Real-Time Weakly Supervised Video Anomaly Detection

Hamza Karim
University of South Florida
hamzakarim@usf.edu

Keval Doshi
University of South Florida
kevaldoshi@usf.edu

Yasin Yilmaz
University of South Florida
yasiny@usf.edu

Abstract

Weakly supervised video anomaly detection is an important problem in many real-world applications where during training there are some anomalous videos, in addition to nominal videos, without labelled frames to indicate when the anomaly happens. State-of-the-art methods in this domain typically focus on offline anomaly detection without any concern for real-time detection. Most of these methods rely on ad hoc feature aggregation techniques and the use of metric learning losses, which limit the ability of the models to detect anomalies in real-time. In line with the premise of deep neural networks, there also has been a growing interest in developing end-to-end approaches that can automatically learn effective features directly from the raw data. We propose the first real-time and end-to-end trained algorithm for weakly supervised video anomaly detection. Our training procedure builds upon recent action recognition literature and trains a large video model to learn visual features. This is in contrast to existing approaches which largely depend on pre-trained feature extractors. The proposed method significantly improves the anomaly detection speed and AUC performance compared to the existing methods. Specifically, on the UCF-Crime dataset, our method achieves 86.94% AUC with a decision period of 6.4 seconds while the competing methods achieve at most 85.92% AUC with a decision period of 273 seconds.

1. Introduction

This paper investigates the effectiveness of end-to-end training for real-time weakly supervised video anomaly detection (wVAD). As opposed to the unsupervised VAD problem [15], in which only nominal videos are used in training, anomalous videos with video-level labels are also available for training in wVAD [16]. The lack of frame-level anomaly labels differentiates wVAD from supervised VAD. Recent wVAD approaches extract features from a video using large pre-trained video models and process them using elaborate deep neural networks [10], [17], [19], [6]. These feature aggregation networks are trained together with a shallow model using a deep metric learning loss, such as the multiple-instance learning (MIL) loss [16] to detect anomalous events. These models process a full video together during inference to refine the features across the entirety of the video. Hence, they focus on offline anomaly detection after observing the entire video with decision period typically being equal to the video length. Figure 1 shows the impact of reducing the decision period during inference on three recent state-of-the-art methods [17], [19], [6] on the UCF-Crime dataset [16]. Results indicate that the performance of state-of-the-art models is proportional to the number frames processed at a time, with significantly reduced performance for real-time or near-real-time performance as shown by the dashed rectangle in Figure 1 [1]. In this paper, we show that a much improved real-time detection performance is possible (shown by star in Figure 1.)

Figure 1. Performance drop in state-of-the-art methods with decreasing decision period. The proposed method (shown with star) significantly improves timely detection performance with 86.94% AUC and 6.4 sec decision period.

An important premise of video anomaly detection in many applications, including video surveillance, is real-time or near-real-time detection to enable timely response. This poses a question: *Can wVAD models be trained with a*

[1] While the notion of real-time decision heavily depends on the application, we consider less than 30 sec to be real-time or near-real-time for video anomaly detection.

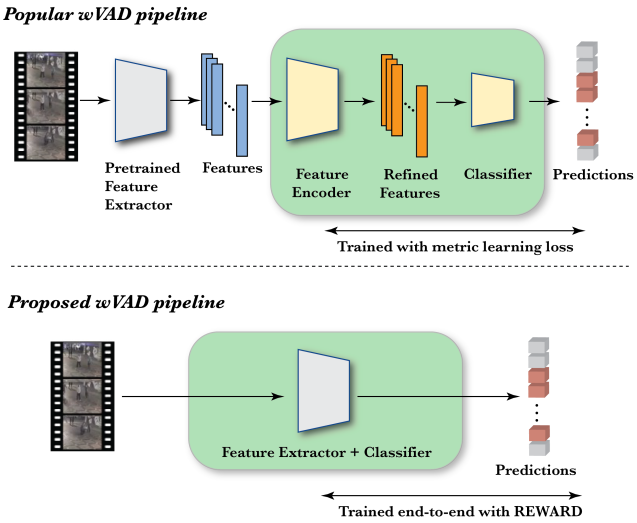

Figure 2. Proposed approach simplifies and outperforms the popular wVAD approaches which only train the feature encoder and classifier networks in an ad hoc manner. End-to-end training of feature extractor in our approach provides compactness and superior real-time performance.

short decision period to enable real-time inference?

