the feature. This research group inspired us to work here to extract all possible kinds of features to increase the anomaly detection rate. To overcome the challenges, we proposed here a multi-stage graph and general DL feature enhancement-based anomaly detection system. In the study, we proposed including CNN and ViT-based pre-trained features, temporal features, graph-based temporal features, and spatial enhancement of the features.

III. DATASET

Anomaly detection datasets play a crucial role in developing and evaluating algorithms aimed at identifying irregular or unexpected events within data streams. These datasets provide diverse scenarios, allowing researchers to train and test their models under various conditions. Also, there are many datasets available for anomaly detection, and we used the following most usable benchmark datasets, such as ShanghaiTech [41], the UCF-Crime [8], and XD-Violence [29], which offer different scales, backgrounds, and types of anomalies, catering to different research needs. By utilizing these datasets, researchers can benchmark their anomaly detection methods, assess their performance, and contribute to advancing the field of anomaly detection in real-world applications.

A. ShanghaiTech DATASET

This dataset represents a medium-scale anomaly dataset comprising 317,398 frames of video clips. These clips capture scenes from various locations within the ShanghaiTech Campus. The dataset includes 13 distinct background scenes, consisting of 307 NVs and 130 anomaly videos. The earliest dataset [41], serves as a common benchmark for VAED. In this dataset, the training set contains NVs, while the testing set contains both normal and anomalous videos. In order to create a weakly supervised training set that encompasses all 13 background scenes, Zhong et al. [6] reorganized the dataset. Their approach involved selecting a subset of anomalous testing videos and using them as training data. We followed the procedure outlined by Zhong et al. [6] to transform the ShanghaiTech dataset into this weakly supervised setting.

B. UCF-CRIME DATASET

This dataset consists of a large-scale anomaly detection dataset that includes 1900 untrimmed videos collected from real-world street and indoor surveillance cameras. There are 128 hours of video in total. The dataset covers 13 different real-world anomalies such as abuse, arrest, arson, assault, accident, burglary, explosion, "fighting", "robbery", "shooting", "stealing", "shoplifting", and "vandalism". Unlike the static background in the ShanghaiTech dataset [41], the UCF-Crime [8] dataset has more complicated and diverse backgrounds. The training set of the UCF-Crime dataset contains 1610 videos, with 800 labeled as normal and 810 labeled as anomalous. The testing set contains 290 videos, with 150 labeled as normal and 140 labeled as anomalous, and includes frame-level labels.

C. XD-VIOLENCE

XD-Violence dataset comprises a variety of media formats, specifically videos and audio. The dataset encompasses a diverse range of backgrounds, such as movies, games, and live scenes. It consists of a total of 4754 videos, with 3954 videos designated for training purposes and equipped with video-level labels. Additionally, 800 testing videos have been labelled at the frame-level [29].

IV. PROPOSED METHOD

In our proposed model for video-based anomaly detection, we leverage a sophisticated combination of state-of-the-art technologies to enhance the accuracy and robustness of anomaly identification. Figure 1 demonstrated the proposed model where we used a multi-stage DL approach. This study is mainly designed to extract characteristics that are more indicative of anomalies. In our approach, similar to previous work [28], [29], [42], we extract features from each video with 10-crop augmentation for the UCF-Crime and Shanghai-Tech datasets and 5-crop augmentation for the XD-Violence dataset using pre-trained models. We divided the untrimmed video into non-overlapping snippets by utilizing a 16-frame sliding window. Then, we introduce a multi-backbone framework, combining a CLIP model trained on Kinetics with an I3D model also pre-trained on Kinetics. This dual-backbone approach leverages the strengths of both architectures to enhance the feature extraction process for video anomaly detection. Subsequently, this enhanced feature set is streamlined via TCA, CNN, UR-DMU and a two-layer Multilayer Perceptron (MLP), optimizing it for further analysis or applications. In the procedure, we employed three stages. Where the first stage was constructed with two streams and the initial stream, we utilised CLIP (Contrastive Language-Image Pre-training) and selected the top-k features, which are considered as the feature of the first stream. In the second stream, we used a pre-trained I3D (Inflated 3D ConvNet) network to extract rich semantic information that fed into the Temporal Contextual Aggregation (TCA) mechanism for integrating contextual information across frames, effectively capturing temporal dependencies. Then, we employed the CNN module for spatial enhancement of TCA output to extract spatiotemporal features from video frames as a feature of the second stream. We concatenated the CLIP-based first stream and the I3D-based second stream feature that fed into the UR-DMU [29] model, which employs dual memory units to learn representations of regular data and discriminative features of abnormal data simultaneously. This model incorporates both global and local structures through GCN [14], [15], [16] and Global/Local Multi-Head Self Attention (GL-MHSA) modules, facilitating the capture of associations in videos. In the third stage, we used a feature reduction module using two-layer multilayer perception (MLP) integrated with PEL to refine and learn discriminative features through knowledge-based prompts. This integration of non-linear mapping further enhances the model's ability to differentiate between normal and anomalous behaviour. Finally, we employed a classifier module to predict the snippet-level anomaly scores. In the training phase, the based function transfers the snippet-level scores into bag-level predictions for learning high activation in anomalous cases. By integrating these cutting-edge technologies, our model offers a comprehensive approach to video-based anomaly detection, promising superior performance in real-world applications.

A. PEPROCESSING

In WVAED, the training set solely comprises video-level labels. Considered a set of training video can expressed as below $W = V_v, y_{v_{v=1}}^{w}$ where each video $V_v = Frame_{i_{i=1}}^{N_v} \in \mathbb{R}^{N_v \times W \times H}$ represent the sequence of frames N_v and each frame consists with $width = W$ and $height = H$. In addition, the label of each video V_v contained in $y_v = 0, 1$ is associated with the anomaly label. Among the frames of a specific video, we divided it into a set of snippets, which can be expressed

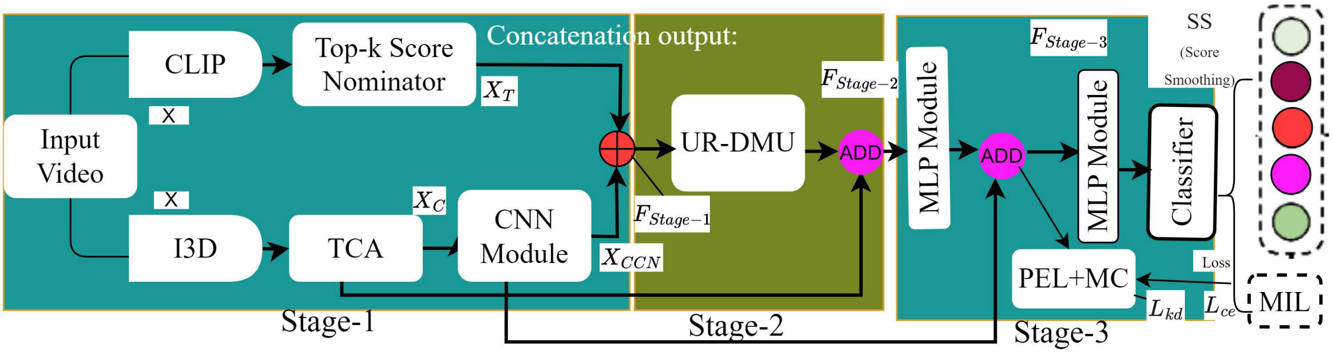


FIGURE 1. Proposed Model.

as $\gamma_{i}{}_{i=1}^{\lfloor\frac{N_v}{\Delta}\rfloor}$ where each snippet contained same number of frames Δ.

