VADOR: Real World Video Anomaly Detection with Object Relations and Action

Halil İbrahim Öztürk
n17238039@cs.hacettepe.edu.tr
Ahmet Burak Can
abc@hacettepe.edu.tr

Computer Science
Hacettepe University
Ankara, TR

Abstract

Anomaly detection in real world videos requires complex scene understanding. Previous works utilize action recognition models as feature extractor, but some anomalies (e.g. robbery) can not be easily understood from basic action information. Our VADOR model leverages action and relationships of objects in the scene to detect anomaly using transformer encoders. Cross-attention between object relation encoder and action encoder helps to fusion of information. Our Anchor based Temporal Action Localization network (TALNet) segments anomalies temporarily by using clip features generated from the encoders. We train VADOR with strong regularization and data augmentation methods. VADOR achieves %83.61 AUC score while achieving %63.09 F1@25 score at temporal segmentation on UCF Crime dataset. Code is publicly available at https://github.com/hibrahimozturk/vador .

1 Introduction

Video anomaly detection from surveillance cameras is important for cities to enhance safety and security. With the ability to detect and alert authorities to criminal activities such as fights, shootings, and robberies, video anomaly detection can help to respond quickly and prevent crimes from occurring. In addition, it can be used to monitor traffic in cities, providing early detection of accidents and incidents, thereby improving emergency response times.

Anomaly detection in surveillance videos has been attempted to be solved by two main approaches. The first approach involves learning normal behaviors from normal videos and then detecting abnormal cases as outliers. The outliers are detected by utilizing the reconstruction error between frames of the input video and reconstructed frames. However, this approach requires observation of a large portion of normal behavior, which may not always be possible. As a result, some normal actions may be mistakenly identified as abnormal. Pre-deep learning study [11] uses dictionary learned from normal videos to reconstruct given video frame. Deep learning based studies [7, 10, 12] proposes encoder-decoder models to detect anomalies. [14] learns normal scenes with Generative Adversarial Network (GAN) [4], learned discriminator is used to detected anomaly.

The second approach involves learning both normal and abnormal behaviors from a training dataset to predict anomalies in videos. First large scale dataset which contains normal and

abnormal videos is UCF Crime [15] dataset. [15, 24] studies use Multiple Instance Learning (MIL) to teach network detecting abnormal video clips. The networks use spatio-temporal features extracted from pre-trained networks like as C3D [18] and I3D [1]. [24] integrates learned motion-aware features as complementary to spatio-temporal features. [22, 23] formulates the problem as supervised learning with noise labels, while [23] employs GCN to decrease noise, [22] uses binary clustering with cluster distance loss to remove noisy labels.

RTFM [17] improves multiple instance learning (MIL) methods for video anomaly detection by training a feature magnitude learning function that leverages temporal feature magnitudes of video snippets. This approach enforces margins between abnormal and normal snippets, resulting in enhanced anomaly detection. In contrast, S3R [20] addresses video anomaly detection by formulating it as an out-of-distribution problem and utilizing self-supervised sparse representation. The S3R combines dictionary-based representation and self-supervised techniques, and incorporates MIL to effectively tackle unsupervised and weakly-supervised video anomaly detection tasks. Additionally, MGFN [2] focuses on capturing long-term context and detecting scene-adaptive anomalies by integrating global-to-local information, allowing for an initial overview of the video before focusing on specific portions for anomaly detection.

Identifying anomalies in real-world scenarios is a challenging task that cannot solely rely on action-based knowledge. For instance, distinguishing between a motorcycle robbery and the rightful owner operating their motorcycle poses a significant challenge. To effectively recognize such complex actions, it becomes crucial to consider the objects involved and their interrelationships within the contextual scenes. Relying solely on static images or features from video clips (e.g. [1]) is insufficient to detect anomalies within video datasets, as dynamic information and temporal context play essential roles in the anomaly detection process.