In this paper, we show that the answer to this question is positive through end-to-end training of modern transformer-based models using a novel self-supervised learning network based on kNN (k-nearest-neighbors) distances and uniform frame sampling. *We show that using a decision period of 6.4 seconds the proposed method, called REWARD, outperforms state-of-the-art methods which use decision periods 43 times ours.*

End-to-end training is not feasible with the existing wVAD methods since the commonly used metric learning losses require uploading two batches of nominal and anomalous videos in a single GPU together. The batch sizes necessary for gradient estimates of sufficient quality cause a memory bottleneck problem. Hence, as summarized in Figure 2, the existing methods limit the training to the feature encoder for feature refinement and do not train the feature extractor, which is typically pre-trained on a large action recognition dataset, such as Kinetics-600.

In the proposed method, the end-to-end trained video model is directly used for inference, greatly simplifying the wVAD pipeline, as shown in Figure 2. Our contributions can be summarized as follows:

- We introduce *a novel end-to-end solution*, called REWARD, for wVAD systems.
- It *enables real-time anomaly detection* on test videos with 6.4 sec decision period.
- It *outperforms state-of-the-art methods* which use 273 sec decision period on popular wVAD datasets such as UCF-Crime and XD-Violence.

2. Related Work

Video Anomaly Detection: The motivation for wVAD is that in some applications it may be possible to roughly label videos as anomalous, without specifying where and when the anomaly happens, to obtain a representative set of anomalies of interest.

Sultani et al. [16] first introduced a deep MIL ranking loss framework to detect anomalous segments. MIST [10] used an encoder-based method that fine-tunes a feature encoder based on the generated pseudo-labels. RTFM [17] uses feature magnitude with the multi-scale temporal scenario from the video to select the top-k segments to determine the abnormality of a segment in a video. This paper also introduced the first feature aggregation MTN network. MTN has since been a staple backbone for current research. S3R [19] uses dictionary-based self-supervised learning to generate en-normal and de-normal features, which also uses the MTN network [17] to retrieve enhanced features to create pseudo-anomaly video features. MGFN [6] introduces a Glance-and-Focus module along with Magnitude Contrastive loss to increase the separability of normal and abnormal features.

There has been some works on real-time anomaly detection in an unsupervised setting [8, 9]. However, in the wVAD setting, the focus has been on offline anomaly detection. Recent works share the popular pipeline which can be summarized as in Figure 2. We propose an end-to-end trained method that outperforms existing methods in real-time detection while providing a more computationally efficient solution to the wVAD problem.

Feature Extractors: Deep neural networks have made substantial progress in action recognition and feature extraction from videos. Two main categories of these models are 3D convolutional neural networks (3D-CNN) and more recent Transformer models. 3D-CNNs leverage 3D convolution to capture spatial and temporal information. Notable examples of 3D-CNNs include C3D [18] and I3D [5], which extract video features effectively. The former employs both RGB images and optical flow in two-stream networks, while the latter uses raw video frames directly, operating as a 3D network. wVAD literature has predominantly focused on utilizing I3D as a feature extractor.

Transformer-based video models, such as TimeSformer [2], Video Swin Transformer [13], ViViT [1], and Uniformer [12] have shown significant improvement in video understanding. While TimeSformer is a completely attention-based model, Uniformer combines attention and convolution to further improve video understanding. Recently, we also see the use of transformer-based models in wVAD research, such as MGFN [6], which reports results using a Swin Transformer [13] model as a feature extractor.

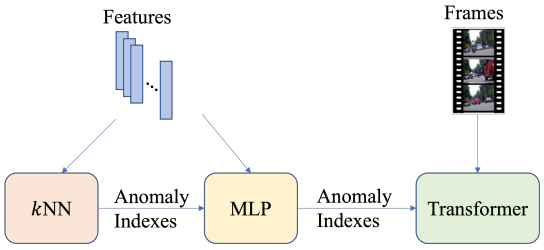

Figure 3. Self-supervised structure of REWARD.

3. Proposed Method

For end-to-end training large video models on wVAD data, we propose a self-supervised learning method called REWARD (**R**eal-Time **E**nd-to-End **W**eakly Supervised Video **A**nomaly **R**elevance **D**etector). As illustrated in Figure 3, it consists of incremental training of a stronger classifier with the help of a weaker classifier in three steps. In the first step, a kNN classifier network is trained on features from a pretrained video model. In the second step, the predicted anomalous frame indexes from the first step are used to train a Multi-Layer Perceptron (MLP) classifier and in the final step, the raw video frames of anomalous frame indexes predicted by MLP are used to end-to-end train a large video model such as a Transformer. Before explaining the details of REWARD, we first discuss the background and motivation.