In the preprocessing, we followed the existing system, and that is first, we divided the untrimmed video into non-overlapping snippets by utilizing a 16-frame sliding window [28], [29], [42]. Then, we extracted features from each sample, using 10-crop augmentation for the UCF-Crime and Shanghai-Tech datasets and 5-crop augmentation for the XD-Violence dataset using pre-trained models in stages 1 [28], [29], [42].

B. STAGE 1: PRETRAINED 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 also pre-trained on Kinetics. It is important to note that our I3D model is configured to process only RGB input. In this architecture, the I3D RGB model extracts features in a 1024-dimensional space, 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.

1) ViT TRANSFORMER BASED CLIP FEATURE EXTRACTION STREAM

In the first stream, we employed a based CLIP model to extract the pre-trained features and then nominate the top score to select the most relevant video snippets.

a: PRETRAINED CLIP FEATURE FEATURE

CLIP leverages a unified framework for understanding both text and image data, enabling it to capture rich semantic information from video frames. Vision-language pre-trained models leverage ViTs to capture the correlations between objects or actions depicted in a video and those described in textual content. These sophisticated models excel at extracting intricate relationships between visual and linguistic elements, thereby facilitating comprehensive understanding and analysis across modalities. There are many researchers used the concept of ViT as a backbone for different kinds of transformers, namely ViLBERT [36], CLIP [35], VisualBERT [37] and data-efficient CLIP [38], aiming to develop different kinds of the language model and the multi-modal vision. Generally, CLIP [38] serves as a multi-modal vision and language model, harnessing a ViT as its foundational framework for extracting visual features. We considered that $d_j = \lceil\frac{\Delta}{2}\rceil$ is a middle frame for the video and from the snipped γ_j, which means we did not consider all frame at a time. In our study, we employ the CLIP on d_j of the snippet γ_j to represent its features as $\phi_{v_{\mathrm{j}}} \in \mathbb{R}^{\aleph}$, here $\aleph$ represented the feature dimension, and final feature vector can be constructed with $\phi_{v_{\mathrm{vit}}} = \phi_{j}{}_{j=1}^{T_v} \in \mathbb{R}^{T\times\aleph}$ [30]. We used pre-trained CLIP models in the first stream to extract effective features from each video. CLIP consists of a multi-backbone framework. The CLIP model provides feature vectors in 512 dimensions.

b: TOP-K SCORE NOMINATOR

The output of the CLIP model is fed into the K-Score Selection Module, which is demonstrated in Figure 2. The top-k score nominator, as described [42], is a crucial component for selecting the most relevant video snippets. These scores are then processed to identify the top most relevant snippets. This method ensures that the snippets with the highest relevance, indicated by their score values, are selected for further processing. This module involved cloning the input vector, which comes from the CLIP model output and is known as a score vector. Then, we added the Gaussian noise and calculated the magnitude. Based on the magnitude value, we selected the top K scores. This top-k score nominator is integral for focusing the model's attention on the most significant parts of the video.

2) CNN BASED I3D FEATURE EXTRACTION STREAM

In the second branch of the first stage module, we employed I3D and then enhanced the information using TCA and CNN modules. The utilization of the I3D module allows for the extraction of robust spatio-temporal features from video sequences. By incorporating spatial and temporal information, this module effectively captures motion and appearance cues, enabling a comprehensive representation of video

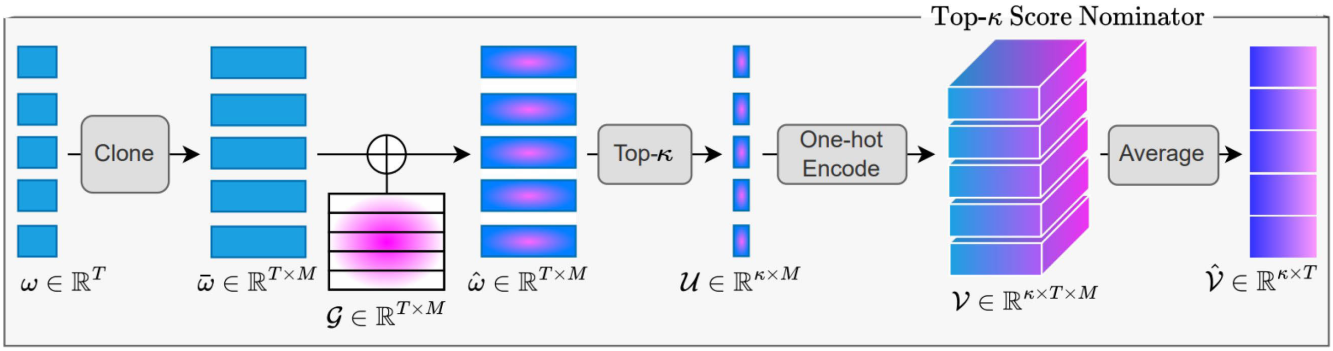

Top-κ Score Nominator
Clone
Top-κ
One-hot Encode
Average
$\omega \in \mathbb{R}^{T}$
$\bar{\omega} \in \mathbb{R}^{T\times M}$
$\mathcal{G} \in \mathbb{R}^{T\times M}$
$\hat{\omega} \in \mathbb{R}^{T\times M}$
$\mathcal{U} \in \mathbb{R}^{\kappa\times M}$
$\mathcal{V} \in \mathbb{R}^{\kappa\times T\times M}$
$\hat{\mathcal{V}} \in \mathbb{R}^{\kappa\times T}$

FIGURE 2. Internal structure of the Top K-Score Selection module [42].

content. TCA plays a pivotal role in integrating contextual information across multiple frames. By considering temporal dependencies within video sequences, TCA enhances the model's ability to discern anomalies. This mechanism ensures that the model can effectively capture dynamic changes over time, improving anomaly detection accuracy. The incorporation of 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.

