Enhancing Feature Representation for Anomaly Detection via Local-and-Global Temporal Relations and a Multi-Stage Memory (Supplementary Material)

Xuan Li, Ding Ma, and Xiangqian Wu(✉)

Faculty of Computing, Harbin Institute of Technology, Harbin, China
xuanli@stu.hit.edu.cn, martin3436@yeah.net, xqwu@hit.edu.cn

1 Contributions of Our Work

- We propose a MMFDL method, which builds a ENPM to memorize easy normal patterns and makes HIF close to the ENPM and away from EAIF to learn discriminative features for WS-VAD.
- We design a LGTRM module, consisting of two subnetworks, *i.e.*, DW-Net and TF-Net. DW-Net aggregates the current clip feature with its adjacent clip features to capture short-range temporal dependencies to augment clip features. TF-Net integrates the current clip features with global context cues to obtain long-range temporal dependencies by the multi-head self-attention mechanism [11] of transformer to enhance clip features.
- Our framework obtains frame-level AUC value of 97.94% on the ShanghaiTech [3] dataset and 87.16% on the UCF-Crime [9] dataset, outperforming the current state-of-the-art methods.

2 Related Work

Weakly supervised video anomaly detection (WS-VAD) methods are widely studied recently. Many researchers [1, 2, 5, 7–10, 12, 14–20] worked on solving this problem through enhancing feature representation for WS-VAD so that abnormal and normal events can be easily distinguished by the classifier. Feature representation enhancing in WS-VAD are mainly implemented via two ways: temporal relations modeling and feature discrimination learning.

2.1 Feature Discrimination Learning

[1, 4, 8, 19] focused on designing a task-specific feature encoder to extract discriminative features from raw video data. Zaheer *et al.* [17] proposed a clustering distance based loss to produce better anomaly representations by encouraging the model to generate distinct normal and anomalous clusters. Sultani *et al.* [9] adopt a framework based on MIL with a ranking loss to solve the WS-VAD

problem. However, their loss only ensures the video-level separability instead of clip-level separability. To address this issue, some works [7,10,12,14] maximized the discrepancy between the HIF and EAIF selected by the top-k method to improve the clip-level separability. The selected EAIF may contain normal features as the fine-grained frame-level labels of abnormal videos are not available. The noise in the EAIF will further degrade the margin between abnormal and normal feature representation. While in our method, we not only push HIF to stay away from EAIF, but also pull HIF close to ENPM to mitigate this problem and achieve interclass dispersion.

2.2 Temporal Relations Modeling

Zhong *et al.* [19] uses the graph convolution network to capture temporal relations. Lv *et al.* [5] proposed a high-order context encoding model to effectively utilize the temporal context by measuring the dynamic variations. Pu *et al.* [7] designed an LA-Net to obtain global correlations and re-calibrate the location preference across snippets. Tian *et al.* [10] proposed to capture short and long range temporal dependencies by a combination of multi-scale dilated convolution and a self-attention module. Li *et al.* [2] used a structure similar to CvT [13] to introduce local and global interframe perception. However, they overlook the order information of video clips. Wu *et al.* [14] proposed to fuse a non-local module with a fixed positional prior to mine the order information and semantic similarity between video clips. Compared with previous works, our TF-Net uses a vanilla transformer encoder which is more powerful in capturing global context and a learnable positional encoding which is more flexible than the fixed one. Besides, our DW-Net makes full use of adjacent clip features to perceive and strengthen local information. The DW-Net and TF-Net jointly form the LGTRM module to encode the local and global temporal relations.

3 Impact of the Maximum Number of Input Video Clips

Following [1,2,9,10,14], we conduct the experiments by using different maximum numbers (*i.e.* 32 [2, 9, 10], 96 [1], 200 [14]) on ShanghaiTech. As shown in the Table 1, we achieve suboptimal results with the small maximum number setting. The reason for this is that our method contains positional encoding so that in the testing stage we need to use sliding windows instead of taking all the video clips as input, causing only short-range dependency. We obtain a suboptimal result with too large maximum number because the training set only contains a few long videos, our positional encoding cannot be learned well.

4 Exploration of the MMFDL Method

We explore our MMFDL method on ShanghaiTech. In Fig. 1, we visualize the similarity matrix $M^{\mathsf{T}}M$ mentioned in $\mathcal{L}_d$ by a heat map. We can observe that

Table 1. Ablation studies of the length of transformer encoder on ShanghaiTech using AUC (%) value.

