IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 March 2023, accepted 5 April 2023, date of publication 11 April 2023, date of current version 17 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3266345

RESEARCH ARTICLE

Contrastive-Regularized U-Net for Video Anomaly Detection

KIAN YU GAN[1], YU TONG CHENG[1], HUNG-KHOON TAN[1], HUI-FUANG NG[1], (Member, IEEE), MAYLOR KARHANG LEUNG[1], AND JOON HUANG CHUAH[2], (Senior Member, IEEE)

[1]Department of Computer Science, Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar 31900, Malaysia
[2]Department of Electrical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur 50603, Malaysia

Corresponding author: Hung-Khoon Tan (thkhoon@utar.edu.my)

This work was supported by the Fundamental Research Grant Scheme (FRGS) through the Ministry of Higher Education (MOHE) of Malaysia under Grant FRGS/1/2018/ICT02/UTAR/02/03.

ABSTRACT Video anomaly detection aims to identify anomalous segments in a video. It is typically trained with weakly supervised video-level labels. This paper focuses on two crucial factors affecting the performance of video anomaly detection models. First, we explore how to capture the local and global temporal dependencies more effectively. Previous architectures are effective at capturing either local and global information, but not both. We propose to employ a U-Net like structure to model both types of dependencies in a unified structure where the encoder learns global dependencies hierarchically on top of local ones; then the decoder propagates this global information back to the segment level for classification. Second, overfitting is a non-trivial issue for video anomaly detection due to limited training data. We propose weakly supervised contrastive regularization which adopts a feature-based approach to regularize the network. Contrastive regularization learns more generalizable features by enforcing inter-class separability and intra-class compactness. Extensive experiments on the UCF-Crime dataset shows that our approach outperforms several state-of-the-art methods.

INDEX TERMS video anomaly detection, weakly supervised learning, contrastive-based regularization, multi-instance learning, deep learning.

I. INTRODUCTION

Recent studies have shown that closed circuit television (CCTV) camera, when strategically installed, leads to a significant drop in crime rate [1]. However, large-scale deployment of CCTV may lead to data overload, making it difficult for the surveillance operators to pick up suspicious or abnormal activities hidden amidst the enormous streams of live CCTV footages. Therefore, there is a need for intelligent systems to automatically detect suspicious or anomalous activities.

Given a video, video anomaly detection (VAD) aims to localize the abnormal segments within a video. Unfortunately, abnormal events are rare and difficult to collect, leading to scarcity of positive samples. To overcome this issue, video anomaly detectors are normally trained with unsupervised learning [2], [3], [4] or weakly supervised methods [5], [6], [7], [8]. In the weakly supervised setting, the labels are provided at the video level. The label only specifies if a video contains anomalous event which may occur at any segments in the video. Fig. 1 shows the pipeline of a weakly supervised learning for VAD. The backbone network, e.g., C3D [9] and I3D [10], extracts segment-level generic features. Then, the anomaly feature extraction block transforms the generic features into specialized features, which in turn are used by the anomaly classifier to generate segment-level anomaly scores. In the absence of segment-level labels, the multiple-instance loss (MIL) formulation is normally applied on the segment scores for each video.

In this work, we focus on improving two aspects of the VAD network in a weakly supervised setting. First, we explore how to model the local and global temporal

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

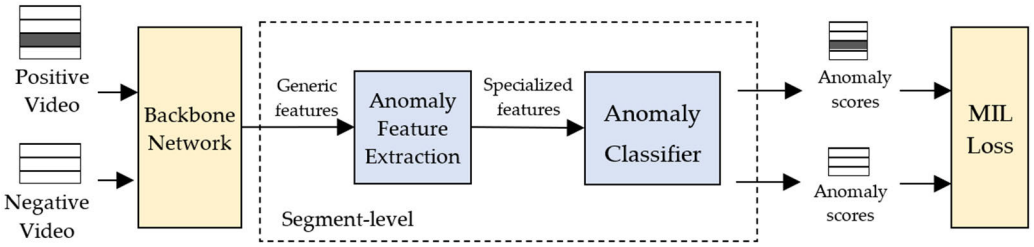


FIGURE 1. General pipeline for weakly supervised learning VAD.

dependencies when generating specialized features in the anomaly feature extraction block. Local information captures immediate anomalous traits, e.g., abrupt motion, suspicious human actions and changes in the environment, while global information provides the global context allowing the network to contrast normal and abnormal scenes in the videos.Previous methods such as stacked RNN [11], temporal consistency [12] and ConvLSTM [13] can only capture short range dependencies. GCN-based methods [14], [15] can model long-range dependencies but they are slower and more difficult to train. RTFM [7] captures both the short and long temporal dependencies using two parallel structures, one for each type. However, the two dependencies are considered separately, neglecting the close relationship between them. In this aspect, we propose to use U-Net like structure [16] to model both local and global dependencies for specialized features generation.

Second, we explore contrastive regularization as a new strategy to reduce overfitting. Overfitting is a dominant issue encountered when training VAD models due to the scarcity of positive samples. Traditionally, regularization is achieved by suppressing the complexity of the network [17], [18], injecting noise into the network [19], [20], [21], [22] or data [23], [24], and augmenting the training set [25]. For VAD, previous work has also applied special heuristics such as sparsity constraint and temporal smoothness [5] to regulate the output of the network. Our model adopts a feature-based approach to regularization where the strategy is to learn more generalizable features. Enhanced separability between normal and abnormal features makes the network less vulnerable to overfitting. To achieve this, we reformulate the contrastive regularization in [26] for VAD.

The main contributions of this work are highlighted as follows:

- We propose a U-Net like structure [16] to perform specialized feature extraction. U-Net has mainly been applied to image segmentation and trained in a supervised setting. In our model, U-Net is novelly used to localize abnormal segments in a video and trained with a weakly supervised setting. The network learns to generate segment-level pseudo labels to facilitate training with only video-level labels. The interaction between the two types of dependencies are embedded naturally in the network structure - the encoder learns global dependencies on top of local dependencies through successive convolution operations while the decoder propagates these global information back to the local level through transposed convolutions.
- We propose a novel weakly supervised contrastive regularization technique to reduce overfitting. Previously, contrastive regularization [26] was applied in the image domain under a supervised setting. For VAD, contrastive regularization is extended to a weakly supervised setting. Contrastive regularization is reformulated as a multiple-instance learning (MIL) problem where each video is a bag of segments. A negative bag (normal video) only comprises negative instances (segments) while a positive bag (abnormal video) contains both positive and negative instances whose labels are unknown. The loss function learns two sets of centers to represent normal and abnormal events, respectively. By enforcing intra-center compactness and inter-center separability among the samples, contrastive regularization enhances the discriminability and generalizability of the learnt feature. The model has good explainability since the centers represent different kinds of events by mapping a segment to the nearest center, we can justify why the segment is classified as such.
- Our work achieves state-of-the art performance on the UCF-Crime bench-mark. In the experiments, the proposed U-Net-based architecture captures the temporal information more effectively than existing methods, while the contrastive regularization learns more generalizable features, resulting in less overfitting and improved test performance.

The remaining of this paper is organized as follows. Section II discusses related works. Section III explains the proposed U-Net based feature extractor (Section III-A), the anomaly classification block (Section III-B) and weakly supervised contrastive-based regularization (Section III-C). Section IV presents the experimental results over the UCF-Crime benchmark. Finally, Section V summarizes the paper.

II. LITERATURE REVIEW

Video anomaly detection (VAD) is challenging due to the absence of training samples. To alleviate the issue, VAD has

traditionally relied on unsupervised [27], [28], [29] or weakly supervised learning [5], [6], [7] for training. Regularization is another critical step to overcome the overfitting issue prevalent in VAD models. In this section, we review related works in these areas.

