Robust Anomaly Detection through Transformer-Encoded Feature Diversity Learning

Kuldeep Biradar[1], Dinesh Kumar Tyagi[1], Ramesh Babu Battula[1], and P Subbarao[2]

[1] Malaviya National Institute of Technology Jaipur, Jaipur, India-302017
[2] Vignan University, India
{2018rcp9503,dktyagi.cse,rbbattula.cse}@mnit.ac.in, drpsr_it@vignan.ac.in

Abstract. Detecting irregularities in weather data serves multiple practical applications. For example, nowcasting focuses on predicting atmospheric conditions for the next 0 to 4 hours, which is essential for effective emergency response and disaster management. Anomaly detection also plays a crucial role in forecasting extreme weather events. However, the complexity increases when considering both anomalies and varying weather conditions. A common approach to anomaly detection is weakly supervised video-level labeling, which aims to identify frames containing abnormal events and is typically framed as a multiple instance learning (MIL) problem. While existing methods perform well, the prevalence of negative instances significantly hinders their ability to detect positive instances, particularly rare abnormal segments. To address this, we aim to extract distinctive features by enhancing the observable differences between various classes using a single branch. We propose a novel method, Transformer Encoded Feature Video Anomaly Detection (TEF-VAD), which exclusively utilizes attention mechanisms, specifically Multi-Head Attention Learning. This approach combines feature magnitude learning loss, class-specific loss, and a TEF-VAD-enhanced MIL classifier training loss, thereby training a model to effectively identify positive examples and improve the MIL method's robustness for detecting positive instances in abnormal videos. Our extensive experiments demonstrate that the MIL model enhanced by our Transformer method significantly improves sample efficiency and the detection of subtle anomalies, outperforming several state-of-the-art techniques on benchmark datasets like UCF-Crime and ShanghaiTech.

Keywords: Anomaly Detection · Transformer · UCF Crime.

1 Introduction

Identifying irregularities in weather data serves various practical purposes. For instance, nowcasting aims to forecast atmospheric conditions within the next 0 to 4 hours, which is vital for emergency response and disaster mitigation. Similarly, anomaly detection is instrumental in predicting and analyzing extreme

weather phenomena over time. Integrating weather factors with applications such as anomaly detection [1], object detection [2], and segmentation poses significant challenges for researchers. However, when we take into account both anomalies and varying weather conditions, the situation becomes quite complex. To address this, we have various types of datasets, such as UCF-Crime and XD, which encompass multiple weather scenarios along with the associated anomalies. Given that abnormal events are infrequent in videos, recent research has predominantly been conducted within the weakly supervised learning framework [1,3–9], which involves solely video-level annotations. Video anomaly detection aims to predict frame-level anomaly scores at which an abnormal event takes place. Abuse, stealing, aggression and other similar behaviors are examples of anomalies in the context of surveillance.

Despite years of research in video anomaly detection (VAD) with different types of weather, creating a model capable of accurately detecting anomalies in videos remains a challenging task. This difficulty stems from the requirement for the model to comprehend the inherent distinctions between normal and abnormal events, particularly rare and widely varying anomalous events. Earlier studies have addressed VAD as an unsupervised learning task [10–14]. One commonly utilized approach is the concept of reconstructing features from normal training data. Based on the features employed, all available methods typically involve training an autoencoder framework or generative reconstruction-based methods (GAN) with a deep neural network. These methods aim to ensure the reconstruction of normal frame events with minimal reconstruction mistake. Nevertheless, the likelihood of failure increases significantly under diverse weather conditions. For instance, if a model is trained exclusively on data from normal weather conditions and is then tested in hazy or rainy conditions, there is a considerable risk of encountering high reconstruction errors, leading to inaccurate results. Consequently, it becomes evident that nearly all methods based on training set reconstruction of frames cannot ensure the detection of unusual occurrences.

