LOCALITY-AWARE ATTENTION NETWORK WITH DISCRIMINATIVE DYNAMICS LEARNING FOR WEAKLY SUPERVISED ANOMALY DETECTION

Yujiang Pu[1], *Xiaoyu Wu*[1*]

[1]State Key Laboratory of Media Convergence and Communication,
Communication University of China, Beijing, China (*Corresponding author)
{pyj2020, wuxiaoyu}@cuc.edu.cn

ABSTRACT

Video anomaly detection is recently formulated as a multiple instance learning task under weak supervision, in which each video is treated as a bag of snippets to be determined whether contains anomalies. Previous efforts mainly focus on the discrimination of the snippet itself without modeling the temporal dynamics, which refers to the variation of adjacent snippets. Therefore, we propose a Discriminative Dynamics Learning (DDL) method with two objective functions, i.e., dynamics ranking loss and dynamics alignment loss. The former aims to enlarge the score dynamics gap between positive and negative bags while the latter performs temporal alignment of the feature dynamics and score dynamics within the bag. Moreover, a Locality-aware Attention Network (LA-Net) is constructed to capture global correlations and re-calibrate the location preference across snippets, followed by a multilayer perceptron with causal convolution to obtain anomaly scores. Experimental results show that our method achieves significant improvements on two challenging benchmarks, i.e., UCF-Crime and XD-Violence.

Index Terms— Video anomaly detection, weak supervision, multiple instance learning, temporal dynamics learning

1. INTRODUCTION

In recent years, the popularity of surveillance equipment has played an important role in preventing criminal behavior and maintaining public security. Relying on manual efforts to monitor these videos can no longer meet the urgent practical needs. Therefore, autonomous anomaly detection has received high attention due to its wide range of promising applications, including violence detection, traffic monitoring, content assessment, etc.

Given that anomalies are relatively scarce in real-life occasions, most previous work focuses on unsupervised algorithms [1–4], which learn the frequently occurring events as normal using one-class classifiers. Anomalies are then identified as outliers according to their departure from the learned representations of the normal class [5–8]. However, it is infeasible to collect all kinds of normal scenarios in real cases,
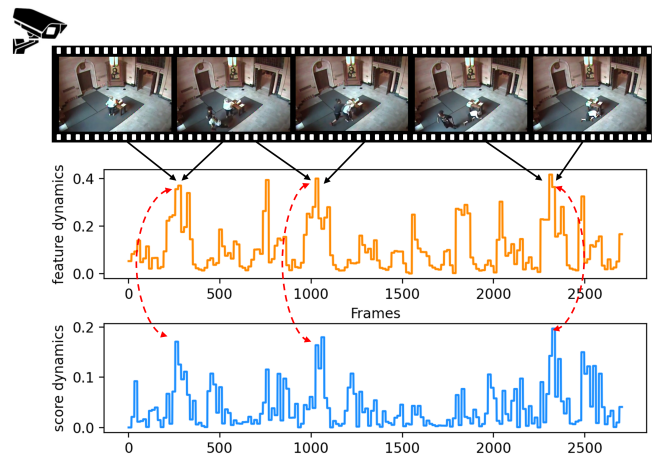

Fig. 1. Temporal consistency of feature dynamics and score dynamics. The peak of score dynamics caused by feature dynamics generally indicates a sharp difference in the current frame, i.e., a sudden intrusion, which further suggests a potential abnormal event.

and there is a huge intraclass diversity among normal classes, which may lead to incorrect detection of emerging normal categories.

Another widely used paradigm for anomaly detection is weakly supervised learning [9–11]. In this case, all events in a normal video are marked as normal while those in abnormal videos are treated as anomalous, even if some events are actually normal. Such methods reduce the cost of annotations, but these noisy labels present a new challenge for anomaly detection. Sultani et al. [9] propose a multiple instance learning approach to detect anomalies, where each video is regarded as a bag consisting of consecutive snippets. Anomaly videos are treated as positive bags, and normal ones are seen as negative bags. A multiple ranking loss with two constraints is introduced to distinguish positive and negative bags. In [12], Wan et al. construct a center regression loss to smooth anomaly scores while aggregating the activation of normal samples. Wu et al. [13] develop an HL-Net to capture both long-range dependencies and local distance relations with auxiliary au-

dio information. A MIST framework is built in [14] to generate pseudo labels for anomalous samples and employ a self-guided attention module to encode anomalous regions.