3.1. Motivation

End-to-End Training refers to the process of training all the neural network parts together to perform a task directly from the raw input to the final output, without any intermediate, separately trained feature refinement steps. This approach can bring about significant advantages. It can simplify the overall machine learning pipeline by eliminating the need for feature refinement. Moreover, it can lead to better real-time performance as the model can learn task-specific representations directly from the raw input data without the need for feature aggregation. It has been applied successfully in various domains, including computer vision, natural language processing, and speech recognition [3, 7].

Memory Bottleneck of Metric Learning Loss: Metric learning losses are a family of loss functions used in machine learning in cases where the output labels for training instances are not sufficient for the use of loss functions used in supervised learning, such as binary cross entropy and mean squared error [11]. These losses are specifically designed to learn a metric space, where distances between samples in the learned space correspond to their semantic dissimilarity. The aim of metric learning losses is to minimize the distance between samples of the same class and maximize the distance between samples of different classes,

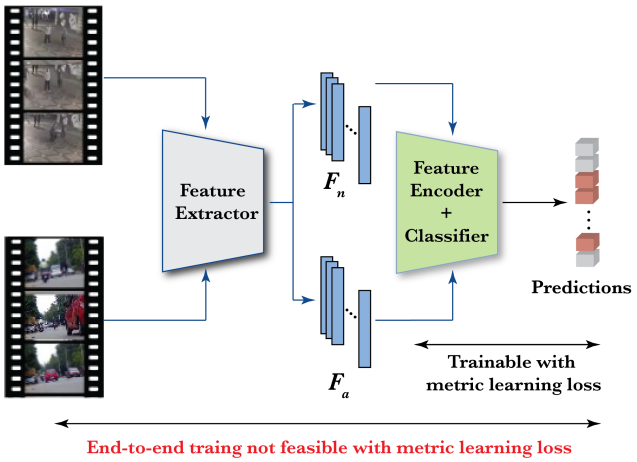

Figure 4. To train the feature encoder and classifier using a metric learning loss, F_n and F_a are processed together to promote the separability between normal and anomalous segments. Nevertheless, training a feature extractor end-to-end using a metric learning loss is not feasible due to two reasons: processing together large number of video clips in a single GPU entails huge memory requirements and feature extractor has orders of magnitude more parameters than feature encoder.

thereby improving the quality of features.

In recent works, MIL ranking loss, a metric learning loss, has played a key role in wVAD systems [16]. This loss function has a requirement that is impracticable for end-to-end training of feature extractors in a wVAD setting (Figure 4). MIL loss requires an equal number of anomaly videos and nominal videos to be trained together at once.

In the existing works utilizing metric learning loss, while training only the feature encoder and classifier with multiple bags of anomalous and normal features together is feasible on a single processing unit, training the feature extractor end-to-end using a metric learning loss would require raw video clips to be processed together (Figure 4). This is infeasible because the memory requirement of processing a large number of raw video clips at once is inordinate. Moreover, the number of trainable parameters in feature extractor is orders of magnitude higher than that of feature encoder.

3.2. End-to-End Training via REWARD

We propose a solution to the memory bottleneck problem by converting the wVAD problem with video-level labels into a classification problem with frame-level pseudo-labels. Specifically, we generate segment-level pseudo-labels and replace metric learning loss functions with the more memory-friendly loss functions such as BCE (binary cross entropy).

To find the anomalous segments and perform end-to-end training, we propose the REWARD network with architec-

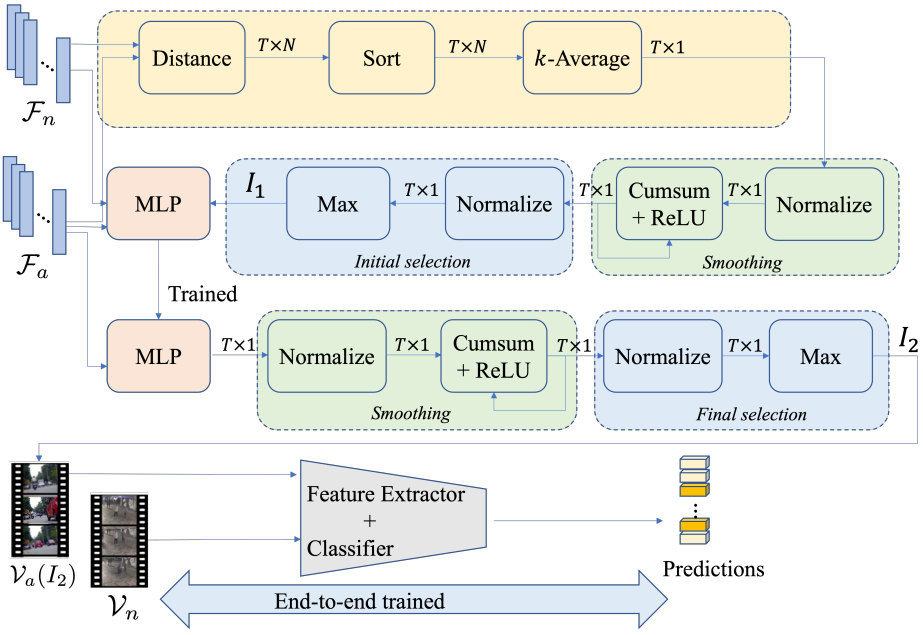

Figure 5. Proposed REWARD network for self-supervised end-to-end training of large video models in wVAD problems.