a: I3D FEATURES

I3D excels in capturing spatio-temporal features from video sequences, providing a robust representation of motion and appearance cues within the temporal context [22]. One of the most widely used DL models, (CNN), has a lot of potential for image classification. CNN-based C3D (Convolutional 3D) [21] are the most usable common feature extractors. Tran et al. [36] showed that C3D can model appearance and motion information simultaneously and outperform the 2D CNN features in various video-analysis tasks. Technically, we calculated the I3D features from the T snippets in the dimension $\acute{\aleph}$. Assume $\phi^{'}_{v_{\text{cnn}}} = \phi_{i}{}_{i=1}^{T_v} \in \mathbb{R}^{T_v \times \acute{\aleph}}$ [30] used to extract features where for a specific video V_v contained the T_v number of snippet and feature vector size can be expressed as V_v. In lieu of employing PCA, we opt for the low-variance filter algorithm to reduce the dimensionality of the extracted data. After reducing the dimension, it produces the first stage feature of this stage and that dimension can be expressed with $\phi^{'}_{v_{\text{cnn}}} \in \mathbb{R}^{T_v \times \acute{\aleph}}$ which comes from the $\phi_{v_{\text{cnn}}} \in \mathbb{R}^{T_v \times \aleph}$ where $\aleph$ is the feature dimension extracted from T snippets.

b: TEMPORAL CONTEXT AGGREGATION MODULE (TCA)

To enhance the temporal contextual information of the I3D features, we used the TCA model [28]. TCA facilitates the integration of contextual information across multiple frames, enhancing the model's ability to discern anomalies by considering temporal dependencies effectively. This mainly used as a video representation learning framework that incorporates long-range temporal information between frame-level features using the self-attention mechanism [26], [28]. It mainly captures temporal relationships from both local and global perspectives. Figure 3 demonstrated the TCA calculation procedure where X is the output of the I3D module, which is projected here in the latent space utilizing various linear layers and finally produces the similarity matrix as below:

$$M = f_q(X) \cdot f_k(X)^{\top} \tag{1}$$

$$A^g = \text{softmax}\left(\frac{M}{\sqrt{D_h}}\right) \tag{2}$$

$$X^g = A^g \cdot f_v(X) \tag{3}$$

Here, query, key and value are represented by $f_q(.), f_k(.)$ and $f_v(.)$, $\top$ is denoted by the transpose operation, and the dimension of hidden spaces represented by D_h. In addition, A^g denotes the global attention, and X^g represents the global context features [26], [28]. We enhanced the similarity matrix with the dynamic position encoding (DPE) approach according to the following Equation 4:

$$\mathbf{G} = exp(-|\gamma(i-j)^2 + \beta|) \tag{4}$$

where i and j denote the absolute positions of two snippets, and γ and β are learnable weights and bias terms.

In contrast, we also calculated the local attention and local context features according to the below formulas:

$$A^l = \text{softmax}\left(\frac{\tilde{M}}{\sqrt{D_h}}\right) \tag{5}$$

$$X^l = A^l \cdot f_v(X) \tag{6}$$

Here A^l denotes the local attention, and X^l represents the global context features where $\tilde{M}$ represent masking output of the similarity matrix from Equation (1) [26], [28]

Then we concatenated the global attention head X^g and local attention head X^l to produce the final feature X^o using Equation (7).

$$X^o = \alpha \cdot X^g + (1-\alpha) \cdot X^l \tag{7}$$

After normalizing the features, we concatenated with the skip connection to overcome lost information. Finally, we employed a linear layer and produced the output of the TCA module feature according to Equation (8).

$$X^c = \mathrm{LN}(X + f_h(\mathrm{Norm}(X^o))) \tag{8}$$

where global weight, local weight and combination of power normalization are represented by α, $(1 - \alpha)$, and $\mathrm{Norm}(\cdot)$ respectively.

CNN Module: These features are then processed through a 1D CNN (Conv1d), followed by a ReLU activation function and a dropout rate of 0.1, reducing the feature dimensionality to 512.

C. FEATURE FUSION

In the first stream, we used top k Score Nominator [42] to select the top k segments based on their CLIP feature relevance to obtain a refined set of 512-dimensional features denoted with X_T. We got the final feature from the FC module in the second stream denoted with X_{CCN}. These features are then concatenated, resulting in comprehensive 1024-dimensional features denoted with $F_{stage-1}$ using the Equation.

$$F_{stage-1} = X_T \oplus X_{CCN} \tag{9}$$

D. STAGE 2: UR-DMU BASED FEATURE

To produce the graph-based temporal enhancement feature, we employed the UR-DMU [14], [15], [29] approach, which is mainly incorporated with dual memory units to simultaneously learn representations of regular data and discriminative features of abnormal data. The main goal is to improve the model's ability to differentiate between normal and anomalous behaviour. It consists of three main components demonstrated in Figure 4. Global and Local Multi-Head Self Attention (GL-MHSA) is crucial for learning both long and short-temporal dependencies of anomalous features. It enhances the transformer structure by integrating global and local structural concepts from graph convolution networks.

$$\mathbf{S} = \sigma(\frac{\mathbf{X}\mathbf{M}^t}{\sqrt{D}}), \mathbf{M}_{aug} = S\mathbf{M} \tag{10}$$

where $\mathbf{X}$ is a feature obtained from GL-MHSA. $\mathbf{M}$ is the memory bank number, D is the number of dimensions of output, $\mathbf{M}$ is querying memory banks, σ is sigmoid activation, and $S \in \mathbb{R}^{NM}$ is the query score. Following that, M_{aug} is used to represent the memory augmentation feature produced by a read operation. We define a dual memory loss as consisting of four binary cross-entropy (BCE) losses in order to train the dual memory units.

$$\begin{aligned} L_{dm} = BCE(\mathbf{S}^n_{k;n}, \mathbf{y}^n_n) + BCE(\mathbf{S}^n_{k;a}, \mathbf{y}^n_a) \\ + BCE(S^a_{k;n;k}, y^a_n) + BCE(S^a_{k;a;k}, y^a_a) \end{aligned} \tag{11}$$

where $\mathbf{S}^n_{k;n}$ is a normal memory score, $\mathbf{y}^n_n = \mathbf{1} \in \mathbb{R}^N$, $\mathbf{S}^n_{k;a}$ is a anomaly memory score, $\mathbf{y}^n_a = \mathbf{0} \in \mathbb{R}^N$. And the means of $\mathbf{S}^a_{k;n}$, $\mathbf{S}^a_{k;a}$ top-K result along the first dimension are $S^a_{k;n;k}, S^a_{k;a;k} \in \mathbb{R}^N$. y^a_n, y^a_a are labels and the value is 1. This helps distinguish hard samples better by comparing feature similarities with stored templates. Normal Data Uncertainty Learning (NUL) uses a Gaussian distribution to constrain the latent normal representation. It's an approach not commonly used in weakly supervised video anomaly detection, drawing on concepts from unsupervised anomaly detection methods. For training and testing, pairs of videos with equal amounts of normal and abnormal footage are processed. The model generates a score for each video snippet, using Binary Cross-Entropy (BCE) loss and five auxiliary losses for discrimination between normality and anomaly. During testing, the model utilizes only the mean-encoder network of the DUL module to obtain feature embeddings, which are then used to label the video snippets and finally produce the UR-DMU features, which is denoted F_{urdmu}. Then we produce the final feature of stage 2 $F_{Stage-2}$ by adding the feature of the UR-DMU F_{urdmu} with the TCA X_c using the following Equation 12.