However, the above methods specialize in the discrimination of the snippets themselves without delving into the variation of adjacent snippets, which we call temporal dynamics. We observe that most anomalies typically occur at locations where the temporal dynamics changes dramatically, and such changes include both feature dynamics and score dynamics. As shown in Fig. 1, if abnormal behavior suddenly appears in the current frame, its corresponding feature dynamics usually change sharply, further leading to a large difference in the anomaly scores. The score dynamics in the positive bag should be greater than that in the negative bag, and the two types of temporal dynamics are shown to be causally consistent over the temporal dimension.

To address the above issues, we propose a novel Discriminative Dynamics Learning (DDL) method including two loss functions, of which the dynamics ranking loss captures the variation magnitude of anomalies by enlarging the accumulated score dynamics between positive and negative bags while the dynamics alignment loss achieves implicit ordering by aligning feature dynamics with score dynamics over the temporal dimension. Besides, a Locality-aware Attention Network (LA-Net) is built to model the long-range dependencies and re-calibrate the locality preference of adjacent snippets. Extensive experiments on two challenging benchmark demonstrate the superiority of our proposed framework over the current state-of-the-art methods.

2. PROPOSED METHOD

2.1. Formulation

In this paper, we consider video anomaly detection as a multiple instance learning task under weak supervision. Specifically, a video snippet bag $\mathcal{X} = \{x_i\}_{i=1}^{T}$ and its corresponding video-level label $y \in \{0, 1\}$ are given, where $y = 0$ indicates that the current video is normal otherwise $y = 1$ means that there exists anomalies in the video. In the test stage, the predicted anomaly score of the video is expressed as $\mathcal{S} = \{s_j\}_{j=1}^{T}$, where s_j is the anomaly score of the jth snippet. The goal of anomaly detection is to determine whether the video contains anomalies based on the predicted anomaly score while pinpointing the interval.

The overall structure of our method is shown in Fig. 2(a). Formally, a long untrimmed video is first divided into non-overlapping snippets through a sliding window of 16 frames. Then, we use a pre-trained I3D network [15] to extract snippet features and concatenate them over the temporal dimension to obtain video feature $X \in \mathbb{R}^{T \times D}$, where T is the number of snippets and D is the feature dimension. Subsequently, the LA-Net is used to model the global dependencies and re-calibrate the local preference. The calibrated feature $\widetilde{X}$ is then

fed into a Multilayer Perceptron (MLP) to capture robust snippet representation X^F, followed by a causal convolution layer to obtain the anomaly score $\mathcal{S}$. A Multiple Instance Learning (MIL) loss is used to monitor the discrimination of the snippets. The robust representation and the anomaly scores are further utilized in DDL for auxiliary optimization.

2.2. Locality-aware Attention Network

The Locality-aware Attention Network (LA-Net) aims to model the long-range dependencies across snippets while re-calibrate the contextual relations of local distance, as shown in Fig. 2(b). Following the self-attention mechanism [16], we construct the global attention map as

$$\mathcal{A}_{ij} = \frac{exp\{\mathcal{R}_k(x_i, x_j)\}}{\sum_{n=1}^{T} exp\{\mathcal{R}_k(x_i, x_n)\}}, \tag{1}$$

where $\mathcal{R}_k = \phi(x_i)^T \psi(x_j)$ is a multi-head kernel function to capture diversity semantic patterns, $\phi(\cdot)$ and $\psi(\cdot)$ are two linear functions applied on input snippets, and k is the head numbers.