Here, we tackle the subject of weakly supervised video anomaly detection (WS-VAD), where acquiring video-level labels are often more practical and can yield more robust outcomes compared to unsupervised techniques. By giving different types of data at the time of training and equalizing the amount of abnormal-normal snippets evenly across the training set, WSVAD techniques trained using multiple-instance learning (MIL) algorithms have recently addressed the aforementioned issues [1,3–5]. The normal videos are used to select the normal snippets at random, while the abnormal videos are used to select the snippets with the greatest anomaly scores.

MIL partially resolves the previously mentioned issues but introduces following problems: Selecting normal snippets at random from normal videos might lead to relatively simple modeling, potentially hindering convergence of training; if a video contains multiple abnormal snippets, the opportunity for a more effective training session that includes multiple abnormal snippets per video is lost; relying on classification scores provides a feeble training signal that may

not facilitate a clear distinction among normal and abnormal snippets. Model trained with single weather with same anomaly. All of these problems are made worse by techniques that ignore important temporal dependencies.

In order to tackle the MIL challenges mentioned earlier, we introduce a new method called Transformer-Encoded Feature-Based Video Anomaly Detection(TEF-VAD). In this approach, we utilize features encoded using a transformer and select the top-k snippets inspired by RTFM [6] of anomaly to train MIL which works on different types of weather it may be rain, fogg, day time, or it may be night time etc. Also we learned class-specific loss to ensure that the features of each class are patterned in a similar manner. Using k instances of the abnormal and normal films with the greatest classification scores, the MIL method trains a classifier. Proposed method addresses the challenges associated with MIL in the following ways: Robust Features encoded by using a transformer encoded feature; it increases the likelihood of selecting truly anomalous frames from abnormal videos; by choosing hard negative normal snippets from normal videos, which are more difficult to model, it promotes better training convergence; it allows the inclusion of a greater number of anomalous frames for each abnormal video in the training process.

We assess the efficacy of proposed TEF-VAD method on two benchmark anomaly detection datasets, ShanghaiTech [11] and UCF-Crime [1]. Our results demonstrate the significant outperformance of our method compared to the current state-of-the-art techniques across all benchmarks. We also establish that our method achieves markedly improved sample efficiency and subtle anomaly discriminability compared to widely used MIL methods. Furthermore, our proposed method showcases effective anomaly snippet detection as indicated by the AUC metric.

Our contributions can be outlined as:

- The Transformer encoded feature VAD (TEF-VAD) architecture for anomaly detection, which aligns with the concept of predicting frame-level anomaly scores during anomalous events, utilizing video-level annotation.
- We achieve the extraction of robust features using a transformer, facilitated by a multi-head attention mechanism.
- The learning of TEF-VAD involves the utilization of three different types of losses: the Transformer encoded feature magnitude learning loss; class-specific loss; and TEF-VAD-enabled MIL classifier training loss.
- The robustness of our method is demonstrated through experiments conducted on two benchmark datasets which improves 2.21% in ShanghaiTech and 1.44% in UCF-Crime anomaly dataset.

2 Related Work

The video anomaly detection research landscape can be categorized into two main classes: unsupervised and weakly-supervised VAD.

Unsupervised VAD: Techniques that only use unlabeled training data or that carry out direct training and testing on testing data are referred to be

unsupervised methods. A method to identify changes in a video sequence by identifying unique frames was proposed by Del et al. [15], while Tudor et al. [16] introduced unmasking technology [17] to train a binary classifier iteratively in order to identify the most discriminant features. A recent method by Zaheer et al. [18] establishes cross-supervision between a generator and a discriminator to take advantage of the low frequency of anomalies. Moreover, One-Class Classification (OCC) approaches treat the problem in an unsupervised way and assume the availability of just normal training data. Usually, when building a model, researchers use only normal data and identify events that deviate from the model; this allows them to identify anomalies. Previous works made use of hand-crafted look and motion aspects [19–23]. Recent techniques are using features from pre-trained deep neural networks to construct an anomaly classifier, thanks to developments in deep learning [10, 24]. Moreover, self-supervised feature learning techniques exist [**?**, 25–27]. One well-liked strategy uses temporal prediction [11, 28, 29]. Unsupervised approaches, however, can raise false alarms for normal patterns that are not seen because it is not feasible to include every possible kind of normalcy in a single dataset.