ture shown in Figure 5. It consists five types of modules: kNN distance calculation, smoothing, selection, multi-layer perceptron (MLP), and a large video model for feature extraction and classification.

We first divide each video into T segments by uniformly sampling at a rate of 5 frames/sec and extract features using a pretrained feature extractor to get feature representations for nominal and anomalous videos as $F_n \in \mathbb{R}^{T\times P}$ and $F_a \in \mathbb{R}^{T\times P}$ respectively, where P is the dimension of the feature vector. All features from nominal and anomalous videos are collected into the sets

$$\mathcal{F}_n = \{F_n^i\}_{i=1}^{N_n T},\ \mathcal{F}_a = \{F_{aj}^t\}_{j=1,t=1}^{N_a,T},$$

where N_n and N_a are the numbers of nominal and anomalous videos in the training set, and j denotes the anomalous video index.

Distance Calculation: The distance calculation module consists of four layers. In the distance layer, each segment $F_{aj}^t, t = 1,\ldots,T$, in an anomalous video is compared to $\mathcal{F}_n$ by computing the Euclidean distance δ_{ti} between F_a^t and all $F_n^i \in \mathcal{F}_n$, yielding a $T \times N$ distance matrix

$$\Delta = [\delta_{ti}]_{t=1,i=1}^{T,N},$$

which integrates the temporal information in the anomalous video. Then, in the sorting layer, the N distances $[\delta_{ti}]_{t,i=1}^{N}$ in each row t of Δ are sorted in the ascending order. Finally, an average pooling layer is applied to each row only by taking the average of the first k elements (i.e., distances to the nearest k neighbors), resulting in the T-dimensional kNN distance vector δ_t. Each kNN distance δ_t gives us an estimate of the similarity of each segment t within an anomaly video to the nominal segments. Small values of δ_t mean a higher degree of similarity to normal action and vice-versa.

Smoothing: Considering the temporal continuity of actions in video, a smoothing step is applied to avoid spurious dissimilarities. We first normalize the distance values around zero by subtracting the average value

$$\tilde{\delta}_t = \delta_t - \frac{1}{T}\sum_{t=1}^{T}\delta_t, \tag{1}$$

and then apply the cumulative sum operation followed by the ReLU activation function in a recurrent way to compute the anomaly evidence for each segment:

$$D_t = \max\{0, D_{t-1} + \tilde{\delta}_t\},\ D_0 = 0. \tag{2}$$

Initial Selection: After smoothing, the anomaly evidence time series D_t takes values around zero for segments similar to the nominal segments and positive values for segments not so similar to the nominal ones. Since it is known that some segment(s) in each anomalous video are anomalous, the ones with largest D_t values are good candidates. We select a small percentage of segments by focusing on the segments with largest anomaly evidence D_t. After normalizing D_t by its largest value,

$$\tilde{D}_t = \frac{D_t}{\max_t D_t} \in [0,1], \tag{3}$$

we pick the segments with normalized evidence greater than $\lambda \in (0.5, 1)$ as an initial anomalous set:

$$I_1^j = \{t : \tilde{D}_t \geq \lambda,\ t = 1, \ldots, T\}, \tag{4}$$

where λ is a threshold close to 1 (e.g., 0.8 used in the experiments) and j denotes the anomalous video index. Combining the initial sets from all anomalous videos we form the set $I_1 = \bigcup_{j=1}^{M} I_1^j$, where $\bigcup$ denotes the union of sets and M is the number of anomalous videos.