The generated global attention map develops the long-distance relation across snippets but ignores the inherent location preference of the video sequence. Thus, a Gaussian-like location prior is introduced to re-calibrate the local contextual correlations, which can be expressed as

$$\mathcal{G}_{ij} = exp(-\frac{|i - j|^2}{2\sigma}), \tag{2}$$

$$\widetilde{\mathcal{A}} = \mathcal{A} + \mathcal{G}, \tag{3}$$

where i and j are the relative positions of the snippets, and σ is a hyperparamter controlling the deviation of the center location. Generally, a higher $\mathcal{G}_{ij}$ indicates a higher prior probability of the ith snippet to the jth snippet. By adding $\mathcal{G}$ as a bias term to the global attention map, the representation of the current snippet can be explicitly enhanced while suppressing the interference of long-distance. Correspondingly, our Locality-aware Attention Network is formulated as

$$\widetilde{X} = Norm(stack(\widetilde{\mathcal{A}}XW_k)W_h + X), \tag{4}$$

where $W_k \in \mathbb{R}^{D \times \frac{D_h}{k}}$ and $W_h \in \mathbb{R}^{D_h \times D}$ are two linear projection layers, and $Norm(\cdot)$ denotes the layer normalization.

2.3. Multiple Instance Learning

Subsequently, a two-layered MLP $\mathcal{F}^{\Theta}$ with GELU activation and dropout operation is followed to obtain robust snippet representation X^F, as shown in Fig. 2(c). To acquire reliable anomaly scores, we use a 1×1 causal convolution layer to capture historical observations. The process is expressed as

$$\mathcal{S} = \sigma(W\mathcal{F}^{\Theta}(\widetilde{X}) + b), \tag{5}$$
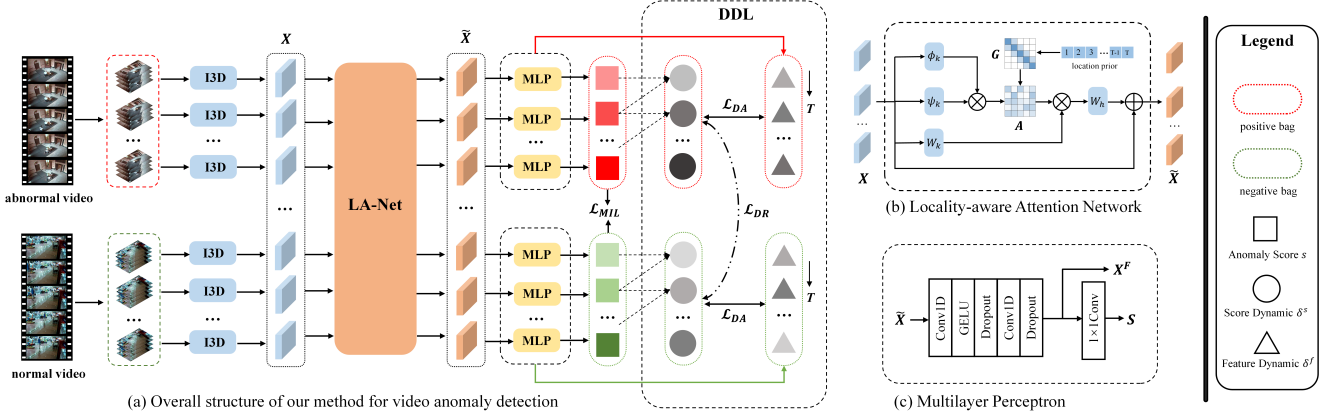
Fig. 2. Overview of the proposed video anomaly detection framework. It comprises a Locality-aware Attention Network (LA-Net), a Multilayer Perceptron (MLP) and Discriminative Dynamics Learning (DDL).

where W is a 1×1 convolution with kernel size K and b is a bias term. $\sigma(\cdot)$ denotes a sigmoid function and $\mathcal{S} \in \mathbb{R}^{1 \times T}$ is the temporal anomaly scores of the video bag.

