Dy-MIL: Dynamic Multiple-Instance Learning Framework for Video Anomaly Detection

Chen Li
Nanjing Vocational College of Information Technology
Mo Chen (chenmo@njcit.cn)
Nanjing Vocational College of Information Technology

Research Article

Keywords: Multiple-instance learning, Video anomaly detection, Dynamic ranking, Weakly supervised learning

Posted Date: May 18th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2906577/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Multimedia Systems on January 13th, 2024. See the published version at https://doi.org/10.1007/s00530-023-01237-0.

Dy-MIL: Dynamic Multiple-Instance Learning Framework for Video Anomaly Detection

Chen Li and Mo Chen

* Nanjing Vocational College of Information Technology, No. 99, Wenlan Road, Nanjing, 210023, Jiangsu, China.

*Corresponding author(s). E-mail(s): chenmo@njcit.cn;
Contributing authors: lichen@njcit.cn;

Abstract

Anomaly detection is an extremely challenging task in the field of visual understanding because it involves identifying events that deviate significantly from normal patterns. One of the primary reasons for the difficulty of this task is the diversity and complexity of anomalous events. Therefore, it is impossible for us to collect all types of anomalies and label them. In recent work, weakly supervised methods becomes one of the optimal solutions for anomaly detection. Thus, in this paper, we focus on weakly supervised learning and propose a dynamic multiple-instance learning framework for video anomaly detection, which develops a dynamic ranking method combined the $\boldsymbol{k}$-max-selection scheme to enlarge the inter-class distance between anomalous and normal instances by only using video-level labels. Experimental results demonstrate that our framework achieves superior improvements on three benchmark datasets, including the ShanghaiTech dataset, UCF Crime dataset and NUT dataset.

Keywords: Multiple-instance learning, Video anomaly detection, Dynamic ranking, Weakly supervised learning

1 Introduction

Video anomaly detection (VAD) is a challenging task, which faces open real-world scenarios with complex environments and various types of anomalies. It is impossible for us to collect all abnormal events, therefore, the semi-supervised learning becomes one of the optimal solutions for anomaly detection by using only normal training data.

The semi-supervised models [1] [2] learn the regularity of normal events, thus these models regard the events deviating from learned regularity as anomalies during testing.

The reconstruction-based method is one of the common representative methods in the semi-supervised learning. With continuous video frames as inputs, the reconstruction-based method [1] usually utilizes Auto-Encoder (AE) [3] with an encoder and a decoder to reconstruct the current input data. The encoder compresses the input data to obtain the feature representation, while the decoder expands the dimension of the feature representation to obtain the reconstructed frames with the same size as the input data. The model uses the reconstruction errors between the input data and reconstructed data as the basis of anomaly detection. However, due to the powerful generalization capacity of AE, some anomaly data can be also reconstructed well. In view of this, the prediction-based method [4] is proposed. According to this method, normal events can be predicted by continuous video frames at previous moments and conform to certain developmental regularity, while abnormal events deviate from this regularity. Compared with the reconstruction-based method, the prediction-based method uses the AE, the Generative Adversarial Network (GAN) [5], the Variational Autoencoder (VAE) [3], the Conditional VAE (CVAE) [6] or other neural networks with generative functions to generate the future video frame. The prediction error is used to measure whether the future video frame is abnormal. Recent works combine the above two mainstream paradigms to form hybrid models for anomaly detection task [7] [8] [2], and we believe that the hybrid models can be more robust to abnormal events.

One of the key challenges is that the collection of enormous labeled data is labor intensive and time-consuming. Therefore, in order to meet the challenge, amount of studies focus on weakly supervised methods that only needs video-level labels instead of frame-level labels. This is significant because it is much easier to annotate a mount of videos by assigning only video-level labels. Weakly supervised learning methods commonly take the anomaly detection as a Multiple Instance Learning (MIL) task. These methods could be traced to the work in [9], which introduced the UCF Crime dataset and a MIL method for VAD. Afterwards, a multiple instance self-training method [10] was developed for the discriminative representation by using weak labels. In order to weaken the inter-batch combination, a novel random batch-selection approach [11] was proposed in basis of a clustering assisted weakly supervised network. Specially, some works combined the weekly supervised methods with attention mechanisms in both spatial and temporal domains. Further, iterative supervised or weak/self-supervised learning methods [12] [13] were presented for effective anomaly detection.