In response to these challenges, we propose VADOR, a method understands complex scenes through the integration of action information and object relations. VADOR leverages two separate encoders: the object relation encoder and the action encoder. The object relation encoder [8] uncovers relationships between objects in the video clips based on their positions and features extracted from a pre-trained object detection model. On the other hand, the action encoder processes the action features of the video clips extracted from a pre-trained I3D model [1]. VADOR fuses action and object relation features using cross-attention layers positioned after the mid-level of the encoders (Figure 1). This approach enables VADOR to detect complex anomalies that are challenging to identify based on actions alone, thereby enhancing the accuracy and effectiveness of anomaly detection in untrimmed videos.

Rather than focusing solely on frame-level anomaly detection performance, our objective is to enhance the temporal localization performance of anomaly segments in a timeline. The final component of our method, VADOR, is a temporal action localization network (TALNet) that utilizes extracted clip features to accurately segment anomalies over time. TALNet is composed of temporal convolutions that operate on the sequence of clip features. To further improve the segmentation performance, we incorporate the Bi-directional Feature Pyramid Network (BiFPN) introduced in [16]. TALNet consists of consecutive BiFPN blocks, and the output of the last block is fed to classification and regression heads.

For evaluating our method, we use the UCF Crime dataset [15], the most widely used real-world anomaly dataset. Our model demonstrates superior performance in terms of F1 score for temporal segmentation, while also achieving a frame-level AUC score comparable to SOTA techniques. VADOR achieves an impressive F1@25 score of 63.03 on the UCF Crime dataset, surpassing the closest competitor ADNet which is temporal anomaly

detection model with a score of 51.85 F1@25. Since there is currently no other dataset with time-labeled annotations apart from UCF Crime, we assessed the performance of VADOR on the XD-Violance dataset, which was trained using UCF Crime. To ensure fair comparisons, we utilized UCF Crime-trained models RTFM, S3R and our TALNet(without encoders) for evaluating on XD-Violance. The results clearly indicate that VADOR exhibits improved generalization capability. Furthermore, we conducted an investigation to analyze the impact of cross-attention within the VADOR model.

The present paper makes the following contributions:

- VADOR is a novel method for detecting anomalies in videos by integrating action and object relations. Cross attention layers are employed between action and object relation encoders to enable information fusion.
- Our proposed TALNet adopts anchor-based temporal action localization and employs multi-stage Bi-directional Feature Pyramid Network (BiFPN) blocks. In particular, a mid-level BiFPN block is designed to generate useful features for predicting anomalies in the anchor area, by incorporating a classification loss in the block.

2 Method

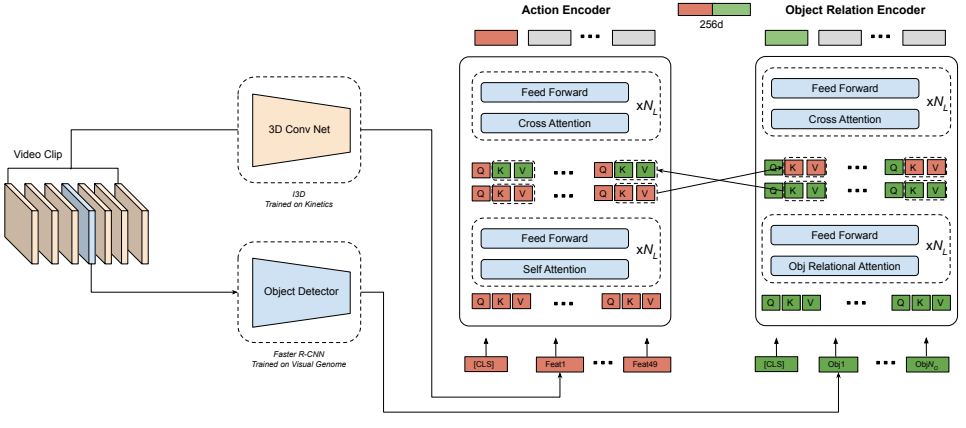


Figure 1: *Overview of clip anomaly encoders*. Objects are detected from center frame of the clip. 3D ConvNet (I3D) extracts spatio-temporal feature from the video clip. Detected object features are fed to object relation transformer. Spation-temporal features are reshaped and fed to action transformer. Cross-attention layers fuse memories after N_L layers. Output feature vectors corresponding to [CLS] tokens are concatenated to get clip feature.