$$X_{Stage-2} = F_{urdmu} + X_c \tag{12}$$

E. STAGE 3: FEATURE REDUCTION WITH PEL THROUGH MLP MODULE

To select the effective features from the graph-based UR-DMU $F_{stage-2}$ features, we employed MLP with the PEL module, which is described below.

a: MLP

To achieve high-level semantic representations by selecting the effective feature from the graph-based $F_{stage-2}$ features, we employed a two-layer MLP for feature reduction. MLP serves as a powerful tool for non-linear mapping and feature transformation, enabling the model to learn complex decision boundaries and refine the extracted features for final anomaly detection. This MLP incorporates two Conv1d layers, two GELU activations, and two PDropout mechanisms [43]. Prior to the first Conv1d layer, we integrate features from TCA. Following the first Conv1d layer, we append a 512-dimensional feature derived from I3D. Each Conv1D layer is succeeded by a GELU activation function and a dropout operation. This methodology is symbolized as following Equation 13:

$$\begin{aligned} F_{MLP-1} &= Dropout(GELU(Conv1D(F_{stage-2}))) \\ F_{Stage-3} &= Dropout(GELU(Conv1D(F_{MLP-1}))) \end{aligned} \tag{13}$$

Finally, we utilize a causal convolution layer to produce the anomaly scores, integrating both present and past observations for enhanced reliability. The classifier is represented as:

$$S = \sigma\left(f_t(X_s)\right), \tag{14}$$

where $f_t(\cdot)$ denotes the causal convolution layer with a kernel size of Δt, $\sigma(\cdot)$ represents the sigmoid function, and s_i signifies the anomaly score of the i-th snippet. Finally, we employed the multi-layer instance learning (MIL) as a loss

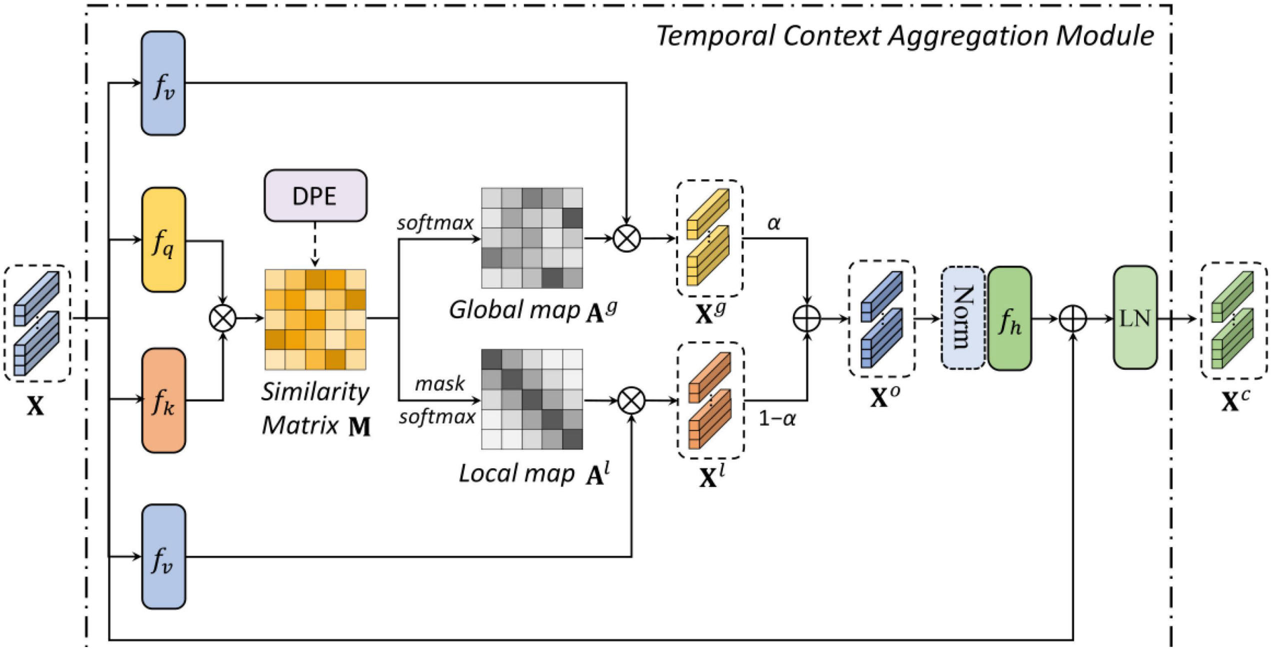


FIGURE 3. Working structure of the TCA module [28].

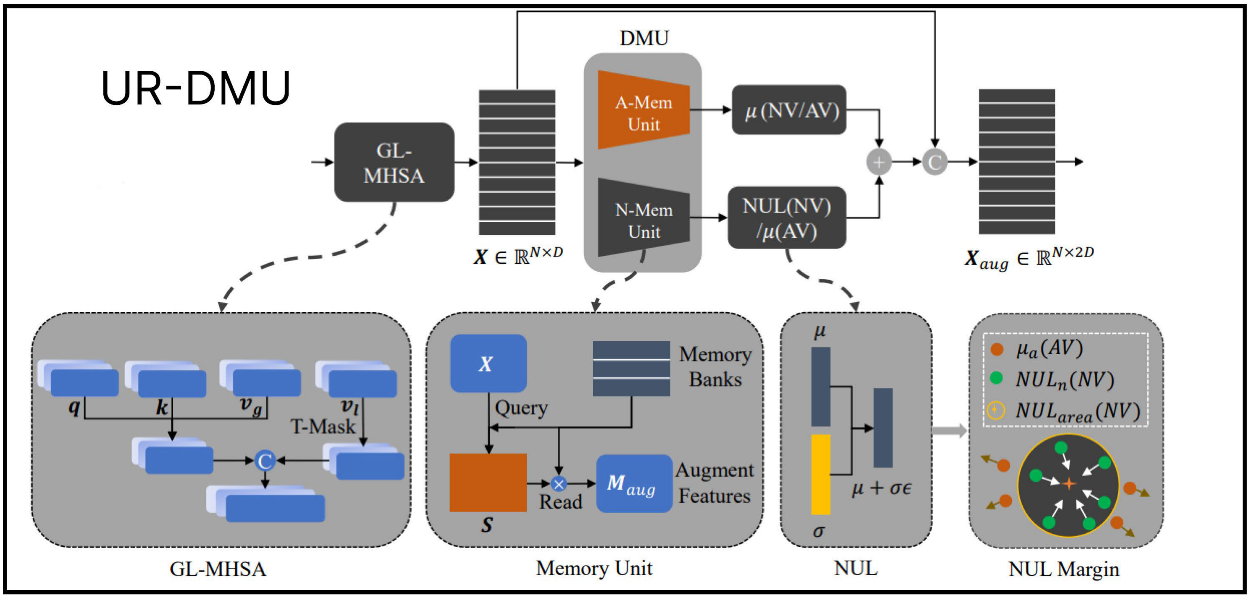


FIGURE 4. Working diagram of the UR-DMU module [29].