A. UNSUPERVISED ANOMALY DETECTION

Unsupervised anomaly detection focuses on one-class classification where the model is trained exclusively on normal training data. The strategy is to build a model that specializes at reconstructing normal samples with small reconstruction error. These methods make the assumption that unseen abnormal videos are difficult to reconstruct accurately and regard samples with high reconstruction errors as an anomaly. Reconstruction can be done through sparse coding [27], [28], [30], [31] or auto-encoder [2], [12], [29], [32]. Sparse coding encodes normal patterns with a dictionary and the sample is reconstructed by linearly combining the dictionary bases [30], [31] such that the reconstructed feature is as close to the original feature. A sparse representation allows the model to represent high-dimensional samples with less training data. For the auto-encoders, the encoder compresses a sample into an encoded representation, which in turn is used by the decoder to reconstruct it. More complex schemes such ALOCC [35] and AVID [36] use adversarial training to train the encoder-decoder network to reconstruct an input data that fools the discriminative network into thinking that it is the original one. To do this, ALOCC enhances the inliers and distorts the outliers to enhance their separability while AVID in-paints the input data to remove pixel-wise irregularity from the input frame. To enhance the result, [32] proposes a two-stage cascade classifier based on sparse filtering and auto-encoder network such that anomalous regions have low sparsity value and high reconstruction cost. To tackle the lack of positive samples, G2D [37] uses generative adversarial network (GAN) to generate outliers. However, the generated outliers are not based on true realistic anomalous events. In general, unsupervised models are unable to handle complex or unseen environments. They typically suffer from higher false positive since it is unrealistic to capture all normal samples.

B. WEAKLY SUPERVISED ANOMALY DETECTION

Current state-of-the-art VAD systems are based on weakly supervised approaches [5], [6], [7], [14], [15] where video-level labels are leveraged to train the network. The multiple-instance learning (MIL) formulation is typically employed to cater for the absence of segment-level labels. For example, the classical multi-instance ranking loss trains the network to rank the top segment in a positive (abnormal) video to be higher than that in the negative (normal) video [5]. However, the top segment may not be an abnormal segment as desired, and the max operator cannot handle videos with multiple abnormal segments. To resolve this, Zhu et al. [38] extends the ranking loss by incorporating a temporal mechanism to localize anomalies. Several recent works trains the network to generate segment-level pseudo labels as supervisory signals [6], [7]. This allows the VAD to be trained with classical supervised learning. For example, Yu Tian et al. [7] selects top-k segments with the highest feature magnitude whereas Feng et al. [6] trains a MLP-based structure to generate pseudo-labels via multiple instance learning.

C. TEMPORAL DEPENDENCIES

It has been shown that temporal relationship between segments is critical towards the performance of VAD models [7], [11], [12], [14], [15]. Different networks have been used to capture the temporal information including recurrent models in [11], graph convolutional network methods (GCN) in [14], [15] and transformer in [7]. The recurrent models mainly model short-term relationship effectively whereas the transformer focuses on long-range relationship. Reference [15] enforces temporal consistency to clean up noisy labels where by propagating supervisory signals from high-confidence snippets to its neighbouring low-confidence snippets. High-order Context Encoding [8] enriches the features by using a moving window to capture the local dynamics of a video. RTFM [7] models the local and global temporal dependencies explicitly with two parallel structures. The pyramid of dilated convolutions (PDC) [39] is used to model local temporal dependency. The other branch uses the temporal self-attention module (TSA) [40] to capture the global temporal dependencies. However, RTFM neglect the close relationship between the two dependencies. Although local dependency is good at capturing local temporal dynamics variations, some anomalies are subtle and may not be discernable unless viewed in relationship to the other parts of the video. In this paper, we propose a unique approach to model global and local dependencies in a unified structure based on the U-Net architecture.

Regularization. Overfitting is a long-standing issue for VAD systems, mainly due to the scarcity of positive samples. Therefore, regularization is the key to successful training of VAD models. Structure-based regularization methods regularize the networks by manipulating the network structure. For example, L2 penalty [17] or elastic net loss [18] are imposed to suppress the complexity of the network. More recent methods drop subnet [41], path [42], channels [22] or layers [43] to reduce co-adaptation between the various computational units in the network during training. On the other hand, data-based regularization methods increase the diversity and size of the training set by creating transformed version of the training samples [25]. To increase the robustness of the system to handle real-world noise, it is useful to inject artificial noise to the data. For example, cutoff [23] and random erasing [24] cut out random portion of the image to avoid co-adaptation of the features and learn stronger features. Recently, feature-based regularization has been shown to improve generalization performance. Contrastive regularization [26], [44], [45] imposes geometric constraints such

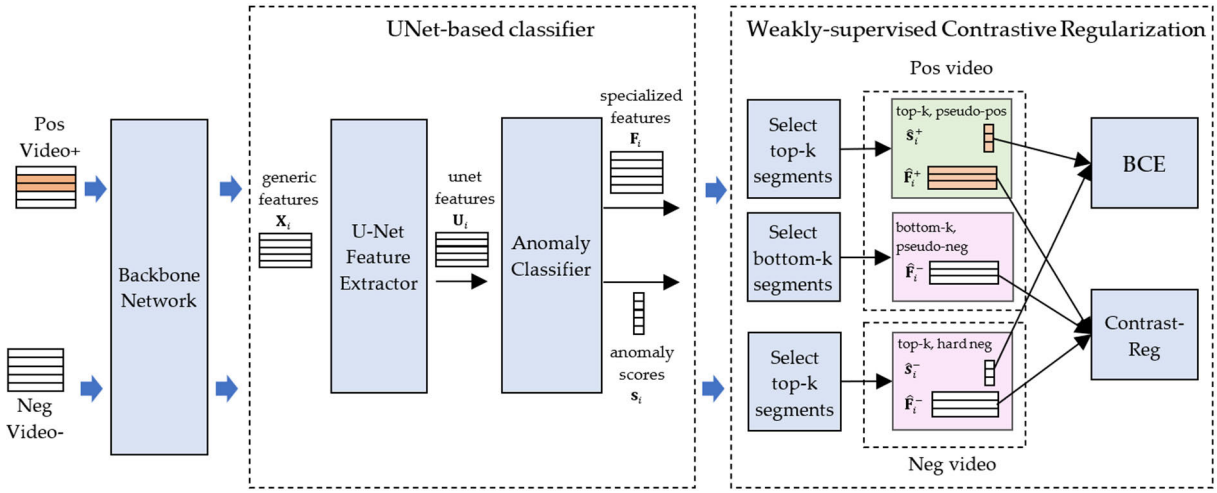


FIGURE 2. Proposed weakly supervised contrastive regularized UNet for VAD.

that the learnt features display good intra-class compactness and inter-class separability. However, contrastive regularizations have mainly been studied under the supervised learning for image classification. In this work, we extend contrastive regularization technique to weakly supervised setting for video anomaly classification.

III. THE PROPOSED METHOD: CONTRASTIVE-REGULARIZED U-NET

Fig. 2 shows the system overview of the proposed contrastive-regularized U-Net method. The network aims to localize abnormal segments using only weak video-level label where segment-level annotations are not available. The Ftraining data is denoted as $\mathcal{T} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{|\mathcal{T}|}$ where each video $\mathbf{X}_i \in \mathbb{R}^{T \times D_g}$ is a sequence of T segment-level features each with a dimensionality D_g, and $y_i \in \{0, 1\}$ is the corresponding video-level label to indicate the absence or presence of anomalous segments in the video. The input features are generic features extracted from a pre-trained 3D-CNN network such as C3D [9] or I3D [10].

The segment-level input features are first processed by the *U-Net feature extractor* to capture the local and global temporal dependencies between the segments in the video. The local dependencies are captured in the lower layers and the global dependencies in the higher layers. The output of the U-Net feature extractor is the U-Net features $\mathbf{U}_i \in \mathbb{R}^{T \times D_u}$ which has been enriched with temporal information. Note that $\mathbf{U}_i$ preserves the temporal resolution as in $\mathbf{X}_i$.

The U-Net features are then passed to the *anomaly classification block* which generates the segment-level anomalous scores $\mathbf{s}_i \in \mathbb{R}^T$ indicating if each segment is normal or anomalous. The block also outputs the segment-level features $\mathbf{F}_i \in \mathbb{R}^{T \times D_f}$ which are used to regularize the network.

The network is trained with a weakly supervised setup where only video-level annotations are provided. To do this, the U-Net based classifier is used to generate pseudo-labels. For *positive* videos, the top-k segments with the highest anomalous scores are selected as pseudo-positive samples and the bottom-k segments as pseudo-negative samples. The features and scores of the selected pseudo-positive and pseudo-negative segments are denoted as $\left\{\left(\hat{\mathbf{F}}_i^+, \hat{\mathbf{s}}_i^+\right), \left(\hat{\mathbf{F}}_i^-, \hat{\mathbf{s}}_i^-\right)\right\}$, respectively. For *negative* videos, only the top-k segments $\left(\hat{\mathbf{F}}_i^-, \hat{\mathbf{s}}_i^-\right)$ are selected and they represent hard negative normal segments that the network has more difficulty fitting (their anomaly scores are higher although they should ideally be lower). The anomaly scores $\hat{\mathbf{s}} = \left\{\hat{\mathbf{s}}_i^+, \hat{\mathbf{s}}_i^-\right\}$ are used to compute the data loss and train the model. Meanwhile, the generated features $\hat{\mathbf{F}} = \left\{\hat{\mathbf{F}}_i^+, \hat{\mathbf{F}}_i^-\right\}$ are used to perform contrastive regularization to reduce overfitting.