VADOR, our anomaly detection framework, employs a two-stage approach, involving clip-level operations followed by video-level operations, to effectively detect anomalies in videos. In the first stage, the input video is divided into consecutive video clips, following the convention of action recognition methods, where each clip consists of 16 consecutive frames. Dedicated video clip encoders are then employed to generate feature vectors for each video clip. A comprehensive description of the encoding process can be found in Section 2.1.

In the second stage, the generated feature vectors of the video clips are organized sequentially to enable the temporal detection of anomalies using our TALNet. TALNet exploits the temporal dynamics present in the video data to accurately identify anomalies over extended time durations. The details of this process are described in Section 2.2.

2.1 Video Clip Encoders

We employ transformer encoders to generate video clip representation vectors based on the extracted object and action features within the clip. Specifically, our approach involves two transformer encoders: the object relation encoder and the action encoder. The object relation encoder focuses on processing the extracted object features and bounding boxes, which are obtained from the center frame (8th frame) of each video clip. Conversely, the action encoder handles the extracted action features.

The initial N layers of the transformer encoders exclusively consist of multi-head self-attention layers, followed by feed-forward layers. Cross attention does not exist in the layers. This design enables the encoding of object relations while keeping the action features distinct and separate.

To further enhance the modeling capability of our approach, we introduce cross-attention layers in the subsequent N layers of the transformer encoders. These cross-attention layers enable the establishment of cross-relations between objects and actions within the same video clip. Specifically, the cross-attention layers retrieve memory (key and value) from the last attention block of the other transformers. In our framework, the action encoder utilizes the memory from the Nth attention block of the object relation encoder, while the object relation encoder incorporates the memory from the Nth attention block of the action encoder, as illustrated in Figure 1.

Our proposed approach integrates transformer encoders with distinct attention mechanisms, enabling the effective representation of both object relations and actions within video clips. This architecture enhances the capability of our model to establish meaningful relationships between objects and actions.

Feature Extraction: The extraction of object and action features is performed using pre-trained networks, where the parameters of these networks remain fixed during training. To detect objects and extract their features, we employ Faster R-CNN trained on the Visual Genome dataset. Since the object detection model operates on static images, we conduct object detection solely on the center frame of each video clip, as depicted in Figure 1. We select the N_o objects with the highest confidence scores from the detector outputs to be forwarded to the object relation transformer encoder.

To extract action features from the video clips, we utilize a pre-trained I3D action recognition network, as illustrated in Figure 1. The extracted action features have a shape of $H \times W \times C$, where H and W are 7, and C is 1024. Since the transformer encoders accept sequential data, we reshape the action features to flattened representation of $(H \times W) \times C$.

Object Relation Encoder: The relationships among detected objects in the center frame of a video clip contain valuable information for anomaly detection. To capture these object relations, we employ the encoder component of the Object Relation Transformer [8]. The object relation encoder leverages self-attention mechanisms, considering both the object features and their relative positions through geometric attention. The relative position between a pair of objects is computed using Formula 1, utilizing the bounding box positions obtained from the object detector. By incorporating this positional information, a 4-dimensional relative position vector is expanded to a 512-dimensional vector using a trained fully connected

layer. The object relation weight, represented as a floating-point number, is computed by performing a matrix multiplication between the 512-dimensional vector and the learned W_G matrix.