MLP and Final Selection: To generalize the selection of anomalous segments for the entire video, we train an MLP using the segments with strong anomaly evidence in the initial selection set I_1 as label 1 and nominal segments as label 0. Once the MLP is trained using BCE loss, all the segments in the anomalous videos are passed through the trained MLP to estimate their probability of being anomalous, $p_t,\ t = 1, \ldots, T$ for each anomalous video. Then, a final smoothing and selection procedure is applied for each anomaly video. Using Eq. (1), p_t is normalized to the zero-mean $\tilde{p}_t$. Then, its rectified cumulative sum R_t is obtained using Eq. (2). Finally, the normalized version $\tilde{R}_t$ of R_t (Eq. (3)) is used to make the final selections. Since the anomaly scores in $\tilde{R}_t$ are more reliable than the evidences $\tilde{D}_t$ in the initial selection module thanks to the trained MLP classifier, we use a lower threshold here, the average value of $\tilde{R}_t$:

$$I_2^j = \{t : \tilde{R}_t \geq \bar{R},\ t = 1, \ldots, T\},\ \bar{R} = \frac{1}{T}\sum_{t=1}^{T} \tilde{R}_t$$

$$I_2 = \bigcup_{j=1}^{M} I_2^j. \tag{5}$$

End-to-End Training: With the pseudo-labelled set I_2 of anomalous segments available, we can now turn the wVAD task into a binary classification problem which lends itself to end-to-end training of the feature extractor. By labeling the video segments $\mathcal{V}_a(I_2)$ as 1 and the nominal video segments $\mathcal{V}_n$ as 0, we train a feature extractor in an end-to-end fashion using the BCE loss. We use a single neuron with the logistic sigmoid activation function at the classification layer to compute the anomaly relevance probability.

Real-Time Inference: After training, the feature extractor with the single-neuron classifier is used in an online fashion with a small decision period (6.4 sec) for real-time inference (see Figure 2). The details of real-time inference and computational efficiency can be found in Sections 3.3 and 4.4.

3.3. Implementation Details

Since we do not use feature aggregation, instead of I3D, we used a more modern transformer-based feature extractor in the experiments, Uniformer-32, that can benefit from end-to-end training. In recent works, consecutive 16 frames are sampled, and the features are extracted using the I3D feature extractor. We also sample consecutive 32-frame segments for the state-of-the-art methods to obtain results with the new Uniformer features for a fair comparison. Then, during training, each video is condensed into 32 segments through linear interpolation. In testing, existing methods can produce an anomaly score for each individual 16-frame segment, but effective feature aggregation (e.g., MTN [6, 17, 19]) requires all or most segments to be passed together at once. The significant drop in the detection performance with smaller number of segments for inference (i.e., decision period) is shown in Figure 1.

To address this, we first uniformly sample each video at 5 frames/sec. Leveraging Uniformer-32, which require an input of 32 frames, a mere 6.4 seconds suffice to accumulate 32 frames, yielding a 6.4-second contextual scope. We set the number of neighbors as $k = 20$, and the initial selection threshold as $\lambda = 0.80$. When training the MLP, which consists of 2 hidden layers each with 1000 neurons, we set the learning rate as $5\text{e}{-5}$ with a weight decay of 0.001. Adam optimizer is used for training all networks.

During the calculation of D_t using kNN distances (Eq. (2)), we observed that a considerable number of anomaly training videos in the UCF-Crime dataset exhibit banners (Figure 6). Typically, these banners appear at the start or end of a video and are static images. In comparison to the nominal video data, banners are relatively uncommon, which leads to the feature representations F_a for such segments being markedly distinct when compared to the nominal feature set $\mathcal{F}_n$. Consequently, this produces anomalously high D_t scores.

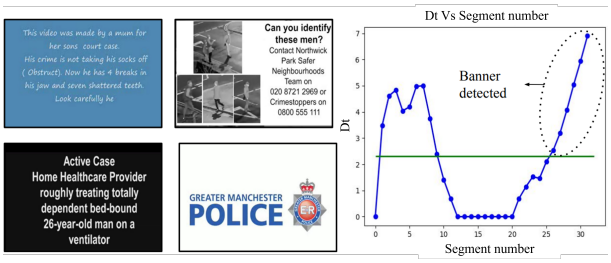

Figure 6. Examples of banners and its effect on D_t.

To mitigate this problem we drop the first and last 20% of total segments from all anomaly videos in the training set for UCF-Crime. Note that this is done during self-supervision in training to identify an initial set of anomalous video frames, hence it does not violate fair comparison with other methods during testing.

For end-to-end training using Uniformer-32 [12], we start with the pretrained model on the Kinetics-600 dataset [4]. It was observed that utilizing all available layers in Uniformer-32 causes the model to overfit rapidly to the

training data, owing to the relatively smaller data sizes of UCF-Crime and XD-Violence as compared to Kinetics-600. In view of this, all but the last block, which comprises 7 attention modules, are frozen in Uniformer-32. The learning rate of $5e{-}5$ is opted for UCF-Crime, and $5e{-}6$ for XD-Violence, with a weight decay of 0.005 for both datasets.