Beside the above methods, several specifically designed methods for VAD were developed. An online anomaly detection strategy [14] was proposed by combining the transfer learning and the any-shot learning with a few labeled normal videos. An adversarial learning scheme [15] was introduced to overcome the lack of anomalous data by using pseudo-abnormal samples. Some works [16] [17] focused on developing novel learning losses to learn discriminative features for VAD models. Specially, a center-guided discriminative learning loss was proposed for an anomaly regression net to reduce the intra-class distances between normal instances.

We are interested in learning discriminative features for VAD models. Different to the work in [17], we develop a dynamic multiple-instance ranking method based on k-max selection scheme for the proposed framework, which focuses on optimizing the learning loss to enlarge the inter-class distance between anomalous and normal instances. Specifically, we adopt k-max selection scheme to optimize the multiple instance ranking process. We propose a novel dynamic ranking function to enforce the positive bags far apart from the negative bags in terms of anomaly scores, and ensure two negative bags closer than one positive bag and one negative bag. In addition, we explore the impact of different components of our framework on the performance of VAD.

Our main contributions are as follow:

- We propose a dynamic multiple-instance learning framework named Dy-MIL for VAD by only using video-level labels.
- Specially, we develop a dynamic ranking method for Dy-MIL to enlarge the inter-class distance between anomalous and normal instances.
- The proposed framework achieves superior improvements on popular benchmark datasets, including the ShanghaiTech dataset, UCF Crime dataset and NUT dataset..

The remainder of this paper is organized as follows: Section 2 introduces the details of the proposed Dy-MIL framework. Section 3 reports the experimental results and ablation studies on three benchmark datasets. Finally, Section 4 gives the conclusions.

2 Proposed Method

2.1 Problem definition

Given a dataset, $V = \{v_m\}_{m=1}^{M}$ denotes the training set with M videos. The MIL scheme only use video-level labels, which are denoted as $Y = \{y_m\}_{m=1}^{M}$, where $y \in (0, 1)$ and $y = 1$ if at least one anomaly presents in the video, else $y = 0$. Each video is represented as a bag, while an anomaly video with $y = 1$ as a positive bag, and a normal video with $y = 0$ as a negative bag. Then, each bag is split into a fixed number of temporal segments that are denoted as instances in the bag. A positive bag is denoted as b_p with positive instances $\{p_i\}_{i=1}^{t_i}$, while a negative bag is denoted as b_n with negative instances $\{n_i\}_{i=1}^{t_i}$, where t_i denotes the number of instances in the i-th bag. The predicted anomaly score of a video bag is represented as $\{s^i\}_{i=1}^{t_i}$, where s^i is the anomaly score of the i-th video instance.

2.2 Dynamic multiple-instance learning framework

2.2.1 Framework overview.

As shown in Fig. 1, we develop a dynamic multiple-instance ranking method for the proposed framework, which focuses on optimizing the learning loss to enlarge the inter-class distance between anomalous and normal instances. First, each video is

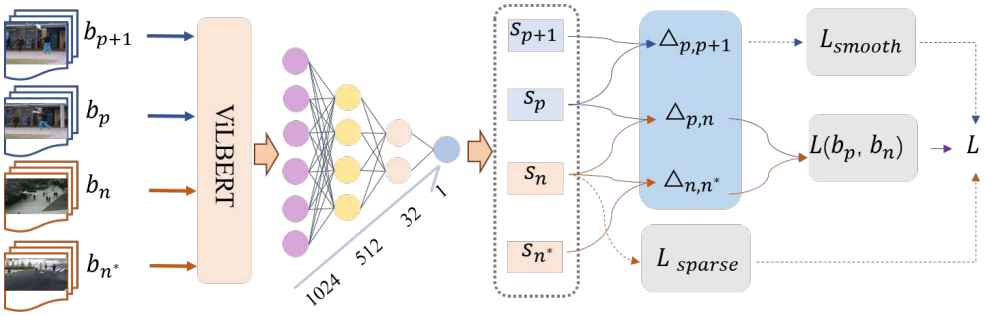

Fig. 1 Overview of Dy-MIL framework. b_p denotes a positive bag, and b_{p+1} denotes the next positive bag in temporal continuous. b_n and b_n^* denotes two different negative bags.The fully connected neural network is trained by adopting the proposed dynamic ranking loss $L(b_p, b_n)$, the smooth loss L_{smooth} and the sparse loss L_{sparse}. s_* represents the anomaly score of the video bag, and Δ_* represents the difference between two anomaly scores.