Weakly supervised VAD: Video-level labels are more useful in weakly supervised learning approaches [1, 3–5, 30] for distinguishing between abnormal and normal events. There are two types of existing weakly supervised VAD approaches: encoder-independent methods and encoding-based methods. **Encoder-independent** techniques only help the classifier get trained. For example, a deep multiple instance learning model that treats a video as a "snippet" and its numerous parts as discrete "instances" was first developed by Sultani et al. [1]. [31] proposed supervised anomaly detection. Wan et al. [4] presented dynamic MIL loss and regularization led by the center, while Zhang et al. [5] introduced inner-bag score gap regularization. Both a feature encoder and a classifier are trained using encoding-based approaches. To encode motion-aware information, Zhu et al. [30] presented an MIL model with attention, integrated with an autoencoder based on optical flow. Weakly supervised VAD was approached as a label noise learning task by Zhong et al. [3], who also used graph convolutional networks (GCNs) to filter label noise for iterative model training. Nevertheless, the method proved to be inefficient and progressed slowly.

Multiple Instance Learning Weakly supervised learning is frequently accomplished through the use of Multiple Instance Learning (MIL). MIL handles a video as a "bag" and its component clips as distinct "instances" in video-related applications [1, 32, 33]. Indirect supervision for instance-level learning can be obtained from video-level labels by using a certain feature/score aggregation function. Many aggregation functions are used, including attention pooling [32, 33] and max pooling [1, 5, 30].

We also use an encoder-based methodology and an online, fine-grained approach in our work. On the other hand, we encode features using a transformer, and then use the top-k snippets for Multiple Instance Learning (MIL) to optimize our solution.

3 Proposed Method

In the context of Weakly Supervised Video Anomaly Detection (WSVAD), the training data for detecting anomaly events at the frame-level typically consists of normal video V_n and abnormal video V_a, each having a label at the video level $Y = [0(Normal), 1(Anomaly)]$. During training, each input video, whether normal or abnormal is divided into T segments, which are then placed in the negative bag (Normal) and positive bag (Abnormal), as illustrated in Fig. 1. Additionally, these snippets provide features of dimension D, obtained from the pre-trained I3D model [34], denoted as f_m^D. Here, f represents the feature of D dimensions, and m equals 1 if the snippet is abnormal and 0 if the snippet is normal. The proposed architecture utilizes encoded features of the transformer, denoted by $E_\alpha(f_m^D)$ where α is a learning parameter denoted in fig 1. Further the model employed by Transformer Encoded Feature VAD (TEF-VAD) is denoted as $F_\beta(E_\alpha(f_m^D))$, and results in a 2-dimensional feature [0 (Normal), 1 (Anomaly)], indicating the classification of the T video segments as normal or anomalous, with the learning parameters α and β specified below which visualize in fig 2. The learning of this model encompasses a collective optimization involving end-to-end Transformer encoded feature magnitude learning, class-specific feature learning, and TEF-VAD-enabled MIL classifier training, utilizing the specified loss represented in equation 1.

$$\min_{\alpha,\beta} \sum_{m,n=1}^{D} l_t(E_\alpha(f_m), E_\alpha(f_n), y_m, y_n) + l_m il(F_\beta(E_\alpha(f_m)), y_m) + l_{cs}(E_\alpha(f_m), y_m)) \tag{1}$$

Where l_t represents Transformer encoded feature magnitude learning loss, l_{mil} represents TEF-VAD-enabled MIL classifier training loss and l_{cs} represents class specific loss

Next, we will elaborate on the rationale behind our suggested TEF-VAD, accompanied by a comprehensive description of the units.