A. MODELING LOCAL AND GLOBAL TEMPORAL DEPENDENCIES WITH U-NET

Temporal relationship has been shown to be critical to the performance of VAD models [7], [11], [12], [14], [15]. Local temporal dependencies capture short-term and tangible anomalous cue (e.g., abrupt motion or scene change, and suspicious actions) while global temporal dependencies provide the global context to expose more subtle abnormal events. Different structures have been proposed to model temporal dependencies, e.g., pyramid of dilated convolutions (PDC) [39], temporal self-attention module (TSA) [40], High-order Context Encoding (HCE) [8], Temporal Consistency Graph (TCG) [15]. However, each of them specializes at capturing either local and global information, but not both. For example, PDC and HCE focuses on local temporal information, while TSA is more effective at modeling global relationship. TCG captures both relationships but the model is inefficient and difficult to train.

In this section, we novelly employ the U-Net [16] to capture both types of dependencies in a unified manner. While

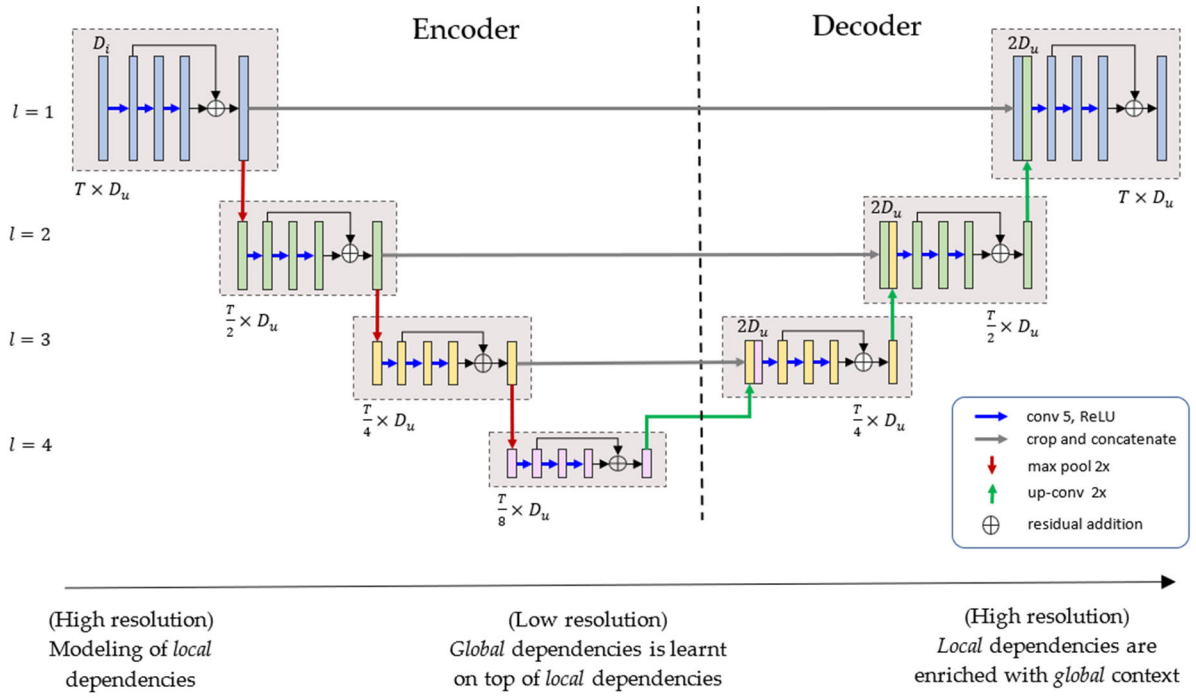


FIGURE 3. Modeling local and global dependencies with UNet.

U-Net has been widely adopted for image segmentation tasks such as medical imaging [46], our work is the first to apply U-Net to localize anomalous segment in a video. Fig. 3 shows the proposed U-Net adapted for VAD. The input to the network is $\mathbf{X}_i \in \mathbb{R}^{T \times D_g}$ which stores the sequence of snippet-level features from the backbone network. U-Net captures the temporal dependencies among the snippets to enrich the output features $\mathbf{U}_i \in \mathbb{R}^{T \times D_u}$. The network comprises a downstream path (encoder) which encodes the input features into a temporally compact representation, and an upstream path (decoder) which restores the segment-level features. The encoder network captures the local dependencies at the shallower layers and global dependencies at the deeper layers. This is because the effective receptive fields in a CNN are local at shallower layers and grows gradually as the network grows deeper [47]. At the decoder network, the activation maps from deeper layers are concatenated with those of the current layer to learn more fine-grained representation. Thus, the decoder combines the global and local information.

The height of the network is fixed to $L = 4$ where the temporal resolution $T^{(l)} = \frac{T}{2^{l-1}}$, $l = \{1, \ldots, L\}$ is halved from one height level to another. Different from conventional U-Net which increases the channels with increasing height, our network uses a regular channel size of D_u for all blocks. The network is constructed with a common residual block structure. The block contains 3 1-D convolutional layers activated by the ReLU function and has a skip connection to facilitate training. The convolutional layers are configured such that the activation maps in the same block have a regular shape of $T^{(l)} \times D_u$.

1) ENCODER

The encoder learns the local and global temporal dependencies in the input sequence. The local dependencies are captured through 1-D convolutional operations where the optimal kernel size is empirically determined to be 5. Stacking multiple convolutional operations allows the network to learn the global context in a hierarchical manner. The effective *receptive field* in the network grows incrementally from one layer to the next, allowing the network to learn increasingly global information from local ones in previous layers. Following conventional U-Net design, the temporal resolution is halved from one layer to another through max pooling. In the end, the output of the encoder is a temporally compact representation with high semantic value and global temporal coverage. This encoding process has been shown to be effective at removing noise and capturing common patterns that are representative of the training samples.

2) DECODER

The decoder takes the encoded message and restores it to segment-level features. Transposed convolution is applied to increase the temporal resolution from one layer to the next. The up-sampled features are concatenated with the encoder output at the same height level, and then passed to the decoder block for further feature extraction. Through this process, high-level global information is propagated through the layers back to the local segments. Thus, the segment-level features generated by decoder are infused with high-level local and global temporal information.

3) DESIGN CONSIDERATION

Both Convolutional Neural Networks (CNN) and Transformers are able to capture local and global information in their layer representations. However, they have very different characteristics. By design, a CNN-based network such as U-Net attends only locally in the lower layers and gradually builds up global-level attention at higher layers. In contrast, the transformer receptive field spans the whole temporal sequence in a single layer. It has been shown that a properly trained Visual Transformers (ViT) contains local heads and global heads even in the lowest layers [47]. However, without large-scale pre-training, ViT is empirically found to be weak at attending locally in the earlier layers, leading to much lower performance. For VAD, training data is scarce and therefore, U-Net structure is better suited for VAD than transformers under such constraint. For a more detailed reading on the difference between how transformers and CNN attends to local and global information, please refer to [47].

Another design consideration is the number of channels. Conventional U-Net increases the channels in the downstream path, and then decreases it in the upstream path. In our network, the number of channels in all blocks is fixed to D_u. This is because the input to a conventional U-Net are low-level 3-dimensional features (raw image input) whereas our network receives high-level high-dimensional features from the backbone network. While the conventional U-Net is originally designed to convert low-level input to high-level features, our network serves a different purpose of enriching an already high-level input features with temporal information. In addition, doubling the channels results in an unreasonably massive network and aggravates overfitting. Another option is to add a reduction layer before U-Net. However, the design results in information I and leads to lower performance in practice. Among these options, fixing the number of channels to D_u is found to yield the best performance.