denoted as a bag. Here, b_{p+1} denotes the next positive bag of b_p in temporal continuous. b_n and b_n^* denotes two different negative bags. Then, we extract visual features for each video instance by using Vision-and-Language Bidirectional Encoder Representation from Transformers (ViLBERT) [18], which is a vision transformer model. Finally, we train a 3-layer fully connected neural network by adopting the proposed dynamic ranking method, which computes the dynamic ranking loss $L(b_p, b_n)$ as the basis component of instance anomaly scores. s_* represents the anomaly score of the video bag, and Δ_* represents the difference between two anomaly scores. L_{smooth} and L_{sparse} denotes the smooth loss and the sparse loss, respectively. Finally, the dynamic ranking loss, the smooth loss and the sparse loss are combined to the final loss L. Noted that the proposed MIL framework is based on the dynamic ranking method, thus named Dy-MIL.

2.2.2 Dynamic MIL ranking.

For anomaly detection, we train models to obtain higher anomaly scores for anomaly instances than normal ones. However, we only have video-level labels instead of instance-level labels. Inspired by the work in [9] that proposes the highest anomaly scores in positive bags and negative bags can be ranked, we propose a novel dynamic ranking method.

Different from [9], we develop k-max selection scheme to obtain the k-max anomaly scores. We adapt $s_p > s_n$ to the following objective function, which is defined as

$$mean(s_p^{k-max}) > mean(s_n^{k-max}), \tag{1}$$

where s_p^{k-max} and s_n^{k-max} represent k-max anomaly scores in the positive bag and in the negative bag, respectively. $mean$ traverses over video instances with k-max

anomaly scores in each bag. The s_*^{k-max} is defined as

$$\begin{cases} s_*^{k-max} = \{q_i^j | j = 1, 2, ..., k_i\}, \\ q_i = sort(s_i), \\ k_i = \lceil t_i/\beta \rceil, \end{cases} \tag{2}$$

where β is a hyperparameter, t_i is the number of instances in the i-th video bag. Instead of ranking by using only instances with the highest anomaly score in each video bag, we rank in basis of the instances with k-max anomaly scores in the positive bags and in the negative bags.

It should be noted that the value of k depends on the number of instances in the video bag as shown in the last line of Equation 2, thus it varies dynamically. This is why we call this ranking method as dynamic MIL ranking.

By using Equation 1, we aim to enforce the positive bags far apart from the negative bags in terms of anomaly scores. In addition, we want two negative bags are closer than one positive bag and one negative bag. Therefore, we develop the Equation 1 as follows

$$|\, mean(s_p^{k-max}) - mean(s_n^{k-max})| > |\, mean(s_n^{k-max}) - mean(s_{n^*}^{k-max})|. \tag{3}$$

Considering $|\, mean(s_n^{k-max}) - mean(s_{n^*}^{k-max})|$ is greater than 0, Equation (3) guarantees Equation 1. Further, in order to prevent over-fitting, we relax the Equation (3) as Equation (4), which is defined as

$$|\, mean(s_p^{k-max}) - mean(s_n^{k-max})| > \mu |\, mean(s_n^{k-max}) - mean(s_{n^*}^{k-max})|, \tag{4}$$

where μ represents the relax parameter.

2.2.3 Loss function

. The ranking loss is defined by the hinge-loss formulation as

$$\begin{aligned} L_{rank} \;=\; & max(0, 1 - mean(s_p^{k-max}) + mean(s_n^{k-max}) \\ & + \mu |\, mean(s_n^{k-max}) - mean(s_{n^*}^{k-max})|). \end{aligned} \tag{5}$$

To ensure the temporal smoothness and the sparsity of scores of positive bags, the loss function is adapted as

$$\begin{aligned} L(B_p, B_n) \;&=\; L_{rank} + \lambda L_{smooth} + \gamma L_{sparse} \\ &=\; L_{rank} + \lambda \sum_{p \in B_p} (s_p - s_{p+1}) + \gamma \sum_{n \in B_n} s_n, \end{aligned} \tag{6}$$

where $\sum_{p \in B_p}(s_p - s_{p+1}))$ is the constraint on the temporal smoothness, $\sum_{n \in B_n} s_n$ is the sparsity constraint, and λ, γ are balance parameters. The losses from k-max scored