$$\lambda(a,b) = (log(\frac{|x_a - x_b|}{w_a}), log(\frac{|y_a - y_b|}{h_a}), log(\frac{w_b}{w_a}), log(\frac{h_b}{h_a})) \tag{1}$$

$$r_{ab} = ReLU(Emb(\lambda(a,b)W_G)) \tag{2}$$

In the self-attention layers, the semantic relation weight between objects is calculated using the key (k), query (q), and value (v) vectors extracted from the object features. The object relation encoder combines geometric attention weights and semantic attention weights using Formula 3. The attention weights between object a and other objects are multiplied with the corresponding extracted value vectors (v) for those objects. The resulting attentioned vector Y_a is obtained by summing the multiplied values, as shown in Formula 4. This operation is applied to each object feature in the center frame of the video clip.

$$w_{ab} = \frac{r_{ab} + exp(\frac{q_a * k_b^T}{\sqrt{d_{model}}})}{\sum_{l=1}^{N} r_{al} + exp(\frac{q_a * k_b^T}{\sqrt{d_{model}}})} \tag{3}$$

$$Y_a = \sum_{l=1}^{N} w_{al} * v_l \tag{4}$$

It is worth to mention that, due to the utilization of relative positions in the geometric attention layers, the object features are not summed with sinusoidal positional encoding vectors before being passed to the encoder.

Action Encoder: Recognizing actions within the video clip is crucial for detecting abnormal situations. To process the extracted action features, we employ a transformer encoder [19] with self-attention mechanisms. Unlike the Object Relation Encoder, the input feature vectors of the action encoder are summed with positional encoding vectors.

$$\begin{aligned} Y_{action}^{i+1} &= Softmax(\frac{Q_{action}^{i} * (K_{obj}^{N_L})^T}{\sqrt{d_{model}}}) * V_{obj}^{N/2} \\ Y_{obj}^{i+1} &= Softmax(\frac{Q_{obj}^{i} * (K_{action}^{N_L})^T}{\sqrt{d_{model}}}) * V_{action}^{N/2} \end{aligned} \tag{5}$$

Cross Connections: In our architecture, the cross attention layers play a crucial role in establishing connections between the encoders. These layers retrieve the key (K) and value (V) vectors, representing the memory, from the cross transformer encoder, while the query (Q) vectors flow directly without crossing the encoders. This mechanism is illustrated in Formula 5. The cross attention layers replace the self-attention layers in the second half of the encoder layers.

Unlike the first half of the object relation encoder, the cross attention layers do not incorporate geometric attention. This is because the relative spatial distance between objects and actions is not considered in the cross attention layers. Therefore, geometric attention is not utilized in this particular part of the encoder architecture.

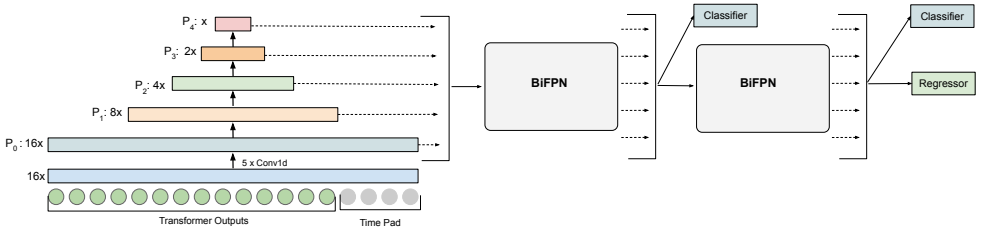


Figure 2: *Overview of Temporal Action Localization Network (TALNet).* Video clip features are bring together consecutively for each video. The clip feature window is fed to base network and then downscaling layers. Output of the downscaling layers are forwarded to BiFPN blocks. Classification network makes prediction from output of first BiFPN block. Output of second BiFPN block is fed to regression and classification networks. Abnormal segments are determined from regression and classification outputs with predefined anchors.