Following [12, 13, 17], we use the k-max approach to calculate multiple instance learning (MIL) loss. Specifically, if the video is abnormal, the average of the top-k prediction results in S is used as the video anomaly score, where $k = \lfloor \frac{T}{16} + 1 \rfloor$. If the video is normal then the highest prediction score is selected, i.e., $k = 1$. Accordingly, our MIL-based loss function is formulated as

$$\mathcal{L}_{MIL} = \frac{1}{N} \sum_{i=1}^{N} -y_i log(p_i), \quad (6)$$

where $\{p_i\}_{i=1}^{N}$ is the video anomaly score and $\{y_i\}_{i=1}^{N}$ is the binary video-level ground truth.

2.4. Discriminative Dynamics Learning

Since the anomaly score in the negative bag is usually close to the positive one at the early stage of training, direct score ranking patterns [9, 10, 18] are likely to produce incorrect decision boundaries, which may lead to further deviations from the ground truth. To alleviate this problem, we design a dynamics ranking loss that amplifies the magnitude of abnormal score changes while suppressing score fluctuations.

Specifically, the anomaly score of a video bag is denoted as $S = \{s_1, s_2, \ldots, s_t\}$ and its corresponding score dynamics is expressed as $\Delta S = \{\delta_1^s, \delta_2^s, \ldots, \delta_{t-1}^s\}$, where $\delta_t^s = |s_t - s_{t+1}|$. Then, we use the top-$k$ dynamics in the bag to calculate the accumulation of score dynamics, which is expressed as

$$\mathcal{E}_{\Delta S} = \frac{1}{k} \sum_{t=1}^{k} |\delta_t^s|^2, \quad (7)$$

where $k = \lfloor \frac{T}{16} + 1 \rfloor$ if the video is abnormal otherwise $k = 1$. Considering that the dynamics accumulation of a positive bag

should be remarkably larger than that of a negative bag, the dynamics ranking (DR) loss is naturally formulated as

$$\mathcal{L}_{DR} = max(0, \zeta - \mathcal{E}_{\Delta S}^a + \mathcal{E}_{\Delta S}^n), \quad (8)$$

where $\mathcal{E}_{\Delta S}^a$ and $\mathcal{E}_{\Delta S}^n$ are dynamics accumulation of positive and negative bags, respectively, and ζ is a margin hyperparameter. This loss function enlarges the dynamics gap between positive and negative bags, further leading to a suppression of dynamics accumulation in the negative bag while increasing that of the positive bag.

Besides, snippets at abnormal boundaries are more variable, resulting in a sharp jittering at the junction. This jittering is reflected in the change of anomaly scores after forward discrimination, that is, the score dynamics and feature dynamics have a causal consistency over the temporal dimension. To this end, we introduce a dynamics alignment strategy to model this tendency. Given the robust snippet representation after two-layered MLP denoted as $X^F = \{x_1^F, x_2^F, \ldots, x_t^F\}$, its corresponding feature dynamics can be expressed as $\Delta X = \{\delta_1^f, \delta_2^f, \ldots, \delta_{t-1}^f\}$, in which the δ_t^f is calculated as

$$\delta_t^f = 1 - \frac{x_t^F x_{t+1}^F}{\|x_t^F\| \|x_{t+1}^F\|}, \quad (9)$$

Considering that the feature dynamics and the score dynamics are in two different semantic spaces, we use Kullback-Leibler (KL) divergence rather than mean square error to learn this property. Thus, our dynamics alignment (DA) loss is formulated as follows:

$$\mathcal{L}_{DA} = \frac{1}{N \times (T-1)} \sum_{i=1}^{N} (\sum_{t=1}^{T-1} -\delta_t^s log(\delta_t^f + \epsilon))_i, \quad (10)$$

where $\{\delta_t^s\}_{t=1}^{T-1}$ and $\{\delta_t^f\}_{t=1}^{T-1}$ are score dynamics and feature dynamics in the same video bag, respectively, and ϵ is a

very small scalar to prevent feature dynamics from being zero. This loss makes the feature dynamics distribution consistent with the score dynamics over temporal dimension, while implicitly increasing the discriminability of normal and abnormal snippets within the video bag.