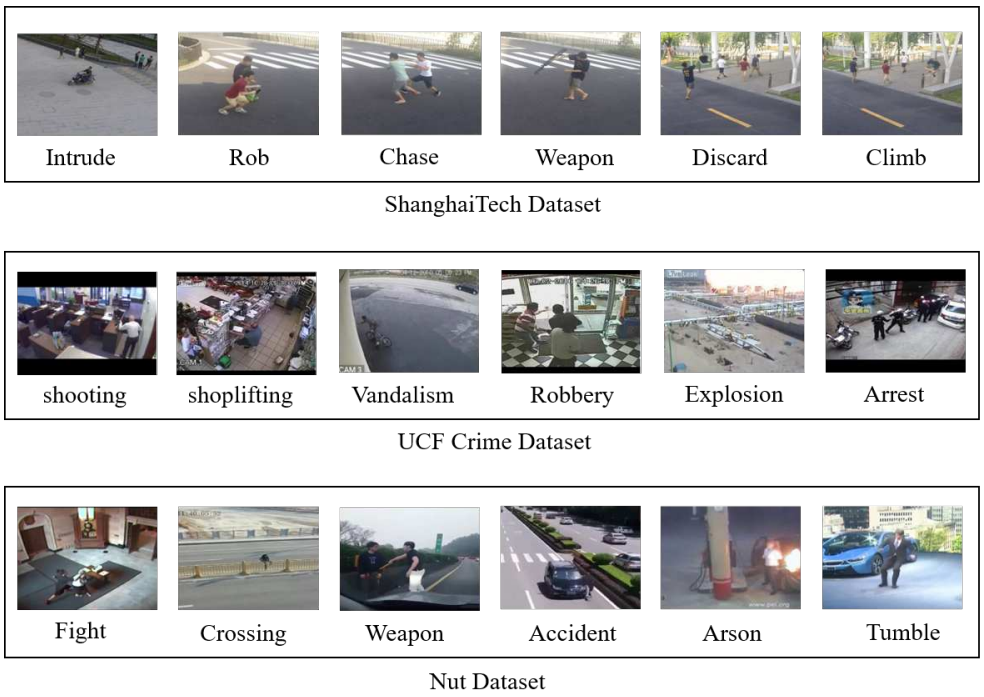

Fig. 2 Anomaly examples from three benchmark datasets. The first row is from the ShanghaiTech dataset, the second row is from the UCF Crime dataset, and the last row is from the NUT dataset.

video instances on both positive and negative bags are back-propagated, which forces the network to train a generalized model to predict high scores for anomaly instances.

Finally, the final loss function is given as

$$L(W) = L(B_p, B_n) + \omega||W||_F, \tag{7}$$

where W is the model weight and ω represents a balance parameter. $||*||_F$ denotes the Frobenius norm.

3 Experimental Results and Analysis

We evaluate our method on three challenging datasets, including the ShanghaiTech dataset [19], the UCF Crime dataset [9], and the NUT dataset [16]. Table. 1 shows the detailed differences between these three benchmark datasets, and Fig. 2 presents anomaly examples from these three benchmark datasets. We compare our proposed method with the previous methods, and evaluate the effects of different modules through ablation experiments.

Table 1 Comparison of popular benchmark datasets.

Datasets	Frames	abnormal videos	Scenes	Length	FPS
ShanghaiTech	317398	130	13	217min	24
UCF Crime	13769300	950	13	128h	30
NUT	38044	100	10	29min	15

3.1 Datasets

ShanghaiTech dataset [19]. It consists of 437 videos with 330 for training and the remaining 107 for testing. This dataset involves 130 abnormal events including objects other than pedestrians and violent sports in 13 different scenes, such as vehicle intrusion, chasing and fighting. This dataset is 220 minutes long.

UCF Crime dataset [9]. It contains 1900 real-world surveillance videos with 13 real-world anomalies, and half of the data are normal and the other half are abnormal. Specifically, the training set consists of 800 normal videos and 810 abnormal videos, while the testing set contains the remaining videos. This dataset is the largest scale one for VAD with 128 hours long in total.

NUT dataset [16]. It is a multi-views real-world anomaly detection dataset with large variations in scenes. It covers 10 abnormal events with 10 videos for each abnormal event, and includes 70 normal videos and 100 abnormal ones, e.g. weapon, intrusion, and arson. The training set covers 50 normal videos and 80 abnormal videos, while the testing set contains the remaining ones. Totally, this dataset is 29 minutes long. The shortest video is 1.7 seconds duration, and the longest one is 58.8 seconds duration.