function [28], [44]. Specifically, we determine the video-level prediction p_i by computing the mean value of the top-k anomaly scores. For positive bags, we set $k = \lfloor \frac{T}{16} + 1 \rfloor$, and for negative bags, we set $k = 1$. Given a mini-batch containing B samples with video-level ground truth y_i, the binary cross-entropy is formulated as:

$$L_{\text{ce}} = -\frac{1}{B}\sum_{n=1}^{B} y_i \log(p_i). \qquad (15)$$

b: PROMPT-ENHANCED LEARNING (PEL)

In this study, we employ Prompt-Enhanced Learning (PEL) proposed by Joo et al. [28], [42] to enrich visual representations by integrating knowledge-based contextual information, improving anomaly detection in complex scenarios. It involves three key steps: prompt construction, fore-background separation, and cross-modal alignment. The prompt construction mainly selected the common relation among the categories to form prompts that focus on the high

TABLE 1. Ablation study performance AUC (%).

I3D	TCA	CLIP	Top-K	MHSA	DMU	PEL	MC	SS	UCF(%)	XD(%)	SH(%)
✓									81.37	74.82	93.15
✓	✓								80.88	75.63	92.66
✓	✓	✓							83.54	75.63	94.04
✓	✓	✓	✓						83.94	76.01	92.91
✓	✓	✓	✓	✓					86.60	79.71	97.76
✓	✓	✓	✓	✓	✓				87.47	79.74	97.89
✓	✓	✓	✓	✓	✓	✓			89.68	83.68	98.37
✓	✓	✓	✓	✓	✓	✓	✓		89.68	86.37	98.54
✓	✓	✓	✓	✓	✓	✓	✓	✓	90.09	86.48	98.69

occurrence categories and make a relevant semantic relationship dictionary. Then, based on the output of $F_MLP - 1$ and cross-entropy loss L_{ce} information, it produces the video label background and foreground features. Finally, the effective feature is prompted based on the enhanced fine-grained semantics of visual features. That means PEL assess the likelihood of a visual feature matching a particular prompt across several anomaly classes and 1 normal class. Overall, the PEL module's integration of textual and visual modalities enables a more nuanced and context-aware approach to anomaly detection in video data. Finally, the cross-modal alignment loss is computed using the Kullback-Leibler divergence, compelling the network to discern between the visual content of the video representing abnormal behavior (foreground) and irrelevant content (background). The loss function is formulated as follows:

$$L_{\text{kd}} = \mathbb{E}_{p\sim p(v)}[\log p_{v2t}(v) - \log q_{v2t}(v)], \tag{16}$$

where $p_{v2t}(v)$ and $q_{v2t}(v)$ denote the similarity score and semantic consistency label of the video-prompt pair, respectively. For a positive pair, $q = 1$; otherwise, $q = 0$. We added the Magnitude Contrastive (MC) Loss [45] with the L_{kd} to enhance the effectiveness of the loss calculation procedure.

V. EVOLUTION AND PERFORMANCE

To evaluate the proposed model, we used three benchmark anomaly detection datasets: ShanghaiTech [41], the UCF-Crime [8], and XD-Violence [29].

A. TRAINING AND TESTING PROCEDURES

During training, we optimize the objective function $L = L_{\text{ce}} + \lambda L_{\text{kd}}$, where λ adjusts the alignment loss. This enables our model to generate discriminative representations of positive and negative snippets, improving generalizability. In the testing phase, we mitigate transient noise impact with a score smoothing (SS) strategy using distinct pooling operations by following Equation (17).

$$\tilde{s}_i = \frac{1}{\kappa}\sum_{j=\kappa}^{i+\kappa-1} s_j \tag{17}$$

It also helps us to suppress biases and reduce false alarms by smoothing prediction scores. Also, we skipped feature-length normalization of the extracted video feature vectors, assuming independence among videos. These vectors underwent TSAN processing, producing reweighed attention features. These features were then fed into the snippet association network and an MLP-based converter to obtain anomaly scores. Each score, ranging from 0 to 1, indicates the anomaly probability of the corresponding snippet. To maintain the original video order for evaluation against ground truth labels, each score was replicated Δ times to match the video's usual frame length.

TABLE 2. Performance result.

Dataset Name	AUC (%)	Anomaly AUC (%)	AP (%)	FAR (%)
UCF-Crime	0.9009	0.7456	0.4090	0.0204
XD-Violence	0.9509	0.8626	0.8648	0.0013
Shanghai	0.9869	0.8228	0.7780	0.0000

1) ENVIRONMENTAL SETUP AND EVALUATION METRICES

The system was developed on a machine with a GeForce RTX 4090 24GB series GPU, running CUDA version 11.7 and NVIDIA driver version 515. The system utilized 64GB of RAM. During the training processing, the learning set employed a learning rate of 0.005 and a batch size of 32. The training process lasted for 300 epochs using the Adam optimizer on the same RTX 4090 machine. For efficient implementation of graph convolution and attention with low computational cost, the Python environment included the following packages open cv pickle package, panda package, a [46], [47], [48], these all packages facilitated initial data processing and model development [47], [48].

We compare the results with the area under the curve (AUC) of the frame-level receiver operating characteristics (ROC) for UCF-Crime and ShanghaiTech to the WS-VAD performance. For XD-Violence, on the other hand, the AUC of the frame-level precision-recall curve (AP) is utilized. In ablation experiments, the False Alarm Rate (FAR) and the anomaly subset consisting of only abnormal data are also utilized. The FAR (false alarm rate) we displayed today was different from the normal implementation. I used the "Learning Prompt-Enhanced Context features for Weakly-Supervised Video Anomaly Detection" implementation as is, but this FAR is limited to normal video. In other words, it is the FAR for video where all frames are 0. Also, in the ShangihaiTech dataset, this FAR was exactly 0. This is probably due to the high AUC of 98.6%. And FAR was not used much as an

TABLE 3. State-of-the-art comparison of the proposed model for the UCF crime and ShanghaiTech dataset.