4. Experiments

4.1. Datasets

We evaluate our approach against the state-of-the-art methods using two datasets: UCF-Crime [16], and XD-Violence [20]. We adopt the datasets to fit the evaluation criteria by following the procedures outlined in previous studies [10], [17], [19], [6]. UCF-Crime [16] involves 1900 surveillance videos depicting 13 different types of real-world anomalous events such as abuse, robbery, arson, explosion, and road accidents. The training data consists of 1610 videos with 810 videos labelled as nominal and 800 videos labelled as anomalous. The dataset also includes 290 videos for testing with a mix of nominal and anomalous videos. The XD-Violence [20] dataset is more recent and larger than UCF-Crime. It consists of 4754 untrimmed videos with accompanying audio. The videos cover a diverse range of sources, including surveillance footage, movies, dash-cam recordings, and video games. The training set for the wVAD setting includes 3954 videos with over 1900 anomalous videos and over 2000 nominal videos. To fairly evaluate the performance of our model, we only utilized video information and discard audio information.

4.2. Metrics

Following the common practice in the literature, we use the Area Under ROC Curve (AUC) metric for evaluating our performance on the UCF-Crime dataset and the Average Precision (AP) metric, which is the area under the precision-recall curve, on the XD-Violence dataset. By putting more emphasis on number of anomalies, AP presents a more suitable metric for the inherently class-imbalanced anomaly detection problems than AUC. Hence, it has been preferred over AUC in the literature for the more recent XD-Violence dataset. The results are reported by averaging over four trials. We compute these metrics with window sizes (i.e., number of segments used for inference) ranging from 4 to 256 for three state-of-the-art methods, RTFM [17], S3R [19], MGFN [6], to evaluate the change in their detection performance when they are made to perform in real-time. We also present computational efficiency results for our algorithm in terms of frames per second (fps) to evaluate its real-time inference capability.

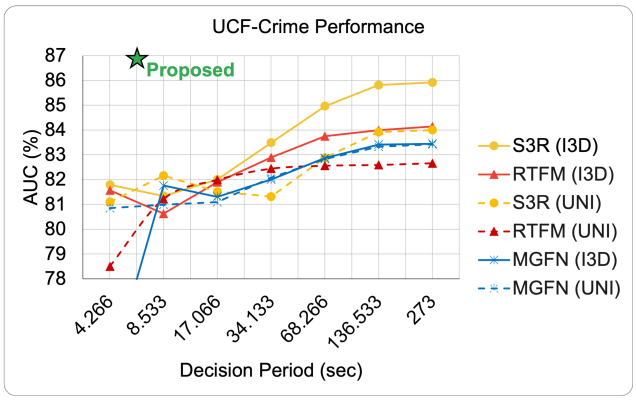

Figure 7. Performance drop of state-of-the-art methods with decreasing decision period on UCF-Crime. Proposed method (REWARD) provides a significantly improved real-time detection performance.

Figure 8. Performance drop of state-of-the-art methods with decreasing decision period on XD-Violence. Proposed method (REWARD) provides a significantly improved real-time detection performance.

Method	feature	AUC %	Decision Period (sec)
S3R [19]	I3D	81.34	8.533
S3R [19]	Uniformer-32	82.16	8.533
RTFM [17]	I3D	80.63	8.533
RTFM [17]	Uniformer-32	81.22	8.533
MGFN [6]	I3D	81.76	8.533
MGFN [6]	Uniformer-32	80.99	8.533
REWARD	Uniformer-32	**86.94**	**6.4**

Table 1. Real-time detection performance of the proposed method is 4.78% higher than the state-of-the-art methods on UCF-Crime.

4.3. Results

Figures 7 and 8 present the performances of S3R, RTFM, and MGFN as a function of decision period using I3D and Uniformer-32 features on the UCF-Crime and XD-violence datasets, respectively. Tables 1 and 2 summarize the real-time detection performance of state-of-the-art methods together with our method on both datasets. As seen in the re-

Method	feature	AP %	Decision Period (sec)
S3R [19]	I3D	70.14	8.533
S3R [19]	Uniformer-32	69.51	8.533
RTFM [17]	I3D	72.6	8.533
RTFM [17]	Uniformer-32	65.35	8.533
MGFN [6]	I3D	73.17	8.533
MGFN [6]	Uniformer-32	67.46	8.533
REWARD	Uniformer-32	**77.71**	**6.4**

Table 2. Real-time detection performance of the proposed method is 4.54% higher than the state-of-the-art methods on XD-Violence.

sults, the proposed approach outperforms the state-of-the-art methods by a wide margin, especially on UCF-Crime, where REWARD with 6.4 sec decision period improves upon S3R with a decision period of 273 sec by more than 1% AUC. Note that for a fair comparison we evaluated the state-of-the-art methods with both I3D and Uniformer feature extractors.