3.2 Evaluation metrics and implementation details

For fair comparisons, we choose the same evaluate metrics to previous state-of-the-art methods [20]. First, the frame-level Receiver Operating Characteristic (ROC) is calculated with various thresholds. Then, the corresponding Area Under Curve (AUC) is accumulated for evaluating the performance of our model. The higher the AUC value is, the better the performance is.

We adapt the MIL [9] as the baseline, which only needs bag-level labels. We first divide each video bag into 32 non-overlapping instances, and then extract ViLBERT features [18] for each 16-frame video clip by using the pre-trained model on the conceptual captions dataset [21]. Note that we calculate the average of all clip features within each instance as the instance feature. Finally, these instance features with 1,024-dimensional are fed into a 3-layer fully connected neural network, whose layers has 512 units, 32 units and 1 unit, respectively. We use 60% dropout regularization [22] between these three fully connected layers, and the ReLU activation [23] and Sigmoid activation [24] for the first layer and the last layer, respectively. For optimizing the model, Adagrad optimizer [25] with the starting learning rate of 0.001 is used. A mini-batch with 30 positive bags and 30 negative bags is selected randomly. The learning loss is computed by using the Equation (6) and (7). We set $\mu = \lambda = \gamma = 8 \times 10^{-5}$, $\omega = 1 \times 10^{-2}$, and $\beta = 4$ for the best performance.

Table 2 AUC (%) of the Dy-MIL framework on benchmarks.

Methods	ShanghaiTech	UCF Crime	NUT
Binary classifier [9]	-	50.0	64.2
Conv-AE [26]	60.9	50.6	65.1
Sparse [27]	-	65.5	80.3
AR-Net [28]	85.4	-	81.9
CLAWS [11]	89.7	**83.0**	-
MIL [9]	86.3	75.4	82.1
MILR [29]	88.6	76.7	83.7
MIST [10]	**93.1**	81.4	-
MIL-Atten [30]	87.8	76.2	83.8
Dy-MIL(**ours**)	90.5	78.9	**86.1**

Table 3 Ablation studies of main components on UCF Crime.

Ablation experiments	k-max	new loss	ViLBERT	AUC(%)
Dy-MIL	✓	✓	✓	**78.9**
Dy-MIL$_k$	-	✓	✓	77.5
Dy-MIL$_l$	✓	-	✓	77.1
Dy-MIL$_v$	✓	✓	-	78.0
Dy-MIL$_{kl}$	-	-	✓	76.0
Dy-MIL$_{lv}$	✓	-	-	76.8
Dy-MIL$_{kv}$	-	✓	-	77.2
baseline	-	-	-	75.4

Table 4 Ablation studies of β on the UCF Crime dataset.

β	2	3	4	5	6
AUC (%)	76.3	78.2	78.9	78.4	77.5

3.3 Experimental results and ablation studies

3.3.1 Results and analysis.

To demonstrate the superior performance of the proposed method, we compare our method with current methods. Table 2 presents the quantitative comparison in different methods on AUC. As shown in Table 2, Dy-MIL achieves the best result on the NUT dataset, exceeding the second best one by 2.3%, and the second best result on the ShanghaiTech dataset. Our model outperforms the baseline by about 4% on all three datasets, while visual features of the proposed model is 1024-dimentional compared to 4096-dimentional in the baseline. And our model outperforms MIL-Atten, the baseline combined with attention scheme, which demonstrates the efficient of proposed dynamic scheme compared to the attention scheme. Although our model is not the best compared to CLAWS and MIST, their visual features are 2048 dimensional that is double of our model. Moreover, MIST uses pseudo labels to assist detection.

As shown in Fig. 3, we give some detected anomaly scores in example frames. The first three examples are success ones, and the last one is a failure case. The orange arrows point to the video frames when anomalies occur, and the red arrow points to a false positive. The light blue shadow shows the ground-truth corresponding to