3.1 Transformer Encoder Feature

In standard NLP models, the input sequence is typically processed sequentially, either word by word or character by character. In contrast, a transformer model processes the input sequence in parallel, enabling the model to analyze the entire input sequence at once [35]. A common formulation of the weakly supervised video-level label anomaly detection method is a multiple instance learning (MIL) issue. Finding video segments featuring anomalous incidents is the primary objective. In our scenario, instead of processing individual characters, we pass temporal features, and leverage the capabilities of the transformer to extract robust features. A series of feedforward neural network layers and multi-head self-attention layers make up the transformer model. In the multi-head self-attention mechanism, the input sequence undergoes processing through three linear layers

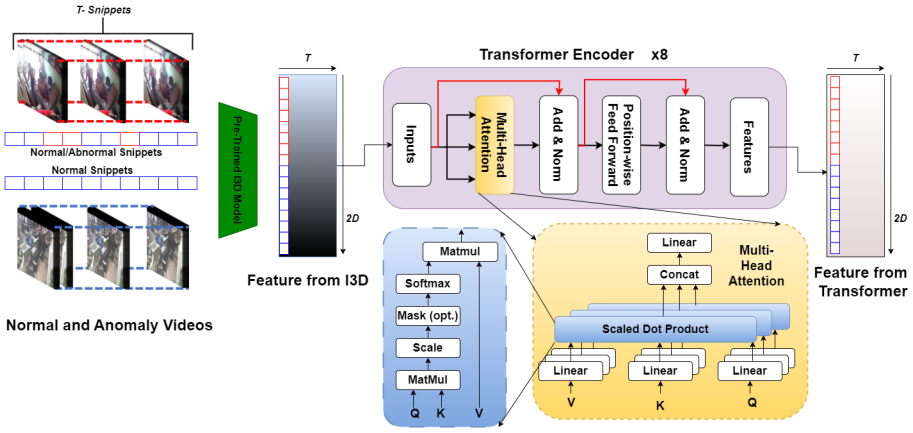

Fig. 1. Videos with weak labels are divided into T snippets and then fed into the Pre-Trained Model for Encoding 2D-Dimentions Features(Anomaly(D)-Normal(D)) using Transformer.

to produce three sets of projections: values(V), queries(Q) and keys(K). The dot product of the query and key projections is calculated and then scaled by the square root of the dimension of the key vectors. Subsequently, a softmax function is employed to generate a set of attention weights for each token in the sequence. These attention weights are utilized to compute a weighted sum of the value vectors, resulting in a context vector for each token in the sequence shown in Fig-1.

3.2 Transformer encoded feature magnitude learning and Class-Specific Learning

Transformer encoded features were further harnessed for VAD using MIL, as depicted in Fig. 2. In this framework, we calculated Transformer encoded feature magnitude learning and Class-Specific Learning losses, illustrated using Equation 2 and Equation 3 respectively. In this method, the largest snippet feature magnitudes from normal videos are reduced, and those from abnormal videos are increased (i.e.using top k samples and top k batch samples), inspired by [6, 36]. The class-specific loss l_{cs} is represented in equation to ensure that the characteristics of each class are patterned in a similar way.

Building on our understanding of statistical normality modeling in BatchNorm within VAD, the mean vector μ serves as a statistical indicator of normality, allowing us to differentiate between normal snippets and potential anomalies.

To promote divergence between potential abnormal features and the mean vector $\hat{\mu}$, while simultaneously clustering the normal features, Specifically, following our l_t criterion, we select K potential abnormal features magnitude from abnormal videos and K normal features magnitude from normal videos. further