Length	32	96	112	200
AUC	96.71	97.92	97.94	97.65

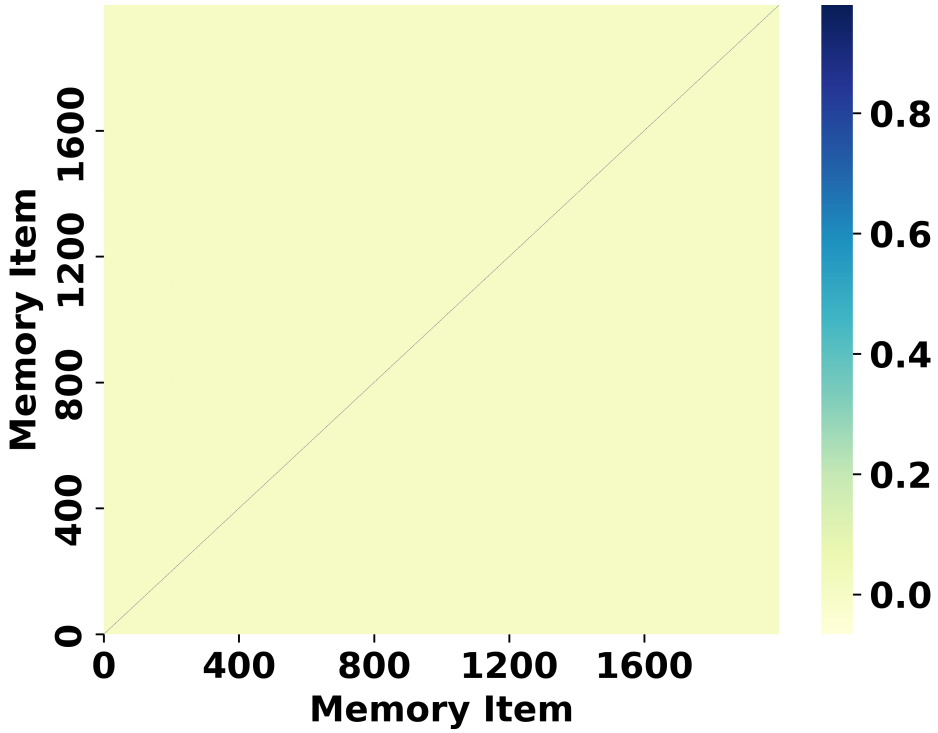

Fig. 1. Visualization of $M^{\top}M$ on ShanghaiTech.

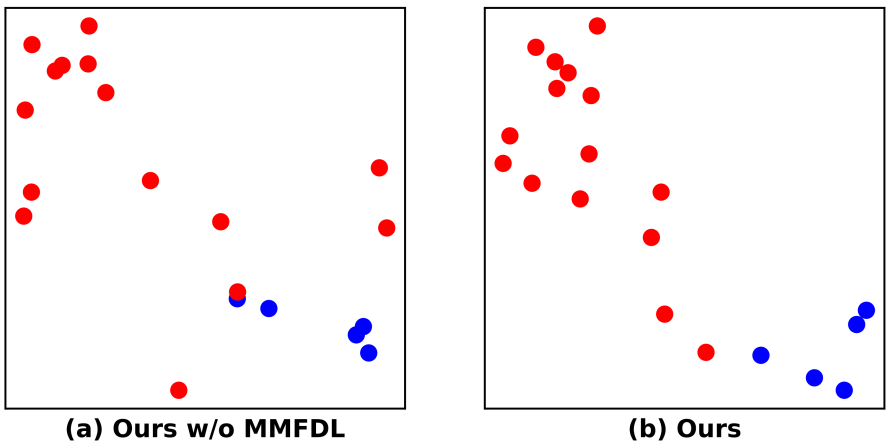

Fig. 2. Visualization of the context aggregated features of our model (a) without and (b) with the MMFDL method via t-SNE [6] on 01_0055 of ShanghaiTech. The blue dots denote normal regions while the red ones are abnormal.

most memory items are highly activated only by themselves, which indicates that the easy normal patterns stored by our ENPM are diverse. We visualize the context aggregated features of our model without/with the MMFDL method via t-SNE [6] in Fig. 2, which indicates the MMFDL method can help our model learn more discriminative features.