Clip Vector: The representation vector of a video clip is derived by concatenating the initial vectors from the output sequences of both the action encoder and the object relation encoder, as illustrated in Figure 1. These first vectors correspond to trainable *[FEAT]* token vectors, which are included in the input sequences. It is important to note that the action encoder and the object relation encoder each possess their own unique trainable *[FEAT]* token. This concatenation process enables the fusion of information from both encoders into a single clip-level representation vector.

2.2 Temporal Anomaly Localization Network (TALNet)

TALNet, plays a crucial role in predicting temporal anomaly segments within a video. It operates by processing the representation vectors of consecutive video clips generated from the input video. The first step in the TALNet process involves applying successive 1D convolution layers to the temporal dimension of the clip features. Notably, each convolution layer reduces the temporal dimension by half, as depicted in Figure 2. The outputs of these convolution layers are then passed through BiFPN stages.

BiFPN, originally introduced in the EfficientDet object detector [16], is adapted in our framework to fuse multi-level features through cross-level connections in the context of 1D temporal features. Importantly, the temporal dimensions of the level features remain unchanged during the BiFPN process. The output level features from the final BiFPN blocks are subsequently forwarded to a shared classifier and regressor.

Input Preparation: During the training phase, we construct the input by utilizing 128 consecutive video clips for each video. To introduce diversity in the input sequences, we randomly crop these 128 clips while ensuring the integrity of abnormal segments is preserved. This random cropping strategy enhances the variability of the input data.

During the inference phase, we pass the entire video clip features of a given video to TALNet without altering the order of the clips. Since our TALNet architecture is fully convolutional, it is capable of processing the entire video at once. The only requirement is that the temporal length of the video clip features must be a multiple of 16. To fulfill this condition, we pad the clip features until the temporal length becomes a multiple of 16. In the training phase, if necessary, we pad the input feature sequence with empty features, as

shown in Figure 2. This ensures that the input satisfies the required condition for TALNet processing.

Temporal BiFPN: The Temporal Bi-directional Feature Pyramid Network (BiFPN) plays a crucial role in processing the temporal features $P_{0...4}$, derived from the outputs of the 1D convolutions. Since the temporal features have varying lengths, the BiFPN is employed to effectively fuse these features using a combination of upsampling, downsampling, and convolution operations. It is important to note that the output shapes of the BiFPN remain consistent with the input vector shapes. Originally designed for object detection tasks in images, we have adapted the BiFPN framework to accommodate 1D operations.

Prediction Heads: The multi-scale features are processed by two consecutive BiFPN blocks, where each prediction in the output levels of the BiFPN corresponds to a segment anchor, following the approach of dense object detection. The outputs of the final BiFPN block are then passed through the classification and regression heads to determine the anomaly score and length of the segments. Notably, each scale level in the BiFPN outputs shares the same classification and regression heads.

To assign the target abnormal segment to the anchors, we calculate the Intersection over Union (IoU) between the anchors and the target segments. If the IoU value exceeds 0.5, we set the target label of the anchor to 1, indicating a positive match. To address the class imbalance between positive and negative classes, we employ the Focal Loss [9] for the output of the classification head.For the regression head outputs, which are responsible for temporal action localization, we utilize a modified version of the Smooth L1 Loss [6].

The output of the first BiFPN block is simultaneously fed into both the classification head and the next BiFPN block. To assign ground truth segments to the anchors, we replace IoU calculation with Intersection over Anchor (IoA) calculation, the rest of the process is same.

2.3 Implementation Details

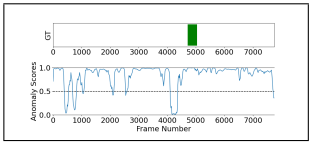
To prevent overfitting TALNet, we apply regularization methods similar to Clip Anomaly Encoders. We drop input clip features to TALNet with 0.2 probability. Also we apply drop block [4] to first temporal convolutions, since the convolutions keep shape of the features same.