B. SEGMENT-LEVEL ANOMALY CLASSIFICATION

The output of the U-Net is then fed into the anomaly classification block to generate the anomalous scores $s_i \in \mathbb{R}^T$ for all T segments in the video. The block is a simple 3-layered multi-layer perceptron (MLP) network. ReLU activation is used for the first and second layers which functions as feature extractor, and sigmoid activation for the last layer which functions as a binary classifier. In addition, the features from the second layer $\mathbf{F}_i \in \mathbb{R}^{T\times D_f}$ are extracted and subjected to contrastive regularization to reduce overfitting. To train the model, we adopt a weakly supervised framework where only weak video level annotations are available to train the segment-level classification network. This will be discussed in the next section.

C. WEAKLY SUPERVISED CONTRASTIVE REGULARIZATION

Since positive samples are scarce, VAD are especially vulnerable to overfitting issues. Therefore, regularization is a crucial step to improve generalization performance of the trained model. In this work, we propose a feature-based approach for regularization. The network is trained to generate more robust feature where events of the same nature generate similar features which are very different from features from other types of events. The more robust the generated features are, the more generalizable and noise-tolerant the model becomes.

1) TRAINING UNDER UNCERTAINTY

We extend contrastive regularization [26] to a weakly supervised setting where only the video label $y_i \in \{0, 1\}$ is provided. A video $\mathbf{X}_i \in \mathbb{R}^{T\times D_g}$ is labeled as positive if any of its segments is anomalous, and negative if none. Since segment-level labels are not available, the location of the abnormal event in a positive video is unknown. To deal with the uncertainty, the segment-level anomaly scores $\mathbf{s}_i \in \mathbb{R}^T$ output by the anomaly classification head can be employed to generate pseudo labels for each segment following the multi-instance learning (MIL) framework. Given a positive video, we extract *pseudo-positive* samples and *pseudo-negative* samples. The former is the set of top-k segments with the highest anomaly scores, while the latter is the set of bottom-k segments. For negative video, only the top-k segments are extracted. They represent *hard* negative samples that the network has difficulty classifying and needs more training. Despite possessing strong labels, not all segments from a negative video are selected to avoid data imbalance issue.

With the generated labels, the network can now be trained with traditional supervised learning. The multi-instance learning framework assumes that there exists common pattern (e.g., sudden movements, scene change, abnormal actions) among positive (abnormal) segments. When correctly selected as pseudo-positive samples, they update the network parameters in a coherent manner. In contrast, negative (normal) segments in different videos tends to be dissimilar to one another (e.g., different scenes and activities). When erroneously selected as pseudo-positive samples, they update the network in an incoherent manner. Consequently, over time, the network gradually becomes more adept at identifying the positive segments.

2) CONTRASTIVE REGULARIZATION WITH PSEUDO-LABELS

Fig. 4 shows how the original contrastive loss [26] is extended to a weakly supervised setting. Suppose the batch size is B. The features for the positive samples $\hat{\mathbf{F}}^+ = \bigcup \left\{\hat{\mathbf{F}}_i^+ | y_i = 1\right\}_{i=1}^B$ are collected by extracting pseudo-positive segments from positive videos in the batch. On the other hand, negative samples $\hat{\mathbf{F}}^- = \bigcup \left\{\hat{\mathbf{F}}_i^-\right\}_{i=1}^B$ are collected from all videos which includes bottom-k segments from positive videos and top-k segments from negative videos. This gives us the training set $\hat{\mathbf{F}} = \bigcup \left\{\hat{\mathbf{F}}^+, \hat{\mathbf{F}}^-\right\}$.

The contrastive regularization regularizes the network by learning discriminative features so that normal features are well separated from abnormal ones. To do this, we define C centers for each class to explicitly model different types

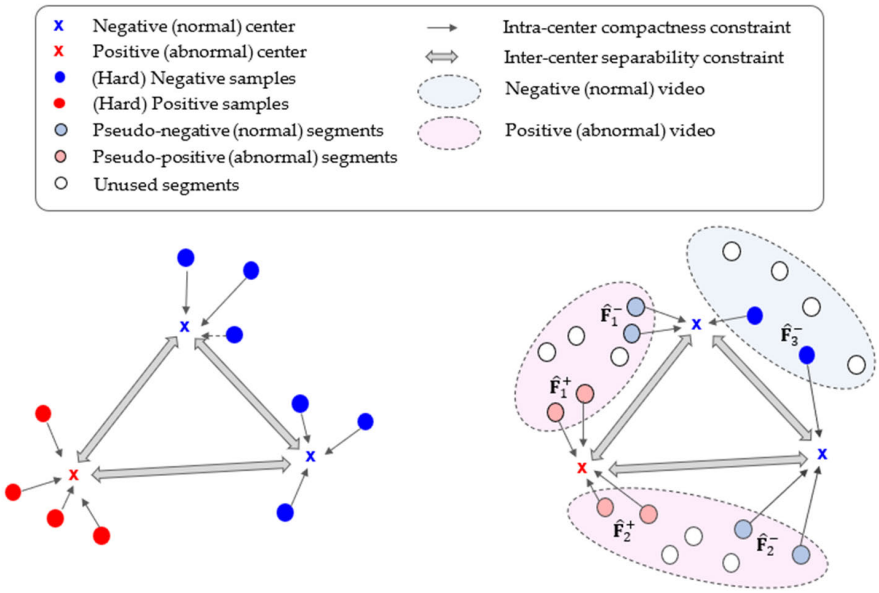


FIGURE 4. (a) Original (supervised) contrastive regularization [26] enforces intra-class compactness such that samples are near to the nearest centers of the same class, and inter-class separability by ensuring all centers are far from each other (b) Weakly supervised contrastive regularization extends [26] to handle weak video-level labels. For positive videos, top-k segments (light red) are selected as pseudo-positive samples, and bottom-k segments (light blue) as pseudo-negative samples. For negative video, the top-k segments (strong blue) are selected so that the network trains with hard normal samples that are difficult to classify. Some segments (white) are not used for training.

of normal and anomaly events, respectively. The centers are essentially network parameters and discovered through training. Let $\mathbf{H} = \left\{\mathbf{H}^+, \mathbf{H}^-\right\}$ where $\mathbf{H}^+ = \left\{\mathbf{h}_i^+\right\}_{i=1}^{C}$ and $\mathbf{H}^- = \left\{\mathbf{h}_i^-\right\}_{i=1}^{C}$ are the set of all positive and negative centers, respectively. The proposed contrastive regularization is given by:

$$
\begin{aligned}
&R_{contrast}\left(\hat{\mathbf{F}}, \mathbf{H}\right) \\
&= \lambda \left(\frac{1}{\left|\hat{\mathbf{F}}^+\right|} \sum_{\mathbf{f}^+ \in \min_{\mathbf{h}^+ \in \mathbf{H}^+} \hat{\mathbf{F}}^+} \left\|\mathbf{f}^+ - \mathbf{h}^+\right\|_2^2 \right. \\
&\quad \left. + \frac{1}{\left|\hat{\mathbf{F}}^-\right|} \sum_{\mathbf{f}^- \in \min_{\mathbf{h}^- \in \mathbf{H}^-} \hat{\mathbf{F}}^-} \left\|\mathbf{f}^- - \mathbf{h}^-\right\|_2^2 \right) \\
&\quad + \beta \frac{1}{C\,(C-1)} \\
&\quad \times \sum_{\mathbf{h}_i \in \mathbf{H}} \sum_{\mathbf{h}_j \in \mathbf{H}, j \neq i} \max\left(0, m - \left\|\mathbf{h}_i - \mathbf{h}_j\right\|_2^2\right) \qquad (1)
\end{aligned}
$$

where λ is the compactness strength, β is the separability strength, $|\cdot|$ is the cardinality of a set, and $\|\cdot\|_2$ is the L2-norm.

The first two terms enforce *intra-center compactness* which minimizes the distance between the sample $\mathbf{f}_i$ and its nearest center $\mathbf{h}_i$ of the same class. The first term ensures that pseudo-positive samples from positive videos are positioned near positive centers, whereas the second term ensures negative samples are distributed around negative centers. There are two types of negative samples, namely pseudo-negative samples from positive videos and hard-negative samples from negative videos. The pseudo-positive and pseudo-negative segments, both from positive videos, are pulled towards different types of class centers, allowing them to distinguish between the abnormal and normal events within the same video. This phenomena is indeed observed in our experiments (cf. Section IV).