Table 2. Efficiency comparison on ShanghaiTech.

Method	#param.	Min FLOPs	Max FLOPs	IT	AUC
RTFM [10]	24.72M	3.21G	27.69G	1.50s	97.21
Ours	30.50M	3.96G	34.14G	1.05s	97.94

5 Efficiency Analysis

We use model size (#param), theoretical computational complexity (FLOPs), inference time (IT) and AUC value to evaluate the efficiency of the video anomaly detection methods. Since the input video is of different lengths, we calculate the maximum and minimum of FLOPs. The experiments are performed on one NVIDIA 3090 GPU. We compare our most similar model RTFM [10] and our method on pre-extracted I3D RGB features of ShanghaiTech, which is shown in the Table 2. Our model achieves a balance among model size, computational complexity, inference speed and AUC value.

6 Datasets

In this section, we describe UCF-Crime [9] and ShanghaiTech [3] datasets in detail.

6.1 UCF-Crime

This is a large-scale anomaly detection dataset that consists of 1,900 long untrimmed videos with 13 categories of anomalous events captured from real-world scenes by CCTV cameras. The training set of this dataset includes 800 anomalous and 810 normal videos and its testing set comprises 140 anomalous and 150 normal videos.

6.2 ShanghaiTech

This is a medium-scale dataset that contains 437 videos captured from 13 different scenes in a university campus. This dataset is originally built for semi-supervised anomaly detection. The training split only contains normal videos. Zhong *et al.* [19] reorganized the dataset for weakly supervised anomaly detection. This new training split of the dataset has 63 anomalous and 175 normal videos and its test split contains 44 anomalous and 155 normal videos.

References

1. Feng, J., Hong, F., Zheng, W.: MIST: multiple instance self-training framework for video anomaly detection. In: CVPR. pp. 14009–14018 (2021)
2. Li, S., Liu, F., Jiao, L.: Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In: AAAI. pp. 1395–1403 (2022)
3. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection - A new baseline. In: CVPR. pp. 6536–6545 (2018)
4. Liu, Y., Liu, J., Zhu, X., Wei, D., Huang, X., Song, L.: Learning task-specific representation for video anomaly detection with spatial-temporal attention. In: ICASSP. pp. 2190–2194 (2022)
5. Lv, H., Zhou, C., Cui, Z., Xu, C., Li, Y., Yang, J.: Localizing anomalies from weakly-labeled videos. IEEE TIP **30**, 4505–4515 (2021)
6. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9** (2008)
7. Pu, Y., Wu, X.: Locality-aware attention network with discriminative dynamics learning for weakly supervised anomaly detection. In: IEEE ICME. pp. 1–6 (2022)
8. Purwanto, D., Chen, Y., Fang, W.: Dance with self-attention: A new look of conditional random fields on anomaly detection in videos. In: ICCV. pp. 173–183 (2021)
9. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: CVPR. pp. 6479–6488 (2018)
10. Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J.W., Carneiro, G.: Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: ICCV. pp. 4955–4966 (2021)

11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS. pp. 5998–6008 (2017)
12. Wan, B., Fang, Y., Xia, X., Mei, J.: Weakly supervised video anomaly detection via center-guided discriminative learning. In: IEEE ICME. pp. 1–6 (2020)
13. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: ICCV. pp. 22–31 (2021)
14. Wu, P., Liu, J.: Learning causal temporal relation and feature discrimination for anomaly detection. IEEE TIP **30**, 3513–3527 (2021)
15. Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z.: Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: ECCV. vol. 12375, pp. 322–339 (2020)
16. Wu, P., Liu, X., Liu, J.: Weakly supervised audio-visual violence detection. IEEE Transactions on Multimedia (2022)
17. Zaheer, M.Z., Mahmood, A., Astrid, M., Lee, S.: CLAWS: clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In: ECCV. vol. 12367, pp. 358–376 (2020)
18. Zhang, J., Qing, L., Miao, J.: Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In: ICIP. pp. 4030–4034 (2019)
19. Zhong, J.X., 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: CVPR. pp. 1237–1246 (2019)
20. Zhou, H., Yu, J., Yang, W.: Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In: AAAI. pp. 3769–3777 (2023)