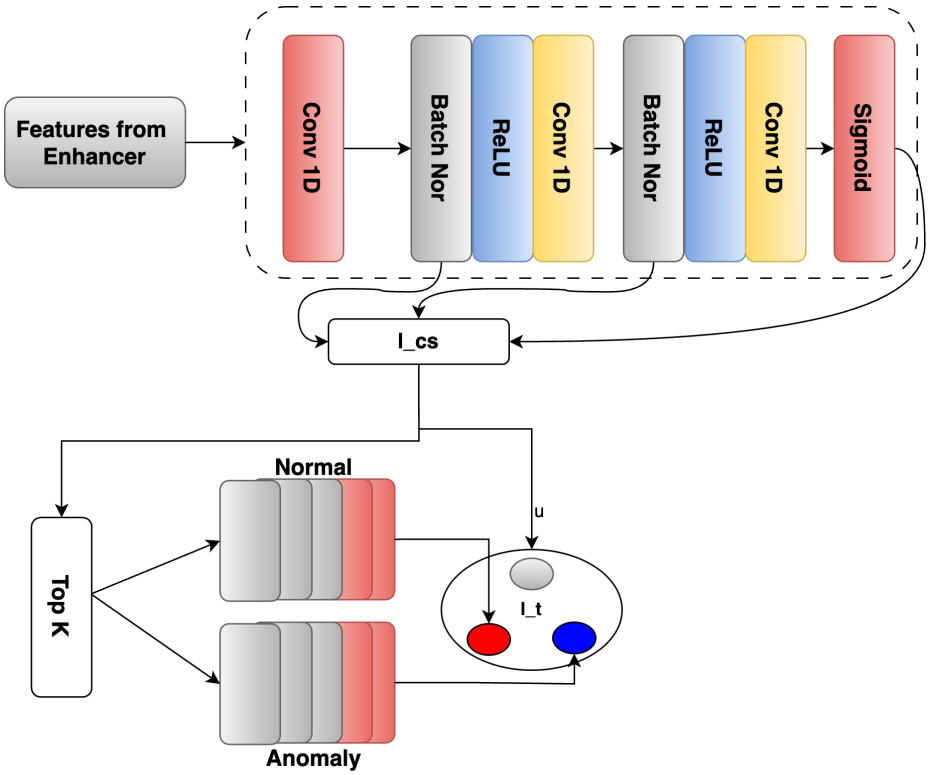

Fig. 2. Learning of proposed Transformer based encoded feature MIL using different type of losses(l_t, l_{cs}).

those substracted with normalised mean $\hat{\mu}$ that exhibit the highest K l_t scores shown in equation 2

$$l_t(E_\alpha(f_a), E_\alpha(f_n), y_m, y_n) = {}^{max(0, m - d_{\alpha,k(X_n-\hat{\mu}, X_a-\hat{\mu})})}$$

$$0 \quad {}_{otherwise} \tag{2}$$

As we have a feature vectors of $E_\alpha(f_n)_k$ and $E_\alpha(f_a)_k$ for normal and anomaly respectively. We took magnitude of

$$l_{cs}(E_\alpha(f_m), y_m)) = L_1(E_\alpha(f_n)_k, y_n) + L_1(E_\alpha(f_a)_k, y_a) \tag{3}$$

where L_1 represents l1-loss , $E_\alpha(f_n)_k$ represents normal top k snipet features from transformer similarly $E_\alpha(f_a)_k$ for abnormal.

In order to train the segment classifier (MIL), We employ a binary cross-entropy classification loss function with a specified set of top k snippets.

4 Datasets and Experiment setup

We assess our model across two varied benchmark datasets, established for anomaly detection in videos labeling: ShanghaiTech, UCF-Crime

4.1 Dataset

The **UCF-Crime** dataset [1] is a collection of 1900 uncut films for anomaly detection, amounting to 128 hours in total. These videos are taken from actual security cameras on the streets and inside buildings presenting dynamic and varied backgrounds as opposed to the still backgrounds observed in the samples present ShanghaiTech dataset. There are an equal number of normal and anomalous videos in the training and testing sets. Together with labels at the video level, the dataset includes 1,610 training videos covering 13 anomaly classes, and it also incorporates 290 test clips with labels at the frame level.