The third term imposes *inter-center separability* such that all centers are well separated. If the distance between any two centers is smaller than m, it will incur a cost. The term ensures that the network learns a more diverse set of features, each representing different types of anomaly or normal events. In addition, explainability of the network is enhanced as the network's classification decision can be associated with the events associated with the nearest class center.

The role of the learnt centers is to generalize the different common patterns within each class and at the same time distinguish the two classes. Therefore, when the network generates feature that are near such centers, this results in reduction of overfitting.

D. LOSS FUNCTION

During training, the network generates pseudo-labels and assembles the batch $\hat{\mathbf{F}} = \left\{\hat{\mathbf{F}}^+, \hat{\mathbf{F}}^-\right\}$ with the corresponding anomaly scores $\hat{\mathbf{S}} = \left\{\hat{\mathbf{S}}^+, \mathbf{S}^-\right\}$. To reduce overfitting, contrastive regularization is performed with the learnt class centers $\mathbf{H} = \left\{\mathbf{H}^+, \mathbf{H}^-\right\}$. The training loss function is defined

as follows:

$$L\left(\hat{\mathbf{F}},\hat{\mathbf{S}},\mathbf{H}\right)=L\left(\hat{\mathbf{S}}\right)+R_{contrast}\left(\hat{\mathbf{F}},\mathbf{H}\right)+\gamma\sum_{\hat{\mathbf{s}}_i\in\hat{\mathbf{S}}}\sum_{t=1}^{T}\left(\hat{\mathbf{s}}_i\left(t\right)-\hat{\mathbf{s}}_i\left(t-1\right)\right)^2+\eta\sum_{\hat{\mathbf{s}}_i\in\hat{\mathbf{S}}}\sum_{t=1}^{T}\hat{\mathbf{s}}_i\left(t\right) \quad (2)$$

$L\left(\hat{\mathbf{S}}\right)$ is the binary cross entropy loss which minimizes the data loss to train the VAD model. $R_{contrast}\left(\hat{\mathbf{F}},\mathbf{H}\right)$ is the contrastive regularization used to regulate the network and reduce overfitting (Eq. 1). The third term is the temporal smoothness constraint such that the anomaly score of adjacent segments should be similar to each other. This avoids the anomaly scores from fluctuating irregularly. Lastly, the fourth term is the sparsity constraint which assumes that only a few segments in a positive video are anomalous. The strength of these different constraints can be finetuned by adjusting the intra-center compactness strength λ, inter-center separability strength β, temporal smoothness strength γ and sparsity strength η.

IV. EXPERIMENTS AND EVALUATION

A. DATASET DESCRIPTION AND EVALUATION MEASURE

1) UCF-CRIME

We evaluate our proposed network UCF-Crime [5], currently one of the largest public VAD dataset. The dataset consists of 1900 surveillance videos with equal number of abnormal and normal videos. It is split into 1610 training videos and 290 testing videos. The dataset is weakly labeled where the training set comes only with video-level labels whereas the testing set comes with frame-level labels for evaluation. The dataset is diverse and challenging. The videos are real-world surveillance data with complex and diverse background. There is a total of 13 types of anomalies in the dataset including explosion, arrest, abuse, fighting, shoplifting, stealing, and vandalism. The videos are untrimmed with a wide range of duration from 8 seconds to 9 hours.

2) EVALUATION

Following previous work on VAD [5], [6], [7], we evaluate our system (CR-U-Net) using the frame-level AUC measure, i.e., the area under the ROC curve. A larger AUC implies better performance. The AUC is computed at the frame level. Since the anomaly score is at the segment (clip) level, the same score is expanded to all frames in the segment for AUC computation. We also provide some qualitative result to evaluate the localization performance of our system and the effect of contrastive regularization. Comparison is made with unsupervised methods [2], [12], [30], [48], [49], [50] and weakly-supervised methods [5], [6], [7], [8], [15] including recent state-of-the-art methods such as MIST [6], RTFM [7] and HCE [8]. In particular, the evaluated weakly-supervised methods employ different structures to model temporal information. For example, GCN [15] uses a similarity graph, RTFM uses a combination of dilated convolution and transformers, and HCE uses windowing approach to encode local variations in time series.

B. IMPLEMENTATION DETAILS

Two types of backbone network, namely C3D [9] and I3D [10] are used to extract the segment-level features from a video. The network extracts features for every 16-frame clip in the video. For C3D, the feature dimension $D_g = 2048$, and for I3D, $D_i = 1024$. Following [5], the clips in a training video are grouped into 32 non-overlapping segments. The segment-level feature is obtained by averaging the clip features in each segment. For short videos (<32 clips), duplicate clips are inserted to the video. Through the process, all training videos have a uniform temporal resolution of 32 segments. For the testing set, no grouping is performed. Each clip is essentially a segment. So, the test videos have varying temporal length. For data augmentation, 10-crop augmentation is performed to bolster both the training data and testing data. The cropped frame size is 210×280 pixels which are 87.5% of the original size.

For the U-Net feature extractor, the number of channels D_u is set to 1024 for all blocks. For the classification layer, the number of neurons in the MLP is set to 512 units, 32 units and 1 unit, respectively. Dropout layer (drop_ratio = 0.7) is inserted between the fully connected layers.

Unless specified otherwise, the learning rate is set to 0.0001 and the model is trained using Adam optimizer for 500 epochs with a weight decay of 0.01 and batch size of 64. In each batch, we ensure that the number of normal and abnormal samples are equivalent to ensure balanced training. For contrastive regularization, the following settings are used: the number of centers $C = 16$, the intra-center compactness strength $\lambda = 0.0001$, the inter-center separability strength $\beta = 0.0001$, and the margin $m = 1.25$. The smoothness and sparsity constraints are set to $\gamma = \eta = 0.008$.

C. RESULTS ON UCF-CRIME

Table 1 shows the AUC results on UCF-Crime, which is currently one of the most realistic and largest VAD benchmark dataset. For the UCF-Crime dataset, we compare our system against unsupervised methods [2], [30], [48] and weakly supervised methods [5], [6], [7], [8], [15]. We use I3D features to build our model.

In general, the weakly supervised methods far outperform the unsupervised methods [2], [30], [48]. This shows that additional supervisory labels, even weak ones, are indispensable to train better models. In addition, I3D features delivers superior performance than C3D features due to a more powerful backbone network and large-scale pre-training.

The proposed CR-U-Net achieves state-of-the-art performance with an AUC of 85.24%. This is the second highest AUC among the evaluated methods. It outperforms MIL-Ranking [5] by 7.32%, MIST [6] by 2.94% and RTFM [7]

TABLE 1. AUC performance on UCF-Crime.

Supervision	Method	Feature	AUC (%)
Unsupervised	Conv-AE [2]	-	50.60
	Subspace-SVDD [51]	-	58.50
	Sparsity-Combination [30]	C3D RGB	65.51
	BODS [48]	I3D RGB	68.26
	GODS [48]	I3D RGB	70.46
Weakly supervised	MIL-ranking [5]	C3D RGB	75.41
	GCN-Anomaly [15]	C3D RGB	81.08
	MIST [6]	C3D RGB	81.40
	RTFM [7]	C3D RGB	83.28
	GCN-Anomaly [15]	C3D RGB	81.08
	MIL-ranking [5]	I3D RGB	77.92
	MIST [6]	I3D RGB	82.30
	RTFM [7]	I3D RGB	84.30
	HCE [8] (Original training set)	I3D RGB	84.44
	HCE + NS + HC [8]	I3D RGB	85.38*
	CR-UNet (Proposed)	I3D RGB	**85.24**

* The best performing model HCE trains on a noise-augmented (NS + HC) dataset. Without noise augmentation, our proposed model CR-UNet outperforms HCE.

by 0.94%. Out of all evaluated methods, HCE [8] delivers the highest AUC of 85.38%. However, HCE's superior performance is attributed to its augmentation strategy where hand-crafted anomalies (HC) and noise simulation (NS) are injected into the training set. Without HC and NS, the performance of HCE drops to 84.44%. Therefore, the proposed CR-U-Net actually outperforms HCE by 0.71% on the same original training set. HCE mainly captures local temporal information. This makes CR-U-Net one of the top performing models for the UCF-Crime dataset. This indicates the efficacy of U-Net whose structure is able to capture both local and global temporal information. In comparison, Although RTFM captures both local and global dependencies, they are implemented in two parallel independent structures and are unable to model the interaction between the two dependencies. In contrast, U-Net models both dependencies more effectively, resulting in state-of-the-art performance.