Model Name	Year	Feature Extractor Name	Shanghai Tech Dataset		UCF Crime Dataset	
			AUC(%)	1-AUC(%)	AUC(%)	1-AUC(%)
Sultani et al. [8]	2018	C3D	0.8317	0.1683	0.7541	0.2459
Sultani et al. [8]	2018	ID3	0.8533	0.1467	0.7792	0.2208
Zhong et al. [6]	2019	C3D	0.7644	0.2356	0.8108	0.1892
Zhong et al. [6]	2019	TSN	0.8444	0.1556	0.8212	0.1788
Zhong et al. [9]	2019	ID3	0.8250	0.1750	0.7870	0.2130
Zaheer et al. [49]	2020	C3D-self	0.8416	0.1584	0.7954	0.2046
Zaheer et al. [7]	2020	C3D	0.8967	0.1033	0.8303	0.1697
Wan et al. [50]	2020	I3D	0.8538	0.1462	0.7896	0.2104
Purwanto et al. [13]	2021	TRN	0.9685	0.0315	0.8500	0.1500
Tian et al. [19]	2021	C3D	0.9151	0.0849	0.8328	0.1672
Majhietal et al. [51]	2021	ID3	0.8822	0.1178	0.8267	0.1733
Tianetal et al. [19]	2021	ID3	0.9721	0.0279	0.8430	0.1570
Wuetal et al. [44]	2021	ID3	0.9748	0.0252	0.8489	0.1511
Yuetal et al. [52]	2021	ID3	0.8783	0.1217	0.8215	0.1785
Lvetal et al. [12]	2021	ID3	0.8530	0.1470	0.8538	0.1462
Fengetal et al. [53]	2021	CD3	0.9313	0.0687	0.8140	0.1860
Zaheer et al. [3]	2022	ResNext	0.8621	0.1379	0.7984	0.7984
Zaheer et al. [54]	2022	CD3	0.9012	0.0988	0.8337	0.1663
Zaheer et al. [54]	2022	3DResNext	0.9146	0.0854	0.8416	0.1584
Joo et al. [20]	2022	C3D	0.9719	0.0281	0.8394	0.1606
Joo et al. et al. [20]	2022	I3D	0.9798	0.0202	0.8466	0.1534
Joo et al. et al. [20]	2022	CLIP	0.9832	0.0168	0.8758	0.1242
Cao et al. [55]	2022	I3D	0.9645	0.0355	0.8587	0.1413
Li et al. [39]	2022	I3D	0.9608	0.0392	0.8530	0.1470
Cao et al. [55]	2022	I3D-graph	0.9605	0.0395	0.8467	0.1533
Tan et al. [56]	2022	I3D	0.9754	0.0246	0.8671	0.1329
Li et al. [39]	2022	VideoSwim	0.9732	0.0268	0.8562	0.1438
Yi et al. [57]	2022	I3D	0.9765	0.0235	0.8429	0.1571
Yu et al. [27]	2022	C3D	0.8835	0.1165	0.8208	0.1792
Yu et al. [27]	2022	I3D	0.8991	0.1009	0.8375	0.1625
Gong et al. [58]	2022	I3D	0.9010	0.0990	0.8100	0.1900
Majhi et al. [59]	2023	13D-Res	0.9622	0.0378	0.8530	0.1470
Park et al. [60]	2023	C3D	0.9602	0.0398	0.8343	0.1657
Park et al. [60]	2023	I3D	0.9743	0.0257	0.8563	0.1437
Pu et al. [28]	2023	I3D	0.9814	0.0186	0.8676	0.1324
Lv et al. [40]	2023	X-CLIP	0.9678	0.0322	0.8675	0.1325
Sun et al. [61]	2023	C3D	0.9656	0.0344	0.8347	0.1653
Sun et al. [61]	2023	I3D	0.9792	0.0208	0.8588	0.1412
Wang et al. [62]	2023	C3D	0.9401	0.0599	0.8148	0.1852
Sharif et al. [30]	2023	I3D+CLIP	0.9866	0.0134	0.8897	0.1103
Proposed Model	-	hybrid model	0.9869	0.0131	0.9000	0.1070

indicator. It was used in two papers, but one paper was not prepared for comparison with the other.

B. ABLATION STUDY

Table 1 demonstrated the ablation study of the proposed model, which also shows the contribution of the multi-backbone pre-trained model. In the ablation study, we systematically evaluated the impact of various technologies on weakly supervised video anomaly detection. The presence of a check mark indicates the utilization of the corresponding technology in our experiments. We observed that integrating I3D alone resulted in a notable improvement in performance across all datasets. Incorporating TCA alongside I3D further enhanced detection accuracy. CLIP integration facilitated even better results, particularly on the UCF crimes dataset. Additionally, employing Top-K selection improved performance consistently. The introduction of UR-DMU significantly boosted detection rates, which is particularly evident in the XD violence and SH tech datasets. Furthermore, the inclusion of PEL and MC contributed to further performance gains. Finally, adopting SS alongside all aforementioned technologies yielded the highest detection accuracy, showcasing the synergistic effect of combining these methodologies.

C. PERFORMANCE RESULT OF THE PROPOSED STUDY

Table 2 presents performance metrics for three different datasets in anomaly detection. The metrics include Area Under the Curve (AUC), Anomaly AUC, Average Precision (AP), and False Alarm Rate (FAR) values. For the UCF-Crime dataset, the AUC is 0.9009, with an Anomaly AUC of 0.7456 and an AP of 0.4090, and the FAR is 0.0204. Similarly, the XD-Violence dataset shows an AUC of 0.9509, Anomaly AUC of 0.8626, AP of 0.8648, and FAR of 0.0013. Lastly, the Shanghai dataset exhibits an AUC of 0.9869, Anomaly AUC of 0.8228, AP of 0.7780, and FAR of 0.0000.

D. STATE OF THE ART COMPARISON FOR UCF CRIME AND ShanghahiTech DATASET

The comparison Table 3 presents an overview of various crime detection models developed over multiple years. Each model is evaluated based on its performance on the

Shanghai Tech and UCF Crime datasets using the AUC (Area Under the Curve) metric. In 2018, Sultani et al. [8] introduced models utilizing the C3D and ID3 feature extractors. These models demonstrated competitive AUC scores on both datasets, indicating their efficacy in identifying crime-related activities in videos. Zhong et al. [6] further advanced the field in 2019 by exploring the use of the C3D and TSN feature extractors. Their experiments revealed varying performance across the datasets, emphasizing the importance of feature extractor selection in model development. Additionally, Zhong et al. [9] investigated the ID3 feature extractor, contributing additional insights into its suitability for crime detection tasks.