XD-Violence imposes some challenges on the action recognition models. As shown in Figure 9, XD-Violence consists of dynamic camera views with constant scene changes, as opposed to the static camera views in UCF-Crime. Also, the heterogeneity in video sources is much larger in XD-Violence (movies, video games, dash-cam recordings, etc.) than UCF-Crime (only surveillance videos). It should be noted that, the performance of REWARD is highly dependent on the used feature extractor since no feature refinement step is involved in the current version. This aspect will be further analyzed in Section 4.5.

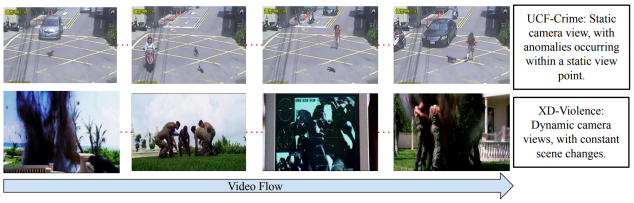

Figure 9. Dynamic camera views in XD-Violence pose challenges for video models compared to the static camera views in UCF-Crime.

Figure 10 shows the final anomaly relevance score of REWARD on sample nominal and anomalous videos from both datasets.

4.4. Computational efficiency

By eliminating the need for feature encoding neural networks, the proposed REWARD method significantly reduces the number of parameters and processing time during inference, leading to better real-time performance compared to the existing methods. Table 3 presents the inference computation rate of 63.3 fps for our method based on Uniformer-32 using a Nvidia RTX-2070 GPU. With 63.3 fps our method is able to process the 32 frames within 0.5 sec, resulting in a total decision delay of 6.9 sec. The kNN computations during training took around 5 hours for Uniformer-32 on an Intel Core i7 8700K CPU for the UCF-Crime dataset.

	Frames processed	Processing time (s)	Fps
Uniformer-32	16.18	1024	63.3
	Decision period (s)	Processing time (s)	Decision delay (s)
REWARD	6.4	0.5	6.9

Table 3. Real-time computation performance of the proposed REWARD method based on Uniformer-32.

4.5. Ablation Study

In this section, we discuss the impact of several factors on the performance of REWARD.

End-to-End Training: To understand the contribution of end-to-end (E2E) training, we compare the wVAD performance of REWARD-E2E with a non-E2E version. Specifically, using the pre-trained features from Uniformer-32 we trained an MLP with the same configuration as the one used in self-supervision using the same labels provided by the self-supervision mechanism, i.e., $\mathcal{V}_a(I_2)$ and $\mathcal{V}_n$ in Figure 5. This transfer learning (TL) approach based on REWARD's self-supervision is called REWARD-TL. On both datasets, end-to-end training improves the performance by a wide margin, 1.64% on UCF-Crime and 2.46% on XD-Violence.

Model	Feature	Dataset	AUC/AP
REWARD-TL	Uniformer-32	UCF-Crime	85.30
REWARD-E2E	Uniformer-32	UCF-Crime	86.94
REWARD-TL	Uniformer-32	XD-Violence	75.25
REWARD-E2E	Uniformer-32	XD-Violence	77.71

Table 4. Impact of end-to-end vs. transfer learning training on wVAD performance.

Offline Performance: Since the existing methods focus on offline detection, we additionally assess the offline anomaly detection performance of our approach by adopting a segmented video analysis strategy. In contrast to a temporal sampling rate of 5 frames per second, we partition both the training and testing videos into 32 segments. Each segment consists of 32 frames, resulting in a consistent total of 1024 frames across videos of varying lengths. The experimental process is reiterated while adhering to the identical implementation details outlined in Section 3.3, with the additional implementation of Savitzky-Golay filter [14] on predictions with a window size of four and a polynomial order of one. Table 5 and Table 6 demonstrate REWARD's offline detection performance on UCF-Crime and XD-Violence datasets, respectively. For long videos, partitioning the video into 32 segments may result in long