3 Experiments

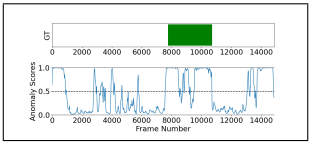
We evaluate VADOR on UCF Crime [15] and XD-Violence [21] datasets, using frame-level AUC and F1 score as performance metrics [13]. Results demonstrate that VADOR outperforms other methods in term of F1 score, while achieving comparable AUC score to SOTA methods. The impact of VADOR's video clip encoders is discussed in the following section.

Datasets: The UCF Crime dataset encompasses various real-world anomalies, including classes such as Explosion, Road Accident, and Burglary. The dataset comprises both normal and abnormal videos in the train and test splits. The training split contains 810 normal and 800 abnormal videos, while the test split consists of 150 normal and 140 abnormal videos. For training VADOR, we utilized second-by-second annotations from the training videos provided by [13].

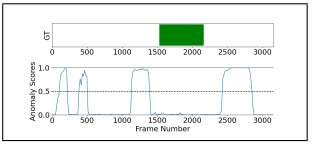
The XD-Violence dataset contains a total of 2405 violent videos and 2349 non-violent videos, with violent cases belonging to 6 classes. Except for the *Riot* class, the other classes are also present in the UCF Crime dataset. Unlike the CCTV camera recordings in the



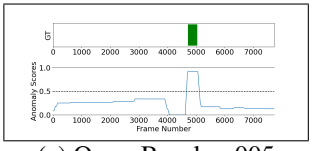
(a) RTFM: Burglary005



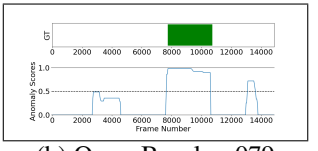
(b) RTFM: Burglary079



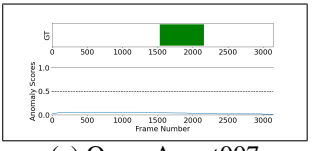
(c) RTFM: Arrest007



(a) Ours: Burglary005



(b) Ours: Burglary079



(c) Ours: Arrest007

Figure 3: *Qualitative Comparison on UCF Crime Videos.* Ground truth segments (green) and anomaly score predictions in timeline are presented

UCF Crime dataset, XD-Violence consists of videos captured from movies, sports events, car cameras, and more. To evaluate the performance of VADOR trained on the UCF-Crime dataset, we only used the test set of the XD-Violence dataset. Training VADOR on the XD-Violence dataset was not possible due to the unavailability of temporal annotations for the videos.

Methods	UCF Crime			
	F1@10	F1@25	F1@50	AUC
Sultani *et al.*[15]	45.20	39.64	32.32	75.41
RFTM [17]	33.55	26.14	16.86	84.44
S3R [20]	43.30	33.43	21.76	**85.99**
ADNet [13]	58.16	51.85	41.29	70.57
TALNet w/o encoders	62.72	57.36	43.40	69.37
VADOR (ours)	**69.79**	**63.09**	**50.28**	83.62

Table 1: *Quantitative comparison on UCF Crime Dataset.*

Results: The experimental results on UCF Crime dataset (Table 1) demonstrate that VADOR achieves state-of-the-art F1 scores. While VADOR's AUC score of 83.62 is lower than the best MIL-based method's score of 85.99, there is a significant difference in F1 score between the two models. This indicates that VADOR is better suited for real-world applications. Qualitative results in Figure 3 further support this conclusion.

In order to evaluate the performance of VADOR as a temporal action localization model, we included other models in the same domain to Table 1. Specifically, we compared VADOR with ADNet [13], which is an adaptation of the MS-TCN [6] model for temporal action localization on the UCF Crime dataset. Additionally, we trained a VADOR's TALNet without video clip encoders for further comparison. Our results indicate that VADOR achieved the highest scores among the evaluated TAL networks, providing evidence that the inclusion of video clip encoders contributes to an improved temporal localization performance.