In 2020, Zaheer et al. [7], [49] introduced novel feature extractors such as C3D-self and C3D, achieving promising results on both datasets. Wan et al. [50] also contributed to advancements by exploring the I3D feature extractor, further diversifying the range of feature extractors used in crime detection models. The year 2021 marked significant progress in the field, with studies by Purwanto et al. [13], Tian et al. [19], Majhi et al. [51], Wu and Liu [44], Yu et al. [52], Lv et al. [12], and Feng et al. [53] introducing various feature extractors and achieving competitive results. These studies highlighted the continuous evolution and improvement of crime detection models. In 2022, the research landscape expanded further with a surge in model diversity. Studies by Zaheer et al. [3], [30], Joo et al. [20], Cao et al. [63], Li et al. [39], Sharif et al. [30], Yi et al. [57], Yu et al. [27], and Gong et al. [58] introduced novel approaches and feature extractors, pushing the boundaries of performance in video-based crime detection. Finally, the proposed hybrid model showcased exceptional performance, achieving remarkably high AUC scores on both datasets. This model represents a culmination of feature extraction techniques and model architecture advancements, underscoring the potential for further improvements in crime detection technology. Overall, the comparison table provides valuable insights into the evolution of crime detection models over the years, highlighting the importance of feature extraction techniques and model architecture design in achieving superior performance. As the field continues to advance, future models are expected to enhance the capabilities of video-based crime detection systems, contributing to improving public safety and security.

E. STATE OF THE ART COMPARISON FOR XD-VIOLENCE DATASET

Table 4 demonstrated the state-of-the-art comparison for the proposed model with the XD-Violence dataset. The proposed model outperforms existing state-of-the-art methods on the XD Violence Dataset, achieving an impressive average precision (AP) score of 86.26%. Sultani et al. [8] achieved an AP of 73.20% using RGB features, while HL-Net [10] attained slightly higher at 73.67%. Notably, incorporating audio features alongside RGB, HL-Net reached 78.64%. RTFM [19] and MSL [39] followed closely with scores of 77.81% and 78.28%, respectively. Pang et al. [64] and ACF [65] leveraged RGB with audio, achieving 81.69% and 80.13%, respectively. However, the proposed model significantly surpasses these benchmarks, demonstrating its efficacy in violence detection.

TABLE 4. State-of-the-art comparison of the proposed model for the XD Violence Dataset.

Method	Feature	AP (%)
Sultani et al. [8]	RGB	73.20
HL-Net [10]	RGB	73.67
HL-Net [10]	RGB+Audio	78.64
RTFM [19]	RGB	77.81
MSL [39]	RGB	78.28
MSL [39]	RGB	78.59
Pang et al. [64]	RGB+Audio	81.69
ACF [65]	RGB+Audio	80.13
MSAF [66]	RGB+Audio	80.51
CUPL [67]	RGB+Audio	81.43
CMA-LA [68]	RGB+Audio	83.54
MACIL-SD [69]	RGB+Audio	83.40
Pu et al. [28]	RGB	85.59
Proposed Model	RGB	86.26

VI. CONCLUSION AND FUTURE DIRECTION

In the study, we proposed a graph and general DL approach to extract discriminative features to effectively distinguish abnormal events from normality in weakly supervised video anomaly detection (WS-VAD) tasks. By addressing the limitations of existing approaches and proposing a multi-stage deep-learning model that integrates cutting-edge technologies, we have demonstrated the effectiveness of our method. Through the utilization of a ViT-based CLIP module, a CNN-based I3D module, an Uncertainty-regulated Dual Memory Units (UR-DMU) model, and GCN and Global/Local Multi-Head Self Attention (GL-MHSA) modules, we have successfully extracted and learned representations of regular and abnormal data simultaneously. The refinement of features in our third-stage module, a CNN-based MLP, further enhances the model's ability to differentiate between normal and anomalous behaviour. Besides anomaly detection, we believe that this model can be used to detect crimes and contribute to crime control automatically. Extensive experiments on multiple datasets have validated the superiority of our approach over state-of-the-art methods, showcasing its potential for real-world applications in anomaly detection tasks. We believe that our comprehensive solution offers significant promise, demonstrating enhanced efficacy and performance in video-based anomaly detection.

ABBREVIATIONS

WS-VAD	Weakly supervised video anomaly detection.
UR-DMU	Uncertainty-regulated dual memory units.
MLP	Multilayer perception.
TCA	Temporal contextual aggregation.
GCN	Graph convolutional networks.
GL-MHSA	Global/Local multi-head self attention.
MIL	Multiple instances learning.
NVs	Normal videos.

AVs	Anomalous videos.
DL	Deep learning.
CNN	Convolutional network.
ViT	Vision transformer.
BCE	Binary cross entropy.

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, Oct. 2019, pp. 1490–1499.

[2] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. Van Den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.

[3] M. Z. Zaheer, A. Mahmood, M. H. Khan, M. Segu, F. Yu, and S.-I. Lee, "Generative cooperative learning for unsupervised video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 14724–14734.

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

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

[6] 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. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1237–1246.

[7] 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 *Computer Vision—ECCV*. Glasgow, U.K.: Springer, 2020, pp. 358–376.

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

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

[10] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Computer Vision—ECCV*. Glasgow, U.K.: Springer, 2020, pp. 322–339.

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

[12] H. Lv, C. Zhou, Z. Cui, C. Xu, Y. Li, and J. Yang, "Localizing anomalies from weakly-labeled videos," *IEEE Trans. Image Process.*, vol. 30, pp. 4505–4515, 2021.

[13] 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. (ICCV)*, Montreal, BC, Canada, Oct. 2021, pp. 173–183.

[14] A. S. M. Miah, Md. A. M. 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.

[15] A. S. M. Miah, M. A. M. Hasan, Y. Okuyama, Y. Tomioka, and J. Shin, "Spatial–temporal attention with graph and general neural network-based sign language recognition," *Pattern Anal. Appl.*, vol. 27, no. 2, p. 37, 2024.

[16] A. S. M. Miah, M. A. M. Hasan, Y. 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.

[17] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognit.*, vol. 77, pp. 329–353, May 2018.

[18] Y. Liu, D. Yang, Y. Wang, J. Liu, J. Liu, A. Boukerche, P. Sun, and L. Song, "Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models," 2023, *arXiv:2302.05087*.

[19] 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. (ICCV)*, Montreal, BC, Canada, Oct. 2021, pp. 4955–4966.

[20] H. Kevin Joo, K. Vo, K. Yamazaki, and N. Le, "CLIP-TSA: CLIP-assisted temporal self-attention for weakly-supervised video anomaly detection," 2022, *arXiv:2212.05136*.

[21] 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. 2013.

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

[23] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "StyleCLIP: Text-driven manipulation of StyleGAN imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, BC, Canada, Oct. 2021, pp. 2065–2074.

[24] K. Vo, S. Truong, K. Yamazaki, B. Raj, M.-T. Tran, and N. Le, "AOE-Net: Entities interactions modeling with adaptive attention mechanism for temporal action proposals generation," *Int. J. Comput. Vis.*, vol. 131, no. 1, pp. 302–323, Jan. 2023.