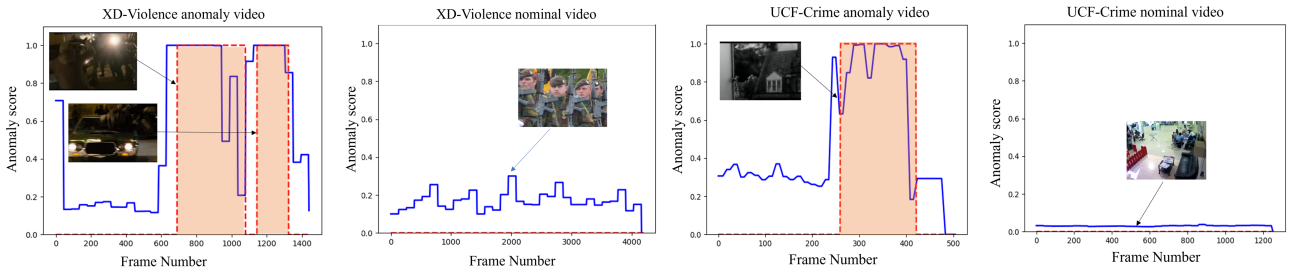

Figure 10. Anomaly score values of end-to-end trained REWARD based on Uniformer-32 on UCF-Crime and XD-Violence test videos. Light orange areas indicate the ground truth anomalous frames.

segment duration (i.e., decision period), which may be prohibitive for online (real-time) detection. Note that longer segments with more contextual information improved REWARD's performance by 0.54% on UCF-Crime and by 2.59% on XD-Violence compared to the online version with a fixed 6.4 sec decision period (cf. Tables 1 and 2).

Method	Feature	Year	AUC(%)
Sultani et al. [16]	I3D	2018	77.92
MIST [10]	I3D	2021	82.30
Wu et al. [20]	I3D	2020	82.44
RTFM [17]	I3D	2021	84.30
S3R [19]	I3D	2022	85.99
MGFN [6]	I3D	2022	83.45
REWARD-E2E	Uniformer-32	2023	**87.48**

Table 5. Offline detection performance comparison between the proposed approach, REWARD-E2E, and the existing methods on the UCF-Crime dataset.

Impact of feature aggregation: Feature aggregation step (e.g., MTN) in existing methods is mainly what prevents high real-time detection performance. Hence, we investigate the impact of feature aggregation on the wVAD performance of existing methods in relation to Uniformer-32 and I3D. Table 7 underscores the substantial deterioration in wVAD performance of existing methods when the critical feature aggregation step is omitted, irrespective of the employed model – I3D or Uniformer-32. Notably, the results suggests a more effective synergy between feature aggregation and I3D.

5. Conclusion

We propose an end-to-end training approach, REWARD (**R**eal-Time **E**nd-to-End **W**eakly Supervised Video **A**nomaly **R**elevance **D**etector), for real-time video anomaly detection in the weakly supervised setting. REWARD trains a large video model for the wVAD task, as opposed to the existing methods based on refining the feature extractors pretrained on action recognition datasets. The novel technique enabling end-to-end training is a self-supervision method based on kNN distance calculations. The experimental results demonstrated that in a delay sensitive setting where real-time decision is important, the proposed end-to-end solution with a decision perior of 6.4 sec outperforms the state-of-the-art methods with a similar decision period of 8.5 sec by 4.78% and 4.54% AUC on the commonly used UCF-Crime and XD-Violence datasets. Moreover, REWARD provides a more computationally efficient pipeline for real-time inference by eliminating the popular feature aggregation/refinement step in the literature.

Method	Feature	Year	AP(%)
Wu et al. [20]	I3D	2020	75.41
RTFM [17]	I3D	2021	77.81
S3R [19]	I3D	2022	80.26
MGFN [6]	I3D	2022	80.1
REWARD-E2E	Uniformer-32	2023	**80.30**

Table 6. Offline detection performance comparison between the proposed approach, REWARD-E2E, and the existing methods on the XD-Violence dataset.

Model	Feature	Aggregation	AUC/AP
S3R	I3D	✗	81.3
S3R	I3D	✓	85.99
S3R	Uniformer-32	✗	80.1
S3R	Uniformer-32	✓	84
RTFM	I3D	✗	81.94
RTFM	I3D	✓	84.30
RTFM	Uniformer-32	✗	79.34
RTFM	Uniformer-32	✓	82.75

Table 7. Impact of feature aggregation on wVAD performance.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[4] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[6] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. *arXiv preprint arXiv:2211.15098*, 2022.

[7] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

[8] Keval Doshi and Yasin Yilmaz. Continual learning for anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 254–255, 2020.

[9] Keval Doshi and Yasin Yilmaz. Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition*, 114:107865, 2021.

[10] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14009–14018, 2021.

[11] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pages 84–92. Springer, 2015.

[12] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*, 2022.

[13] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[14] Jianwen Luo, Kui Ying, and Jing Bai. Savitzky–golay smoothing and differentiation filter for even number data. *Signal processing*, 85(7):1429–1434, 2005.

[15] Bharathkumar Ramachandra, Michael J Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2293–2312, 2020.

[16] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.

[17] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4975–4986, 2021.

[18] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[19] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 729–745. Springer, 2022.

[20] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer, 2020.