D. COMPARATIVE AND ABLATION STUDY

We performed an ablation study on UCF-Crime to evaluate the effectiveness of the proposed U-Net anomaly classifier as well as contrastive regularization. First, we replace the U-Net anomaly classifier with other kinds of structure. The first is a 3-layered MLP network from [5] which does not model any temporal information. The second is a combination of pyramid of dilated convolution (PDC) and transformer structure (TSA) [7]. The former models the local temporal dependencies while the latter models the global temporal dependencies separately in two branch. To ensure fairness, the training of the compared models are standardized using binary cross entropy loss and top-k pseudo labels without contrastive regularization. Lastly, the fourth model applies contrastive regularization on top of the network with U-Net-based anomaly classifier, which represents our optimal model. Table 2 shows the results of the AUC of the three model.

TABLE 2. Comparative and ablation studies.

MLP [5]	PDC+TSA [7]	U-Net	ContrastReg	AUC (%)
✓				81.74
	✓			82.20 (+0.46)
		✓		83.43 (+1.23)
		✓	✓	85.24 (+1.81)

As expected, the MLP-based classifier has the lowest AUC of 81.74%. When the MLP is replaced with PDC and TSA, the AUC improves by 0.46% to 82.20%. This shows that learning temporal dependencies is useful to detect anomalies in VAD. However, PDC+TSA models the local and global temporal dependencies separately, neglecting the close relationship between them. The proposed U-Net-based network resolves this problem by modeling both dependencies in a single structure. Compared to PDC+TSA, the U-Net structure improves the AUC by 1.23% to 83.43%.

Next, we evaluate the effectiveness of contrastive regularization aimed at reducing overfitting by learning more generalizable features. When contrastive regularization is applied, the AUC jumps significantly by 1.81% to 85.24%. This shows that features that are more class separable are indeed more generalizable. Section IV-F provides more analysis on this property of contrastive regularization.

E. FINE-RUNNING THE MODEL

In this section, we evaluate the impact of channel and filter size in the U-Net classification block, and the number of centers in contrastive regularization.

1) NUMBER OF CHANNELS

To study the impact of the channel sizes D_u in U-Net, we evaluate with the following channel sizes: 256, 512, 1024 and 2048. Different from traditional U-Net, in our design, all

TABLE 3. Impact of convolutional channel size.

Channel Size, D_u	AUC (%)
256	84.08
512	84.85
1024	**85.24**
2048	83.78

TABLE 4. Impact of filter size in residual block.

Filter Size, f	AUC (%)
3	83.44
5	**85.24**
7	83.84

convolutional layers have the same channel size. In our implementation, the input to U-Net from the backbone network has a channel size of 1024. So, the first two settings ($f = 256$ or 512) shrink the channels while the last setting ($f = 2048$) expands the channels. Table 3 shows the experimental result.

The best AUC is achieved when U-Net uses the same channel size as the backbone network feature ($D_u = 1024$). Setting to a lower channel size, $D_u = 256$ or 512, lowers the AUC performance. This may be due to information bottleneck where useful information are lost when the channels are compressed. Setting a bigger channels size, $D_u = 2048$, decreases the performance as well because a bigger network is more difficult to optimize and easier to overfit without sufficient training samples.

2) FILTER SIZE

Next, we evaluate the impact of filter size in U-Net. The filter sizes of 3, 5 and 7 are evaluated. The filter size affects the range of the temporal coverage. The filter size should neither be too big nor too small.

As shown in Table 4, the optimal filter size is $f = 5$. Using a larger filter size $f = 7$ makes it less sensitive to subtle local cues. It also incurs more parameters, and therefore takes longer and more resources to train and is easier to overfit when training samples are insufficient. On the other hand, a smaller filter size $f = 3$ has a limited temporal range, hindering it from extracting longer-range temporal dependencies effectively with the current depth level.

3) NUMBER OF CENTERS

The number of centers C in contrastive regularization signifies the number of representative events captured by the model. We evaluate $C = 0$, 2, 4, 8, 16 and 24. When $C = 0$, contrastive regularization is disabled. Table 5 shows the experimental results.

When the contrastive regularization is enabled ($C \geq 2$), the performance of the model improves. With $C = 2$ (one center per class) the AUC improves by 0.87%. The optimal number of centers is $C = 16$ where the AUC improves significantly by 2.1% to 85.24%. Increasing the number of centers any further to 24 does not yield any further performance improvement. This is because 24 centers are more difficult to train and 16 centers are sufficient to explain the anomalies.

TABLE 5. Impact of number of centers on UCF-Crime.

Number of Centers, C	AUC (%)
0	83.14
2	84.01
4	84.09
8	84.01
16	**85.24**
24	84.89

F. QUALITATIVE ANALYSIS

In this section, we perform qualitative analysis of the model. Fig. 5 shows the frame anomaly scores in the test videos produced by our model. Fig. 5(a) – (e) shows the scores for six anomalous videos. In general, the anomaly scores align successfully with the ground truth where the network outputs high anomaly scores for anomalous regions and low scores for normal regions. Fig. 5(f) shows that the network correctly outputs low anomaly scores for all frames in a normal video. Our system is not perfect. Fig. 5(g) shows a missed detection where the network misses a shop-lifting event where a man put a stolen watch into his pocket. Our model fails to detect the shop-lifting event because the action is subtle and the stolen item is occluded, making it difficult to detect even for a non-observant human. Fig. 5(h) shows a false alarm when a group of people made some body contact. The model confuses it to be a fighting event for a short time duration.

Fig. 6 looks deeper into the chronology of events for 2 test videos containing burglary and explosion events. The first video "Burglary037" starts with an empty cashier counter (Frame 78). Then, a man appears and climbs over the counter (Frame 257), passes several stolen wines to his accomplice (Frame 796), ransacks the cash register and then fleets (Frame 1815). The network correctly predicts the initial scene as normal and a higher anomaly scores constantly after the burglar appears. The model predicts a lower anomaly score after the burglars leave. The second video "Explosion033" contains two different explosion events. The anomaly scores are lower for the scenes preceding the two explosions at their respective locations (frames 499 and 1505) and high anomaly scores when the explosions occur (frames 1075 and 2095). This shows that the model is able to detect anomalous events in the frame despite being trained with video-level labels.

G. IMPACT OF CONTRASTIVE REGULARIZATION

To evaluate the effectiveness of contrastive regularization, we compare the distance between a segment feature $\mathbf{f}$ to its nearest normal center $\mathbf{h}^- \in \mathbf{H}^-$ and its nearest anomaly center $\mathbf{h}^+ \in \mathbf{H}^+$. The relative distance is computed as follows:

$$Dist\,(\mathbf{f}, \mathbf{H}) = \min_{\mathbf{h}^- \in \mathbf{H}^-} \left\|\mathbf{f} - \mathbf{h}^-\right\|_2^2 - \min_{\mathbf{h}^+ \in \mathbf{H}^+} \left\|\mathbf{f} - \mathbf{h}^+\right\|_2^2 \quad (3)$$