ShanghaiTech is relatively a medium sized dataset with 437 videos altogether and 13 distinct background scenes that were obtained from street video footage captured from fixed angles. These, 307 videos are categorized as normal, while 130 are classified as anomaly videos. The dataset cited in [11] has been commonly adopted as a benchmark for detecting anomolies. Zhong et al. [3] modified the dataset via utilizing the inclusion of anomalous inferencing videos within the training data, generating a weakly supervised training dataset. This assures that the train and test samples contain representations of all 13 background scenarios. We have adopted the same methodology as described in [3] to adapt ShanghaiTech for the weakly supervised scenario.

The proposed anomaly detection method is structured in two distinct phases: feature extraction and classification. For feature extraction, we employ the I3D (Inflated 3D ConvNet), which excels at capturing spatiotemporal dynamics from video data. However, it is important to note that this multi-step approach may limit its effectiveness for real-time anomaly detection due to inherent processing latencies. To validate our method, we have utilized two widely recognized datasets: UCF-Crime and ShanghaiTech. UCF-Crime, in particular, provides diverse real-time data drawn from CCTV footage, social media, and YouTube, encompassing a variety of anomaly types. While these datasets are popular in the field, we recognize the necessity for broader experimentation. Future work will involve expanding our analysis to additional datasets to strengthen the generalizability of our findings across various anomaly detection contexts. Although our results demonstrate superior performance within these specific datasets.

4.2 Assessment Metrics

Following previous research, we select the evaluation metric to measure the effectiveness of our proposed technique on the UCF-Crime and ShanghaiTech datasets: the area under the curve (AUC) of the receiver operating characteristic (ROC) curve at the frame level.

4.3 Implementation Details

According to [1], the features of the I3D pre-trained network [34] were taken from the 'mix 5c' layers correspond to 2048D. Following feature extraction, there are 32 film clips in each video denoted as $T = 32$. Throughout all investigations, the margin remains fixed at $m = 100$, while $k = 5$ in equation (2). The model delineated in Section 3 fig 2 includes three dense layers, having 512, 128, and 1 nodes. Each node succeeded by a dropout function and a ReLU activation function having the dropout rate of 0.3. Before fully connected we used transformer encoded layer to extract features from I3D features. We use an exhaustive training strategy with our TEF-VAD technique, employing the optimizer Adam [37] with a batch size of 32 over 50 epochs and a weight decay of 0.0001. For the ShanghaiTech and UCF-Crime datasets, a learning rate of 0.001 is set. Samples chosen at random from 32 normal and anomalous videos make up each mini-batch. The method's implementation is completed in PyTorch [38]. We use the same standard setup as in [1, 3–5, 30] to ensure fair comparison, and we use the published results with the same foundation as ours for all baselines.

Table 1. Comparative analysis of frame-level AUC performance with other state-of-the-art (SOTA) methods on the ShanghaiTech dataset. AUC_o represent the AUC calculated on the complete test set.

[t]

Category	Methods	Features	$AUC_o(\%)$
Unsupervised VAD	Conv-AE [10]	-	60.85
	Frame-Pred [11]	-	73.40
	Stacked-RNN [12]	-	68.00
	VEC [39]	-	74.80
	Mem-AE [40]	-	71.20
	MNAD [41]	-	70.50
Weakly-Supervised VAD	GCN-Anomaly [3]	TSN-Flow	84.13
	GCN-Anomaly [3]	C3D-RGB	76.44
	GCN-Anomaly [3]	TSN-RGB	84.44
	Zhang et al. [5]	I3D-RGB	82.50
	Sultani et al. [1]	I3D-RGB	85.33
	AR-Net [4]	I3D-Flow	82.32
	AR-Net [4]	I3D-RGB & I3D-Flow	91.24
	AR-Net [4]	I3D-RGB	85.38
	Proposed	I3D-RGB	93.45

4.4 Results

Table 1 provides a breakdown of the AUC results at the frame level for the ShanghaiTech dataset. Our proposed method demonstrates superior performance as measured against earlier state-of-the-art (SOTA) unsupervised learning techniques [10–12, 39, 41], as well as weakly supervised approaches [1, 4, 5]. Using