[25] K. Yamazaki, K. Vo, S. Truong, B. Raj, and N. Le, "VLTinT: Visual-linguistic transformer-in-transformer for coherent video paragraph captioning," 2022, *arXiv:2211.15103*.

[26] J. Shao, X. Wen, B. Zhao, and X. Xue, "Temporal context aggregation for video retrieval with contrastive learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3267–3277.

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

[28] Y. Pu, X. Wu, L. Yang, and S. Wang, "Learning prompt-enhanced context features for weakly-supervised video anomaly detection," 2023, *arXiv:2306.14451*.

[29] 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, vol. 37, 2023, pp. 3769–3777.

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

[31] 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. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 4489–4497.

[32] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, Nov. 2019.

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

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and G. Krueger, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[36] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, 32, 2019, pp. 1–11.

[37] L. Harold Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," 2019, *arXiv:1908.03557*.

[38] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan, "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm," 2021, *arXiv:2110.05208*.

[39] S. Li, F. Liu, and L. Jiao, "Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 1395–1403.

[40] 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. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 8022–8031.

[41] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection—A new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6536–6545.

[42] 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. (ICIP)*, Kuala Lumpur, Malaysia, Oct. 2023, pp. 3230–3234.

[43] W. Zhu, P. Qiu, O. M. Dumitrascu, and Y. Wang, "PDL: Regularizing multiple instance learning with progressive dropout layers," 2023, *arXiv:2308.10112*.

[44] P. Wu and J. Liu, "Learning causal temporal relation and feature discrimination for anomaly detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3513–3527, 2021.

[45] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, and Y.-C. Wu, "MGFN: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 1, pp. 387–395.

[46] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 32, 2019, pp. 1–12.

[47] S. Gollapudi, *Learn Computer Vision Using OpenCV*. Berlin, Germany: Springer, 2019.

[48] T. Dozat, "Incorporating Nesterov momentum into Adam," in *Proc. 4th Int. Conf. Learn. Represent., Workshop*, 2016, pp. 1–4.

[49] M. Z. Zaheer, A. Mahmood, H. Shin, and S.-I. Lee, "A self-reasoning framework for anomaly detection using video-level labels," *IEEE Signal Process. Lett.*, vol. 27, pp. 1705–1709, 2020.

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

[51] 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)*, Nov. 2021, pp. 1–8.

[52] S. Yu, C. Wang, Q. Mao, Y. Li, and J. Wu, "Cross-epoch learning for weakly supervised anomaly detection in surveillance videos," *IEEE Signal Process. Lett.*, vol. 28, pp. 2137–2141, 2021.

[53] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, "MIST: Multiple instance self-training framework for video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 14004–14013.

[54] M. Z. Zaheer, A. Mahmood, M. Astrid, and S.-I. Lee, "Clustering aided weakly supervised training to detect anomalous events in surveillance videos," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 26, 2024, doi: 10.1109/TNNLS.2023.3274611.

[55] C. Cao, X. Zhang, S. Zhang, P. Wang, and Y. Zhang, "Weakly supervised video anomaly detection based on cross-batch clustering guidance," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2023, pp. 2723–2728.

[56] W. Tan, Q. Yao, and J. Liu, "Overlooked video classification in weakly supervised video anomaly detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2024, pp. 202–210.

[57] S. Yi, Z. Fan, and D. Wu, "Batch feature standardization network with triplet loss for weakly-supervised video anomaly detection," *Image Vis. Comput.*, vol. 120, Apr. 2022, Art. no. 104397.

[58] 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. IEEE Int. Conf. Multimedia Expo (ICME)*, Taipei, Taiwan, Jul. 2022, pp. 1–6.

[59] 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*.

[60] 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. (WACV)*, Jan. 2023, pp. 2664–2673.

[61] S. Sun and X. Gong, "Long-short temporal co-teaching for weakly supervised video anomaly detection," 2023, *arXiv:2303.18044*.

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

[63] C. Cao, X. Zhang, S. Zhang, P. Wang, and Y. Zhang, "Adaptive graph convolutional networks for weakly supervised anomaly detection in videos," *IEEE Signal Process. Lett.*, vol. 29, pp. 2497–2501, 2022.

[64] 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. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 12170–12179.

[65] 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. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 1980–1984.

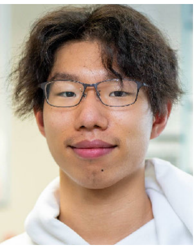
[66] D. Wei, Y. Liu, X. Zhu, J. Liu, and X. Zeng, "MSAF: Multimodal supervise-attention enhanced fusion for video anomaly detection," *IEEE Signal Process. Lett.*, vol. 29, pp. 2178–2182, 2022.

[67] C. Zhang, G. Li, Y. Qi, S. Wang, L. Qing, Q. Huang, and M.-H. Yang, "Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 16271–16280.

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

[69] 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, Oct. 2022, pp. 6278–6287.

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, South Korea, in 1990 and 1994, respectively, and the Ph.D. degree in computer science and communication engineering from Kyushu University, Japan, in 1999, under a scholarship from 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, Japan, in 1999, 2004, and 2019, respectively. He has coauthored more than 400 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, handwriting analysis, recognition, and synthesis. He is a member of ACM, IEICE, IPSJ, KISS, and KIPS. He served as the program chair and as a program committee member for numerous international conferences. He serves as an Editor for IEEE journals, Springer, Sage, Taylor & Francis, *Sensors* (MDPI), *Electronics* (MDPI), and Tech Science. He serves as an Editorial Board Member for *Scientific Reports*. He serves as a reviewer for several major IEEE and SCI journals.



YUTA KANEKO is currently pursuing the bachelor's degree in computer science and engineering with The University of Aizu (UoA), Japan. He joined the Pattern Processing Laboratory, UoA, in April 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.

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, and the Ph.D. degree in computer science and engineering from The University of Aizu, Japan, in 2024, under a scholarship from Japanese Government (MEXT). He assumed the positions of a Lecturer and an Assistant Professor with the Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology (BAUST), Saidpur, Bangladesh, in 2018 and 2021, respectively. He has been a Visiting Researcher (postdoctoral researcher) with The University of Aizu, since April 2024. He has authored or coauthored more than 50 publications in widely cited journals and conferences. His research interests include AI, ML, DL, human activity recognition (HCR), hand gesture recognition (HGR), movement disorder detection, Parkinson's disease (PD), HCI, BCI, and neurological disorder detection.

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

SATOSHI NISHIMURA (Member, IEEE) received the B.E. degree from Tohoku University, in 1987, and the M.Sc. and D.Sc. degrees in information science from The University of Tokyo, in 1989 and 1995, respectively. He is currently a Senior Associate Professor with The University of Aizu. His research interests include computer graphics and computer music. He is a member of ACM and IPSJ.