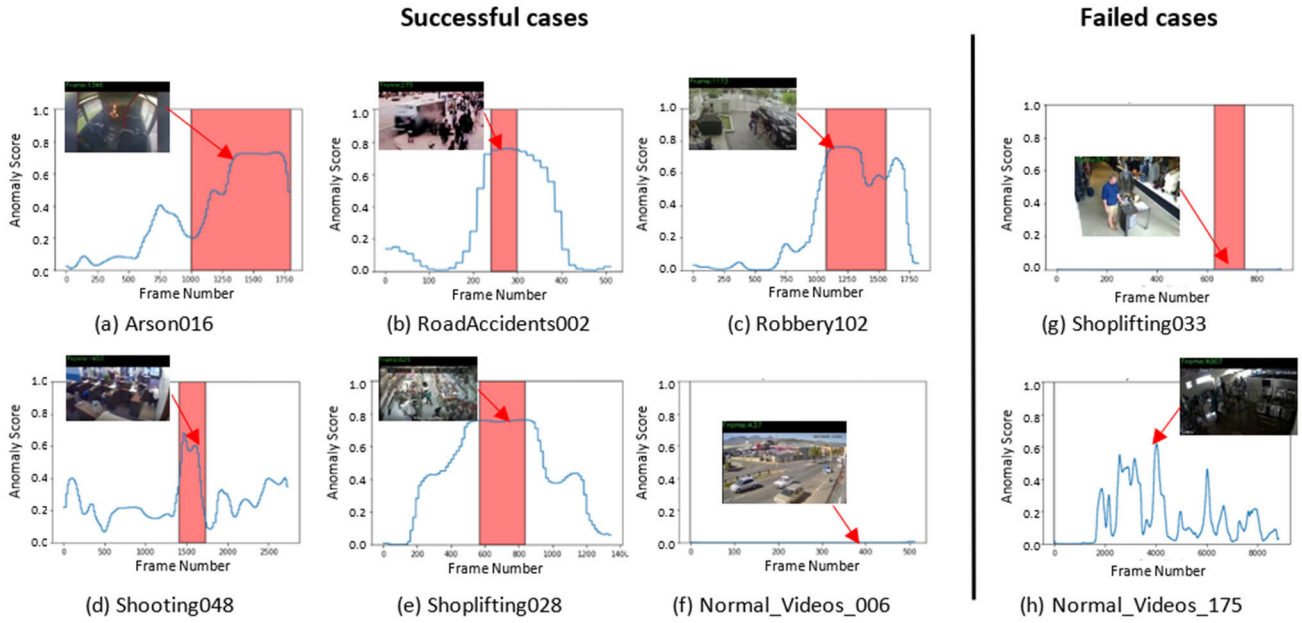


FIGURE 5. The anomaly scores for our method on test videos from UCF-Crime. The red colour region indicates the region where an anomaly event occurs. (a) – (e) shows that the anomaly scores generated by the model aligns with the anomalous regions containing arson, road accident, robbery, shooting and shoplifting. (f) show a normal video without an anomaly event. (g) and (h) show 2 failure cases where (g) is a missed detection and (h) is a false alarm.

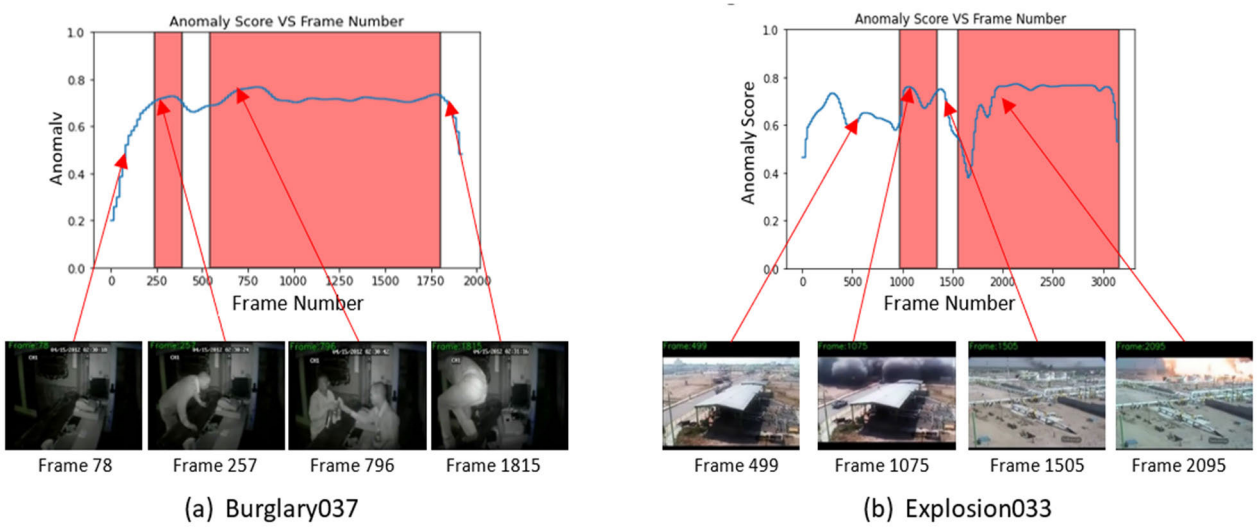


FIGURE 6. Chronology of events and their corresponding anomaly scores for two UCF-Crime test videos.

A negative value indicates that **f** is nearer to a normal center than to an abnormal center. Conversely, a positive value indicates that it is nearer to an abnormal center. Fig. 7 shows the computed relative distance of the segments in eight test videos. Fig. 7(a) - (e) show 5 test videos with anomaly events while (f) show a test video without anomaly events. The embedding feature of the normal event (white region) is closer to the normal center (below the horizontal line) while the embedding feature of the anomaly event (red zone) is closer to the anomaly center (above the horizontal line). This shows that the features generated wth contrastive regularization are more discriminative and aligns very well to the correct event type. Fig. 7(g)-(h) show 2 failure cases on a test video with and without an anomaly event, respectively. For the former, the features clearly fails to capture the subtle anomalous event. For the latter, the features generated shows that the normal actions are occasionally confused as abnormal due to high level of activities in the scene.

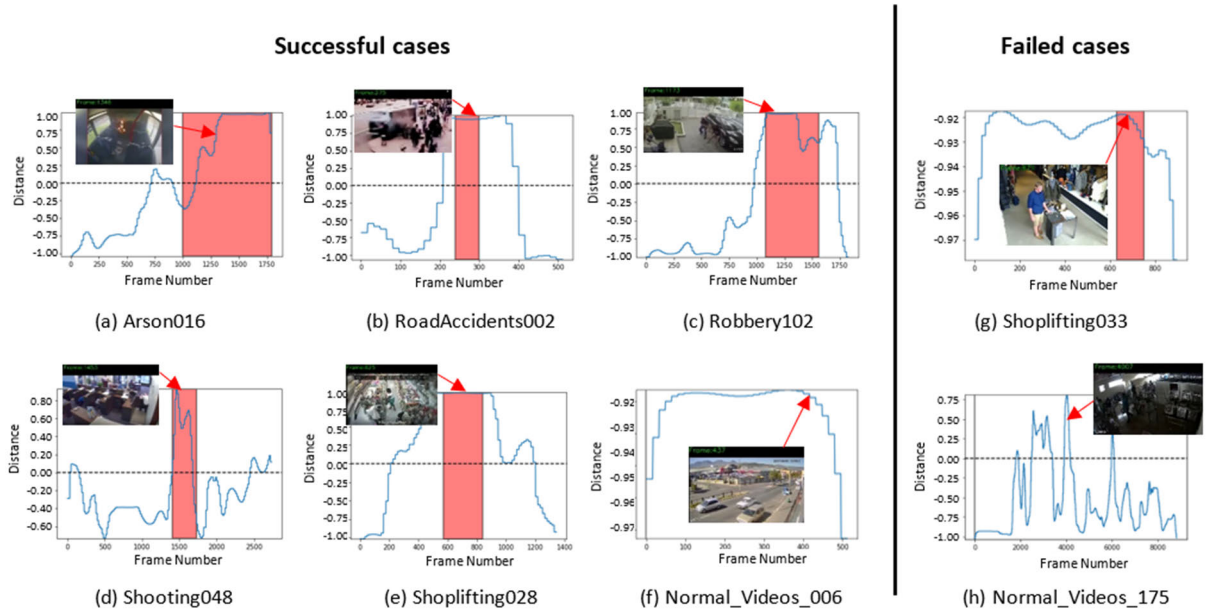


FIGURE 7. Relative distance between video segments and the normal vs abnormal centers. The red colour region shows the groundtruth location of the anomaly event in a video. When a segment's relative distance is above the dotted line (positive value), the segment is nearer to an anomaly center. When it is below the dotted line (negative value), it is nearer to a normal center. (a)-(f) shows the successful cases where the segments are generally nearer to the correct class centers. For the normal video (f), the distance for all segments are all negative (near to a normal center). (g)-(h) show 2 failure cases where the segments are not near to the correct centers.

V. CONCLUSION

This work novelly applies U-Net to capture both local and global temporal information for detecting anomalous segments in a video. U-Net learns global temporal dependencies on top of local dependencies in the encoder, and this global information is then propagated back to the local level in the decoder. This intricate way of modeling the two types of dependencies are found to be superior to RTFM which models the two types of dependencies separately. To reduce overfitting issues, we propose weakly supervised contrastive regularization. The loss function enforces inter-class separability and intra-class compactness. The resultant features are found to be discriminative and generalizable, resulting in improved test performance. For future work, it is interesting to explore 3D U-Net structure which additionally considers the spatial dimension, and to study how different distance metrices affect contrastive regularization.