Combined with the MIL-based loss, the overall objective function of our model is formulated as

$$\mathcal{L} = \mathcal{L}_{MIL} + \lambda_1 \mathcal{L}_{DR} + \lambda_2 \mathcal{L}_{DA}, \qquad (11)$$

where λ_1 and λ_2 are pre-defined weights for dynamics ranking loss and dynamics alignment loss, respectively.

3. EXPERIMENTS

In this section, we evaluate our approach on two benchmark datasets (UCF-Crime and XD-Violence). First, we present the two datasets and their corresponding evaluation metrics in detail. The implementation details of the experiments are then given. Finally, we compare our results with several state-of-the-art methods, a series of quantitative and qualitative experimental results elucidating the effectiveness of our method.

3.1. Datasets

UCF-Crime [9] is a large-scale dataset collected from surveillance with a total duration of 128 hours, containing 1900 long untrimmed videos. The dataset covers 13 types of anomalies in 1610 training videos and 290 test videos, where the training videos have only video-level labels while the test provides frame-level annotations. Following [9, 12, 14, 19], we use the area under the frame-level ROC curve (AUC) to evaluate the performance of the proposed method.

XD-Violence [13] is a newly released anomaly detection dataset that covers 6 types of violence events. The dataset consists of 4754 untrimmed videos including video-level labeled training set and frame-level labeled test set, with a total duration of 217 hours. Unlike UCF-Crime, the dataset is collected from various scenes including movies, sports, surveillance footage, CCTV, etc. As in [13, 20], we use frame-level average precision (AP) to evaluate the proposed method.

3.2. Implementation Details

The head number k of LA-Net is set to 4, and the hidden dimension D_h for UCF-Crime is set to 512, which is adopted as 128 for XD-Violence. The location prior value σ is empirically set to 16 and 6 for UCF-Crime and XD-Violence, respectively. The two Conv1D layers in MLP have 512 nodes and 128 nodes, regularized by dropout with a probability of 0.1 between each layer. The kernel sizes of the 1×1 causal convolution are set to 10 and 5 for UCF-Crime and XD-violence, respectively. Limited by computational resources,

we use a uniform sampling strategy as in [13, 18] to obtain a fixed length of 200 snippets in the training phase.

For a balanced purpose, the weights λ_1 for UCF-Crime and XD-violence are set to 1 and 2, respectively, while λ_2 is adopted as 1. The positive scalar ϵ is set to 10^{-7}, and the margin ζ is set to 0. We use the Adam optimizer with a 128 mini-batch size to train the model for 50 epochs, and the initial learning rate is set to 5×10^{-4} with a cosine decay strategy. All experiments are conducted on an NVIDIA Tesla A40 GPU based on PyTorch.

3.3. Quantitative Comparison with existing Methods

We first compare our approach against several state-of-the-art methods under weak supervision. Table 1 presents the frame-level AUC values on the UCF-Crime dataset. Notably, our method surpasses the previous state-of-the-art MIL-based methods, Sultani et al. [9] by 7.2%, Zhong et al. [11] by 3%, Wu et al. [13] by 2.68%. Even when compared to the latest RTFM [20], our result is also ahead by 0.82%.

Table 2 shows the frame-level AP values on the XD-Violence dataset. Remarkably, our approach substantially outperforms other methods, including both unsupervised and weakly supervised methods. Our result is 5.31% higher than Wu et al. [13] and exceeds RTFM [20] by 2.91%, reaching a new state-of-the-art frame-level AP of 80.72%. Moreover, our method is the only one exceeding 80% in terms of AP on XD-Violence with only RGB features. The superior results on both datasets demonstrate the effectiveness and capacity of the proposed method.

Compared with previous approaches, our LA-Net introduces explicit location prior to model local contextual relations, which suppresses long-distance redundancies and strengthens the local representation of adjacent snippets. Moreover, the proposed DDL method effectively captures the temporal dynamics of the video bag. The accumulation of score dynamics is ranked to increase the discriminability between positive and negative bags, while the feature dynamics are aligned with score dynamics within the bag to achieve causal consistency over the temporal dimension.

Table 1. Frame-level AUC performance on UCF-Crime.