We investigated importance of fusion of relation and action in encoders. We trained

Methods	UCF Crime			
	F1@10	F1@25	F1@50	AUC
VADOR only action	40.85	24.19	14.54	69.36
VADOR only object	65.78	57.78	42.75	74.50
VADOR cross-attention	**69.79**	**63.09**	**50.28**	**83.62**

Table 2: *Effect of cross attention layers.*

VADOR with only action encoder and with only object relation encoder. The results in Table 2 show that encoders with cross attention is important to get better performance. Furthermore, the results show that object relations are more useful than action to recognize anomalies in UCF Crime dataset. Action features are especially useful to recognize abnormal events which cause significant changes in action, such as explosions and fires. However, the majority of abnormal events in UCF Crime dataset occur in small portion of the scene, such as robbery, fight, abuse. Generally, these events do not create significant action changes, resulting in action features provides lower performance than object features on UCF Crime dataset.

Also we investigated generalization ability of our model. We evaluated UCF-Crime trained model on XD-Violance dataset. Similarly we evaluated TALNet and RFTM which trained on UCF-Crime dataset. The results in Table 3, VADOR has better generalization ability.

Methods	XD-Violance			
	F1@10	F1@25	F1@50	AP
TALNet	36.65	26.43	12.67	51.30
RFTM [10]	41.23	31.05	15.28	58.35
S3R [20]	44.26	31.19	14.75	61.96
VADOR	**49.74**	**40.41**	**25.07**	**65.90**

Table 3: *XD-Violance Scores*

Lastly, we compared VADOR with ADNet and our TALNet on UCF Crime V2 [13] dataset, which is an extension of UCF Crime Dataset. The results in Table 4 show that VADOR achieves best score in UCF Crime V2 dataset.

Methods	UCF Crime V2			
	F1@10	F1@25	F1@50	AUC
ADNet [13]	58.89	50.75	34.69	67.63
TALNet	64.31	54.63	42.16	72.19
VADOR	**68.58**	**59.72**	**47.09**	**80.99**

Table 4: *UCF Crime V2 Scores*

4 Conclusion

In this paper, we propose to use transformer encoders to capture object relations and actions in the scene. Fusion of action and object relation information with cross attention layers increases performance of VADOR. TALNet produces temporal abnormal segments of given video. Qualitative and quantitative results show that transformer encoders with cross attention layers provides better temporal anomaly segmentation performance. Delving into object analysis of objects and their interrelationships provides deeper insights into understanding anomalous events. To investigate VADOR's generalization, we evaluated its UCF-Crime trained model on the XD-Violance dataset, revealing superior generalization capabilities.

5 Acknowledgements

This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant No. 119E098 and Hacettepe University Scientific Research Projects Coordination Department under Grant No. FHD-2022-20044.

References

[1] 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.

[2] 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.

[3] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3584, 2019.

[4] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *Advances in neural information processing systems*, 31, 2018.

[5] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[7] 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.

[8] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *arXiv preprint arXiv:1906.05963*, 2019.

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[10] 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.

[11] 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.

[12] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1273–1283, 2019.

[13] Halil İbrahim Öztürk and Ahmet Burak Can. Adnet: Temporal anomaly detection in surveillance videos. In *International Conference on Pattern Recognition*, pages 88–101. Springer, 2021.

[14] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581. IEEE, 2017.

[15] 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.

[16] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[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] 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.

[20] 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.

[21] 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 *European conference on computer vision*, pages 322–339. Springer, 2020.

[22] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, Arif Mahmood, and Seung-Ik Lee. Cleaning label noise with clusters for minimally supervised anomaly detection. *arXiv preprint arXiv:2104.14770*, 2021.

[23] 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.

[24] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211*, 2019.