REFERENCES

[1] E. L. Piza, B. C. Welsh, D. P. Farrington, and A. L. Thomas, "CCTV surveillance for crime prevention: A 40-year systematic review with meta-analysis," *Criminol. Public Policy*, vol. 18, no. 1, pp. 135–159, Feb. 2019.

[2] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.

[3] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11996–12004.

[4] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14372–14381.

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

[6] 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)*, Jun. 2021, pp. 14009–14018.

[7] 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)*, Oct. 2021, pp. 4975–4986.

[8] 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.

[9] 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)*, Dec. 2015, pp. 4489–4497.

[10] 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)*, Jul. 2017, pp. 4724–4733.

[11] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 341–349.

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

[13] W. Liu, W. Luo, Z. Li, P. Zhao, and S. Gao, "Margin learning embedded prediction for video anomaly detection with a few anomalies," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3023–3030.

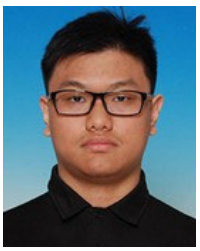
[14] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 322–339.

[15] 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)*, Jun. 2019, pp. 1237–1246.

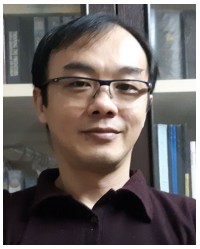
[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Oct. 2015, pp. 234–241.

[17] A. Krogh and J. Hertz, "A simple weight decay can improve generalization," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 1991, pp. 950–957.
[18] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc., B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.
[19] Y. Yamada, M. Iwamura, T. Akiba, and K. Kise, "Shakedrop regularization for deep residual learning," *IEEE Access*, vol. 7, pp. 186126–186136, 2019.
[20] Y. Yamada, M. Iwamura, and K. Kise, "Shakedrop regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2018, pp. 1–4.
[21] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 648–656.
[22] S. Hou and Z. Wang, "Weighted channel dropout for regularization of deep convolutional neural network," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Jan. 2019, pp. 8425–8432.
[23] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.
[24] M. Saran, F. Nar, and A. N. Saran, "Perlin random erasing for data augmentation," in *Proc. 29th Signal Process. Commun. Appl. Conf. (SIU)*, Jun. 2021, pp. 13001–13008.
[25] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Jul. 2019.
[26] M. Tanveer, H.-K. Tan, H.-F. Ng, M. K. Leung, and J. H. Chuah, "Regularization of deep neural network with batch contrastive loss," *IEEE Access*, vol. 9, pp. 124409–124418, 2021.
[27] W. Luo, W. Liu, D. Lian, J. Tang, L. Duan, X. Peng, and S. Gao, "Video anomaly detection with sparse coding inspired deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1070–1084, Mar. 2021.
[28] P. Wu, J. Liu, M. Li, Y. Sun, and F. Shen, "Fast sparse coding networks for anomaly detection in videos," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107515.
[29] B. Wang and C. Yang, "Video anomaly detection based on convolutional recurrent AutoEncoder," *Sensors*, vol. 22, no. 12, p. 4647, Jun. 2022.
[30] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2720–2727.
[31] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *Proc. CVPR*, Jun. 2011, pp. 3313–3320.
[32] M. Sabokrou, M. Fathy, and M. Hoseini, "Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder," *Electron. Lett.*, vol. 52, no. 13, pp. 1122–1124, Jun. 2016.
[33] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2017, pp. 439–444.
[34] X. Wang, W. Xie, and J. Song, "Learning spatiotemporal features with 3DCNN and ConvGRU for video anomaly detection," in *Proc. 14th IEEE Int. Conf. Signal Process. (ICSP)*, Aug. 2018, pp. 474–479.
[35] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3379–3388.
[36] M. Sabokrou, M. Pourreza, M. Fayyaz, R. Entezari, M. Fathy, J. Gall, and E. Adeli, "Avid: Adversarial visual irregularity detection," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Mar. 2019, pp. 488–505.
[37] M. Pourreza, B. Mohammadi, M. Khaki, S. Bouindour, H. Snoussi, and M. Sabokrou, "G2D: Generate to detect anomaly," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2003–2012.
[38] Y. Zhu and S. Newsam, "Motion-aware feature for improved video anomaly detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2019, pp. 1–12.
[39] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2016, pp. 1–13.
[40] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.
[41] T. Yang, S. Zhu, and C. Chen, "Gradaug: A new regularization method for deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Dec.*, vol. 2020, pp. 14207–14218.
[42] G. Larsson, M. Maire, and G. Shakhnarovich, "FractalNet: Ultra-deep neural networks without residuals," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2017, pp. 1–11.
[43] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 646–661.
[44] K. Zhao, J. Xu, and M.-M. Cheng, "RegularFace: Deep face recognition via exclusive regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1136–1144.
[45] M. Tanveer, H. Tan, H. Ng, M. Leung, and J. Chuah, "Batch contrastive regularization for deep neural network," in *Proc. 12th Int. Joint Conf. Comput. Intell.*, 2020, pp. 1–10.
[46] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.
[47] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2021, pp. 12116–12128.
[48] J. Wang and A. Cherian, "GODS: Generalized one-class discriminative subspaces for anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8200–8210.
[49] 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.
[50] G. Yu, S. Wang, Z. Cai, E. Zhu, C. Xu, J. Yin, and M. Kloft, "Cloze test helps: Effective video anomaly detection via learning to complete video events," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 583–591.
[51] F. Sohrab, J. Raitoharju, M. Gabbouj, and A. Iosifidis, "Subspace support vector data description," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 722–727.

KIAN YU GAN is currently pursuing the bachelor's degree in computer science with Universiti Tunku Abdul Rahman (UTAR), Malaysia. His research interests include software development, machine learning, video analysis, and deep learning.



YU TONG CHENG is currently pursuing the bachelor's degree with Universiti Tunku Abdul Rahman (UTAR), Malaysia. His current research interests include the Internet of Things, smart sensors, and deep learning.



HUNG-KHOON TAN received the B.Eng. degree in computer engineering from Universiti Teknologi Malaysia, Malaysia, and the M.Phil. and Ph.D. degrees in computer science from the City University of Hong Kong, Hong Kong. He was a Test Development Engineer and a Senior Design Engineer with Altera, Penang, Malaysia, from 1999 to 2004. He joined the Department of Computer Science, Universiti Tunku Abdul Rahman (UTAR), in 2011. His current research interests include image processing, computer vision, and machine learning.

HUI-FUANG NG (Member, IEEE) received the Ph.D. degree in biosystems and agricultural engineering from the University of Minnesota, USA, in 1996. He was a Software Engineer with PPT Vision Inc., Minnesota, from 1996 to 2003. He was with the Department of Computer and Information Science, Asia University, Taiwan, from 2003 to 2013. He is currently with the Department of Computer Science, University Tunku Abdul Rahman (UTAR), Malaysia. His research interests include image processing, computer vision, and machine learning.

MAYLOR KARHANG LEUNG received the B.Sc. degree in physics from National Taiwan University, in 1979, and the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Saskatchewan, Canada, in 1983, 1985, and 1992, respectively. He is currently a Professor with Universiti Tunku Abdul Rahman, Malaysia. His research interests include computer vision, pattern recognition, image processing, video surveillance for human behavior analysis with respect to robbery, and theft detection.

JOON HUANG CHUAH (Senior Member, IEEE) received the B.Eng. degree (Hons.) from Universiti Teknologi Malaysia, the M.Eng. degree from the National University of Singapore, and the M.Phil. and Ph.D. degrees from the University of Cambridge. He is currently the Head of the VIP Research Group and an Associate Professor with the Department of Electrical Engineering, Faculty of Engineering, University of Malaya. His main research interests include image processing, computational intelligence, IC design, and scanning electron microscopy. He is a fellow and was the Honorary Secretary of the Institution of Engineers, Malaysia (IEM). He was the Honorary Treasurer of the IEEE Computational Intelligence Society (CIS) Malaysia Chapter and the Honorary Secretary of the IEEE Council on RFID Malaysia Chapter. He is the Chairman of the Institution of Engineering and Technology (IET) Malaysia Network. He is also a Chartered Engineer registered under the Engineering Council, U.K., and also a Professional Engineer registered under the Board of Engineers, Malaysia.