Method	Feature	AUC(%)
Sultani *et al.* [9]	C3D RGB	75.41
Zhang *et al.* [10]	C3D RGB	78.66
Motion-Aware [21]	PWC Flow	79.00
Zhong *et al.* [11]	TSN RGB	82.12
Wu *et al.* [13]	I3D RGB	82.44
MS-BSAD [18]	I3D RGB	83.53
RTFM [20]	I3D RGB	84.30
DDL (Ours)	I3D RGB	**85.12**

Table 2. Frame-level AP performance on XD-Violence.

Method	Feature	AP(%)
SVM baseline	-	50.78
OCSVM [22]	-	27.25
Hasan *et al.* [23]	-	30.77
Sultani *et al.* [9]	C3D RGB	73.20
Wu *et al.* [13]	I3D RGB	75.41
RTFM [20]	I3D RGB	77.81
DDL (Ours)	I3D RGB	**80.72**

3.4. Ablation Study

The location prior is a critical component of our LA-Net. Table 3 shows that the introduction of the prior improves the performance of the model on both datasets, with a 0.61% improvement on UCF-Crime and a 0.77% increase on XD-Violence. This indicates that video sequences have a certain location preference, which can be enhanced by adding a specific location prior. Meanwhile, the location prior can further suppresses the long-distance noise information and improve the robustness of the current snippet, which has a positive impact on modeling adjacent temporal dynamics.
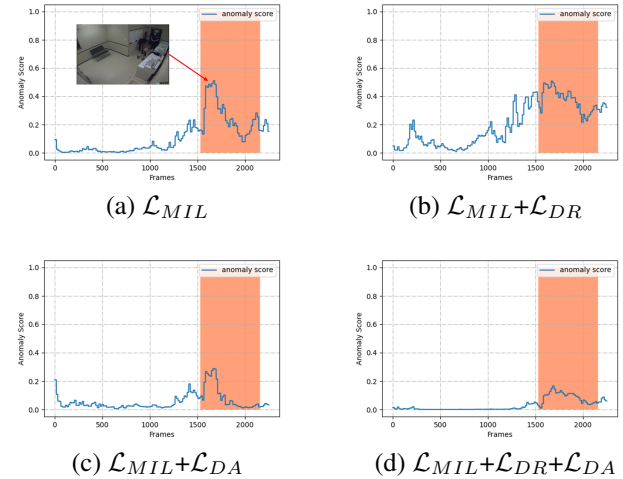
We then compare the effect of different dynamics loss functions, as shown in Table 4. Both $\mathcal{L}_{DR}$ and $\mathcal{L}_{DA}$ loss have distinct improvements on the basis of $\mathcal{L}_{MIL}$. Notably, the effect of $\mathcal{L}_{DA}$ is better than that of $\mathcal{L}_{DR}$, which illustrates the importance of maintaining causal consistency of temporal dynamics within the bag. Moreover, the synergistic effect of both is demonstrated when introduced simultaneously, which are shown to be complementary with the MIL-based loss. In addition, the feature dynamics produce an implicit ordering in the process of alignment with the score dynamics, which further enhances the discriminability of the video snippets.

Table 3. Ablation study of location prior.

Model	UCF-Crime AUC(%)	XD-Violence AP(%)
LA-Net w/o prior $\mathcal{G}$	83.06	78.41
LA-Net w/ prior $\mathcal{G}$	83.67	79.18

Table 4. Ablation study of the DDL method.

$\mathcal{L}_{MIL}$	$\mathcal{L}_{DR}$	$\mathcal{L}_{DA}$	UCF-Crime AUC(%)	XD-Violence AP(%)
✓			83.67	79.18
✓	✓		84.04	80.15
✓		✓	84.33	80.23
✓	✓	✓	**85.12**	**80.72**

(a) $\mathcal{L}_{MIL}$ (b) $\mathcal{L}_{MIL}+\mathcal{L}_{DR}$

(c) $\mathcal{L}_{MIL}+\mathcal{L}_{DA}$ (d) $\mathcal{L}_{MIL}+\mathcal{L}_{DR}+\mathcal{L}_{DA}$

Fig. 3. Visualization results of our DDL method on UCF-Crime. Orange windows indicate that the interval contains an abnormal event.