Table 2. Comparative analysis of frame-level AUC performance with other state-of-the-art (SOTA) methods on the UCF-Crime dataset. AUC_o and AUC_a represent the AUC calculated on the complete test set and solely on abnormal test videos, respectively

Category	Methods	Features	AUC_o(%)	AUC_a(%)
Unsupervised VAD	SVM Baseline		50	50
	Sohrab et al. [42]	-	58.50	-
	Conv-AE [10]	-	50.60	-
	Lu et al [21]	-	65.51	-
	GODS [43]	-	70.46	-
	BODS [43]	-	68.26	-
Weakly-Supervised VAD	Zhang et al. [5]	I3D-RGB	78.66	-
	Sultani et al. [1]	C3D-RGB	75.41	54.25
	Zhang et al. [5]	I3D-RGB	78.66	-
	Motion-Aware [30]	TSN-Flow	79.10	62.18
	GCN-Anomaly [3]	TSN-RGB	82.12	59.02
	Wu et al. [44]	I3D-RGB	82.44	-
	WSAL [7]	I3D-Flow	85.38	67.38
	RTFM [6]	I3D-RGB	84.30	-
	UMIL [8]	I3D-Flow	86.75	68.68
	Proposed	I3D-RGB	83.71	70.12

I3D-RGB features, proposed TEFVAD model achieves the greatest AUC score on this dataset, reaching 93.45%. Furthermore, with the same I3D-RGB features, our Transformer-encoded feature learning MIL approach surpasses existing SOTA MIL-based methods [56, 62, 74] by a margin ranging from 2% to 9%. These results highlight the advancements facilitated by our proposed transformer-encoded feature learning technique.

Table 2 compares our Proposed Method (TEFVAD) with other well-known state-of-the-art (SOTA) techniques, including Unsupervised VAD (UVAD) and WSVAD. When compared to all other approaches, the Proposed Method earns the greatest AUC_A on the UCF-Crime dataset., marking an enhancement of +1.44%. Similarly, for AUC_O, our method demonstrates improvements relative to all competing methods except for those specified (RTFM [6], WSAL [7], UMIL [8]).

In particular, our improvement in the AUC_A metric demonstrates the exceptional performance of the Proposed TEFVAD Method in AUC_O is not only attributed to the existence of clear-cut normal videos, but also arises from its enhanced capability to detect anomalous segments within abnormal videos. Furthermore, across both datasets, Weakly Supervised VAD exhibits substantial enhancements over Unsupervised VAD in terms of AUC_O, empirically confirming the inherent challenges in detecting open-set anomalies in UVAD. Conversely, the improvements in AUC_A are relatively minor, for instance, 54.25% in comparison to 50.00% on UCF-Crime. This highlights the existing bias in current WSVAD methods towards the explicit differentiation between normal and ab-

normal, leading to a high incidence of false positives and negatives in uncertain segments from abnormal videos.

5 Conclusion

In this research, we presented a Multiple Instance Learning (MIL) approach that utilizes Transformer-encoded features to develop an impartial anomaly classifier and a customized representation for WSVAD. This MIL training scheme necessitates a confident set comprising visually apparent normal/abnormal video snippets, derived from Transformer-encoded features with a substantial margin in positive and negative snippets. The application of MIL leads to notable enhancements in classification performance. The combination of these distinctive features and classifier in our approach yields a substantial An enhancement over present state-of-the-art techniques. Our approach is empirically substantiated by achieving top-tier performance on two benchmark datasets for WSVAD.