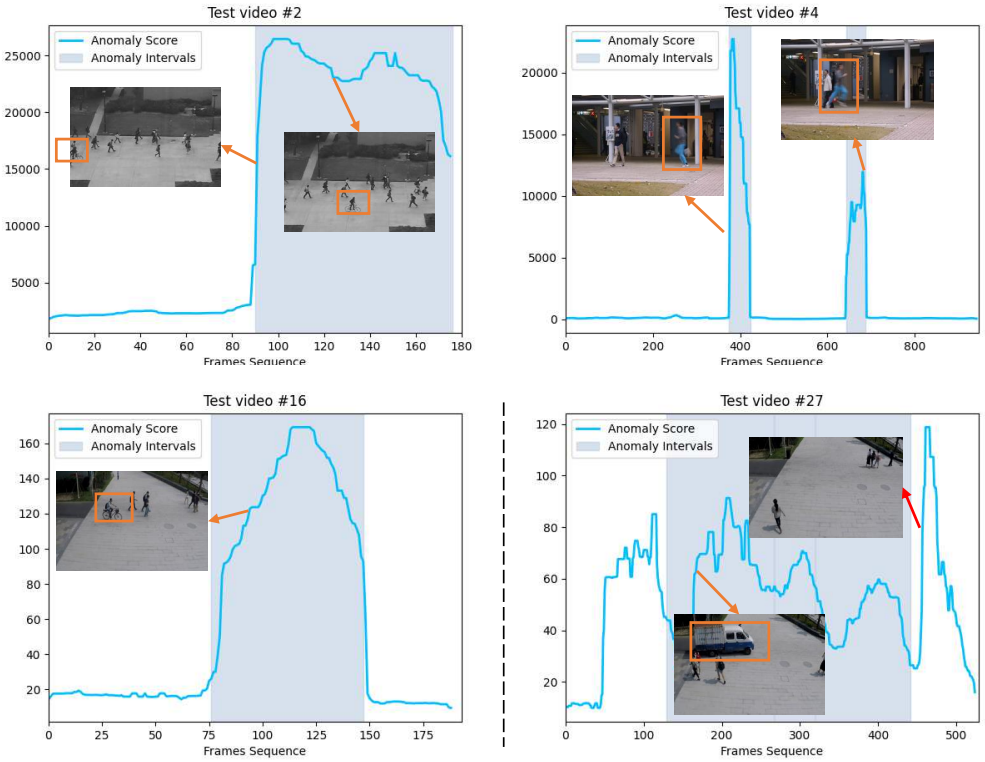

Fig. 3 Detected anomaly scores of Dy-MIL in video frames. The orange arrows point to the video frames when anomalies occur, and the red arrow points to a false positive. The ground-truth corresponding to anomalies are denoted in the light blue shadow. The last one is a failure case.

anomalies. The success examples show that the anomaly scores increase according to anomalies occur and decrease when the anomalies disappear. In the failure case, the video frame is blurred at the moment that leads to an increase in the abnormal score. Thus, our method is not robust to poor video frame quality.

3.3.2 Ablation studies.

In order to validate the effectiveness of main components in our model, we develop six ablation studies, including the k-max scheme, the new ranking loss, and the ViLBERT features. On the first three ablation studies, we replace k-max, *new loss* and $ViLBERT features$ with max, the loss as Equation 1, and C3D features, separately. We denote these three ablations as Dy-MIL$_k$, Dy-MIL$_l$ and Dy-MIL$_v$. On our fourth to sixth ablation studies, we replace two items of k-max, *new loss* and $ViLBERT features$, which are denoted as Dy-MIL$_{kl}$, Dy-MIL$_{lv}$ and Dy-MIL$_{kv}$, respectively. Table 3 shows that all AUC values of six ablation studies are lower than Dy-MIL and higher than the baseline. Therefore, our framework and the new loss

function are feasible. Moreover, we develop studies on the effectiveness of hyperparameter β as shown in Table 4. The results show that the optimal value of β is 4.

3.3.3 Computational time.

All experiments are performed on an NVIDIA GeForce GTX 1080 GPU and an Intel Core (TM) i7-8700K CPU @ 3.70GHz×12. The running speed is about 10fps. Thanks to the low dimensional of ViLBERT features, compared with the baseline, the fully connected neural network has only half of the parameters and FLOPs (0.5M number of parameters and 2.1M FLOPs).

4 Conclusions

To meet the key challenge that the collection of labeled data is labor intensive and time-consuming for video anomaly detection, we proposed a dynamic multiple-instance learning framework, which only required video-level labels for training. Specially, we developed a dynamic ranking method that was combined with the k-max selection scheme and the novel ranking function for the proposed framework, which focused on optimizing the learning loss to enlarge the inter-class distance between anomalous and normal instances. Experimental results showed that our framework achieved superior performance on three challenging benchmark datasets.