3.5. Qualitative Analysis

Finally, we visualize the anomaly detection results before and after discriminative dynamics learning. As shown in Fig. 3(a), the anomaly score predicted with MIL-based loss fluctuates widely before and after the ground truth boundary. With dynamics ranking loss, the anomaly score roughly increases over temporal dimension, as shown in Fig. 3(b), which is due to the increase of score dynamics within the positive bag during the pairwise ranking process. Besides, it smoothes out the anomaly score within the ground truth interval. And after the temporal dynamics alignment, the peaks of the anomaly score in Fig. 3(c) is suppressed and the overall prediction is in a lower range, thanks to the alignment implicitly improving the discriminability of the snippets within the bag. When both losses are added to the optimization, as Fig. 3(d) shows, the predicted interval overlaps with the ground truth, and the anomaly score outside the boundary is significantly suppressed, which further improves the capacity to locate the anomaly intervals.

4. CONCLUSION

In this paper, we propose a Discriminative Dynamics Learning (DDL) method containing dynamics ranking loss and dynamics alignment loss for weakly supervised video anomaly detection. Dynamics ranking loss boosts the response magnitude of anomalous events by enlarging the anomaly score dynamics between positive and negative bags. Dynamics alignment loss explicitly aligns the two types of temporal dynamics within the video bag, thus smoothing out the dynamic variation of the inner bag snippets. In addition, we construct a prior-based Locality-aware Attention Network (LA-Net),

which captures the long-range temporal dependencies across snippets while recalibrating the contextual correlations within the neighborhood. Experimental results on two large anomaly video datasets demonstrate the effectiveness of our approach. In the future, online detection and multimodal information will be further explored.

5. ACKNOWLEDGEMENTS

6. REFERENCES

[1] Antoni Chan and Nuno Vasconcelos, "Ucsd pedestrian dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[2] 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*, 2013, pp. 2720–2727.

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

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4183–4192.

[5] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3379–3388.

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

[7] 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 International Conference on Computer Vision*, 2019, pp. 1705–1714.

[8] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee, "Old is gold: Redefining the adversarially learned one-class classifier training paradigm," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14183–14193.

[9] 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*, 2018, pp. 6479–6488.

[10] Jiangong Zhang, Laiyun Qing, and Jun Miao, "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection," in *IEEE International Conference on Image Processing*, 2019, pp. 4030–4034.

[11] 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 Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1237–1246.

[12] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei, "Weakly supervised video anomaly detection via center-guided discriminative learning," in *IEEE International Conference on Multimedia and Expo*, 2020, pp. 1–6.

[13] 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 *Proceedings of the European Conference on Computer Vision*, 2020, pp. 322–339.

[14] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng, "Mist: Multiple instance self-training framework for video anomaly detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14009–14018.

[15] 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*, 2017, pp. 6299–6308.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[17] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury, "W-talc: Weakly-supervised temporal activity localization and classification," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 563–579.

[18] Yang Zhen, Yuanfang Guo, Jinjie Wei, Xiuguo Bao, and Di Huang, "Multi-scale background suppression anomaly detection in surveillance videos," in *IEEE International Conference on Image Processing*, 2021, pp. 1114–1118.

[19] Kun Liu and Huadong Ma, "Exploring background-bias for anomaly detection in surveillance videos," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1490–1499.

[20] 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," *arXiv preprint arXiv:2101.10030*, 2021.

[21] Yi Zhu and Shawn Newsam, "Motion-aware feature for improved video anomaly detection," in *Proceedings of the British Machine Vision Conference*, 2019.

[22] Bernhard Schölkopf, Robert C Williamson, Alexander J Smola, John Shawe-Taylor, John C Platt, et al., "Support vector method for novelty detection.," in *NeurIPS*, 1999, pp. 582–588.

[23] 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*, 2016, pp. 733–742.