References

1. 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. 2, 3, 4, 8, 9, 10
2. Prafulla Saxena, Kuldeep Biradar, Dinesh Kumar Tyagi, and Santosh Kumar Vipparthi. Richex: A robust inter-frame change exposure for segmenting moving objects. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2172–2176. IEEE, 2022. 2
3. Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1237–1246, 2019. 2, 4, 8, 9, 10
4. Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2020. 2, 4, 9
5. Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4030–4034. IEEE, 2019. 2, 4, 9, 10
6. 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. 2, 3, 6, 10
7. Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE transactions on image processing*, 30:4505–4515, 2021. 2, 10
8. Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8022–8031, 2023. 2, 10

9. Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6074–6082, 2024. 2
10. Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016. 2, 4, 9, 10
11. Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018. 2, 3, 4, 8, 9
12. Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, pages 341–349, 2017. 2, 9
13. Sachin Dube, Kuldeep Biradar, Santosh Kumar Vipparthi, and Dinesh Kumar Tyagi. Mag-net: A memory augmented generative framework for video anomaly detection using extrapolation. In *International Conference on Computer Vision and Image Processing*, pages 426–437. Springer, 2021. 2
14. Kuldeep Marotirao Biradar, Ayushi Gupta, Murari Mandal, and Santosh Kumar Vipparthi. Challenges in time-stamp aware anomaly detection in traffic videos. *arXiv preprint arXiv:1906.04574*, 2019. 2
15. Allison Del Giorno, J Andrew Bagnell, and Martial Hebert. A discriminative framework for anomaly detection in large videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 334–349. Springer, 2016. 4
16. Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu. Unmasking the abnormal events in video. In *Proceedings of the IEEE international conference on computer vision*, pages 2895–2903, 2017. 4
17. Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8(6), 2007. 4
18. M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14744–14754, 2022. 4
19. Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, 30(3):555–560, 2008. 4
20. Borislav Antić and Björn Ommer. Video parsing for abnormality detection. In *2011 International conference on computer vision*, pages 2415–2422. IEEE, 2011. 4
21. Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 4, 10
22. Shu Wang and Zhenjiang Miao. Anomaly detection in crowd scene. In *IEEE 10th International Conference on Signal Processing Proceedings*, pages 1220–1223. IEEE, 2010. 4
23. Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE conference on computer vision and pattern recognition*, pages 935–942. IEEE, 2009. 4

24. Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, Enver Sangineto, and Nicu Sebe. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1689–1698. IEEE, 2018. 4
25. Mohammad Sabokrou, Mahmood Fathy, Mojtaba Hoseini, and Reinhard Klette. Real-time anomaly detection and localization in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 56–62, 2015. 4
26. Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015. 4
27. Kuldeep Marotirao Biradar, Murari Mandal, Sachin Dube, Santosh Kumar Vipparthi, and Dinesh Kumar Tyagi. Triplet-set feature proximity learning for video anomaly detection. *Image and Vision Computing*, 150:105205, 2024. 4
28. Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15425–15434, 2021. 4
29. Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 4
30. Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211*, 2019. 4, 9, 10
31. Kuldeep Biradar, Sachin Dube, and Santosh Kumar Vipparthi. Dearest: deep convolutional aberrant behavior detection in real-world scenarios. In *2018 IEEE 13th international conference on industrial and information systems (ICIIS)*, pages 163–167. IEEE, 2018. 4
32. Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6752–6761, 2018. 4
33. Fa-Ting Hong, Xuanteng Huang, Wei-Hong Li, and Wei-Shi Zheng. Mini-net: Multiple instance ranking network for video highlight detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 345–360. Springer, 2020. 4
34. Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5, 9
35. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5
36. Yixuan Zhou, Yi Qu, Xing Xu, Fumin Shen, Jingkuan Song, and Heng Tao Shen. Batchnorm-based weakly supervised video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 6
37. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 9
38. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 9

39. Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM international conference on multimedia*, pages 583–591, 2020. 9
40. Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1705–1714, 2019. 9
41. Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14372–14381, 2020. 9
42. Fahad Sohrab, Jenni Raitoharju, Moncef Gabbouj, and Alexandros Iosifidis. Subspace support vector data description. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 722–727. IEEE, 2018. 10
43. Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8201–8211, 2019. 10
44. 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. 10