References

[1] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: ICPR, pp. 733–742 (2016)

[2] Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., Hua, X.-S.: Spatio-temporal autoencoder for video anomaly detection. In: ACM MM, pp. 1933–1941 (2017)

[3] Lin, S., Clark, R., Birke, R., Schönborn, S., Trigoni, N., Roberts, S.: Anomaly detection for time series using vae-lstm hybrid model. In: ICASSP, pp. 4322–4326 (2020)

[4] Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection–a new baseline. In: CVPR, pp. 6536–6545 (2018)

[5] Zaheer, M.Z., Mahmood, A., Khan, M.H., Segu, M., Yu, F., Lee, S.-I.: Generative cooperative learning for unsupervised video anomaly detection. In: CVPR, pp. 14744–14754 (2022)

[6] Liu, Z., Nie, Y., Long, C., Zhang, Q., Li, G.: A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In: ICCV, pp. 13588–13597 (2021)

[7] Ye, M., Peng, X., Gan, W., Wu, W., Qiao, Y.: Anopcn: Video anomaly detection via deep predictive coding network. In: ACM MM, pp. 1805–1813 (2019)

[8] Tang, Y., Zhao, L., Zhang, S., Gong, C., Li, G., Yang, J.: Integrating prediction and reconstruction for anomaly detection. PRL **129**(1), 123–130 (2020)

[9] Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6479–6488 (2018)

[10] Feng, J.C., Hong, F.T., Zheng, W.S.: Mist: Multiple instance self-training framework for video anomaly detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14009–14018 (2021)

[11] Zaheer, M.Z., Mahmood, A., Astrid, M., Lee, S.-I.: Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In: European Conference on Computer Vision (ECCV), pp. 358–376 (2020)

[12] Degardin, B., Proena, H.: Iterative weak/self-supervised classification framework for abnormal events detection. Pattern Recognition Letters **145**(1), 50–57 (2021)

[13] Zhong, J., Li, N., Kong, W., Liu, S., Li, T.H., Li, G.: Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1237–1246 (2019)

[14] Doshi, K., Yilmaz, Y.: Any-shot sequential anomaly detection in surveillance videos. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4037–4042 (2020)

[15] Georgescu, M.I., Ionescu, R., Khan, F.S., Popescu, M., Shah, M.: A background-agnostic framework with adversarial training for abnormal event detection in video. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**(1), 1–18 (2021)

[16] Li, Q., Yang, R., Xiao, F., Bhanu, B., Zhang, F.: Attention-based anomaly detection in multi-view surveillance videos. Knowl. Based Syst. **252**(2), 1–11 (2022)

[17] Wan, B., Fang, Y., Xia, X., Mei, J.: Weakly supervised video anomaly detection via center-guided discriminative learning. In: IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2020)

[18] Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems (NIPS), pp. 13–23 (2019)

[19] Luo, W., Liu, W., Gao, S.: A revisit of sparse coding based anomaly detection in stacked RNN framework. In: IEEE International Conference on Computer Vision (ICCV), pp. 341–349 (2017)

[20] Hinami, R., Mei, T., Satoh, S.: Joint detection and recounting of abnormal events by learning deep generic knowledge. In: IEEE International Conference on Computer Vision (ICCV), pp. 3639–3647 (2017)

[21] Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: 56th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 2556–2565 (2018)

[22] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(1), 1929–1958 (2014)

[23] Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: International Conference on Machine Learning (ICML), pp. 1–8 (2010)

[24] He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision (ECCV), pp. 630–645 (2016)

[25] Duchi, J.C., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research **12**(1), 2121–2159 (2011)

[26] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 733–742 (2016)

[27] Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: IEEE International Conference on Computer Vision (ICCV), pp. 2720–2727 (2013)

[28] Wan, B., Fang, Y., Xia, X., Mei, J.: Weakly supervised video anomaly detection via center-guided discriminative learning. In: IEEE International Conference on Multimedia and Expositions (ICME), pp. 1–6 (2020)

[29] Dubey, S., Boragule, A., Jeon, M.: 3D ResNet with ranking loss function for abnormal activity detection in videos. In: International Conference on Control, Automation and Information Sciences (ICCAIS), pp. 1–6 (2019)

[30] Zhu, Y., Newsam, S.: Motion-aware feature for improved video anomaly detection. In: 30th British Machine Vision Conference (BMVC), pp. 1–12 (2020)