Prompt-Enhanced Multiple Instance Learning for Weakly Supervised Video Anomaly Detection

Supplementary Material

A. Appendix

The supplementary material is organized as follows:

- Section B provides more implementation details, including data pre-processing, hyperparameter settings, and training details.
- Section C provides hyperparameter analysis of prompt constraint loss weight and number of scales.
- Section D provides more qualitative results on datasets of XD-Violence [11] and UCF-Crime [9].
- Section E provides precision-recall curves.

B. Implementation details

Data Pre-processing. To extract video and audio features, we follow approaches used in existing methods [7, 10]. Videos features of 1024 dimensions are extracted from *global_pool* layer by RGB-stream I3D [1] video encoder, which is pre-trained on Kinetics [4] dataset. For 128-dimension audio features, we leverage VGGish [3] encoder pre-trained on YouTube [3] dataset. For computation efficiency, each video snippet consists of 16 frames. To ensure a fair comparison, we employ the same augmentation strategy as in [11]. For the UCF-Crime [9] and ShanghaiTech [6] datasets, we utilize a 10-crop augmentation strategy. This strategy involves taking crops from the center, four corners, and their mirrored counterparts. For the XD-Violence [11] dataset, we use a 5-crop augmentation strategy, which includes crops from the center and four corners. For text features, we use CLIP [8] text encoder to encode the class annotations to 512-dimension text features.

Hyperparameter settings. The hidden dimension D_h of temporal feature fusion module (TFF) is set to 128. And initial gate weight α of TFF is 0.5. The window size w of event attention is 9. The dimension D_m of the intermediate layer of MLP is set to 512, which is the same as the dimension of text features extracted by CLIP [8] text encoder. The length of normal context prompt (NCP) is 35 for XD-Violence and UCF-Crime, and 5 for ShanghaiTech. We set the causal convolution kernel size to 9, 3, and 3 for each dataset. In addition, the temperature coefficient τ is initialized to 0.09, 0.05, and 0.2 for each dataset respectively. The scaling factor μ for separation is set to 10. The model parameters are initialized by Xavier [2] uniform initialization.

Training Details. The weight λ and β are 1 and 8 respectively to balance the weights. The scales of s are set to 2 and 3. Weight λ with 0.001 is applied to balance multi-scale loss. During training, the batch size is set to 128 and the initial learning rate is 5×10^{-4} with a cosine decay strategy. The model is trained using Adam optimizer [5]. The total training epochs are 50. For balance between computational efficiency and detection performance, we set the snippet sampling threshold to 200 during the training phase. The epochs of NCP training are set to 10.

β	0	1	3	5	6	8	9
AP(%)	87.45	87.84	88.02	87.74	88.07	88.12	87.86

Table 1. Performance comparison of different prompt constraint loss weights.

$\mathcal{N}$	1	2	3	4
AP(%)	87.19	88.02	88.21	87.78

Table 2. Performance comparison of number of scales.

C. Hyperparameter Analysis

Effect of prompt constraint loss weight β. We evaluate the effect of prompt constraint loss by comparison between different hyperparameters β on XD-Violence, as shown in Table 1. When β is set to 0, the prompt constraint loss weight is not applied, resulting in an AP of 87.45%. As β increases from 0 to 1, the AP improves slightly to 87.84%. As β continues to increase, the AP shows a gradual improvement, reaching 88.12% when β equals to 8. The results prove the effect of prompt constraint loss. Prompt constraint loss can help learnable prompt to learn semantic-related and context-rich text feature effectively. Video features are then enriched by text features, facilitating accurate anomaly detection.

Effect of number of scales $\mathcal{N}$. In Table 2, we evaluate the effect of number of scales $\mathcal{N}$ on XD-Violence. When only a single scale is applied in the prediction head, the result is 87.19% in AP. As the number of scales increasing from 1 to 3, the AP gradually improves to 88.21, which shows the effect of scale-aware prediction head. The results demonstrate that scale-aware prediction head can benefit the model in learning multi-scale abnormal events. Therefore, the scale-aware prediction head can help to detect anomalies of different lengths, facilitating precise anomaly detection.

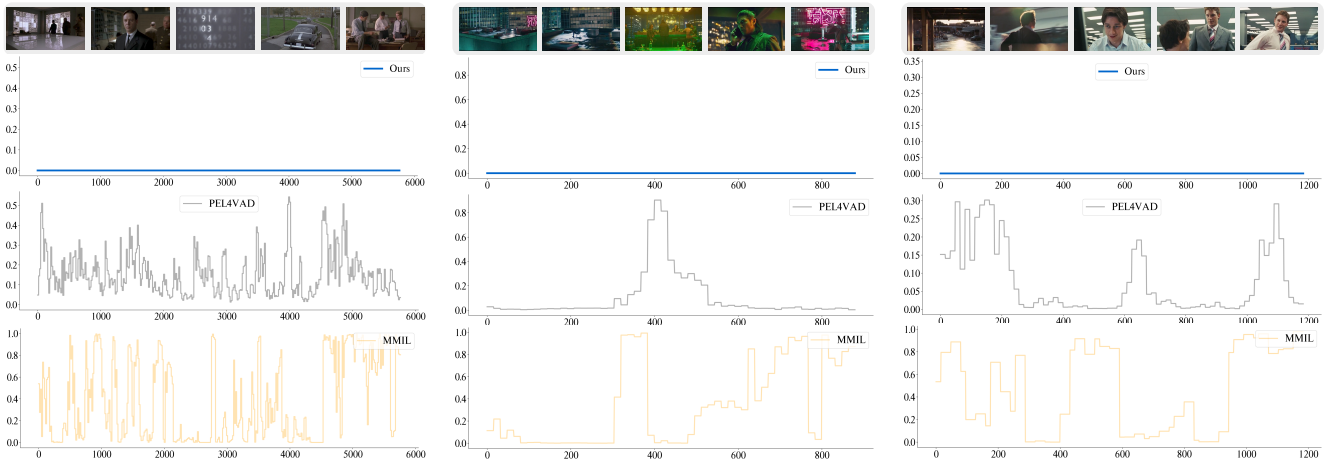
Figure 1. Anomaly scores of abnormal videos on XD-Violence. The Y-axis represents anomaly scores (1 for abnormal and 0 for normal), while the X-axis represents the frame number of videos. The pink area refers to the regions where anomalies take place. The blue lines are the predictions of our method; the gray lines are the predictions of PEL4VAD [7]; the orange lines are predictions of MMIL [9]. The above frames are snapshots from videos and the red ranges indicate the abnormal sections.

Figure 2. Anomaly scores of abnormal videos on UCF-Crime. The Y-axis represents anomaly scores (1 for abnormal and 0 for normal), while the X-axis represents the frame number of videos. The pink area refers to the regions where anomalies take place. The blue lines are the predictions of our method; the gray lines are the predictions of PEL4VAD [7]; the orange lines are predictions of MMIL [9]. The above frames are snapshots from videos and the red ranges indicate the abnormal sections.

D. Qualitative Results

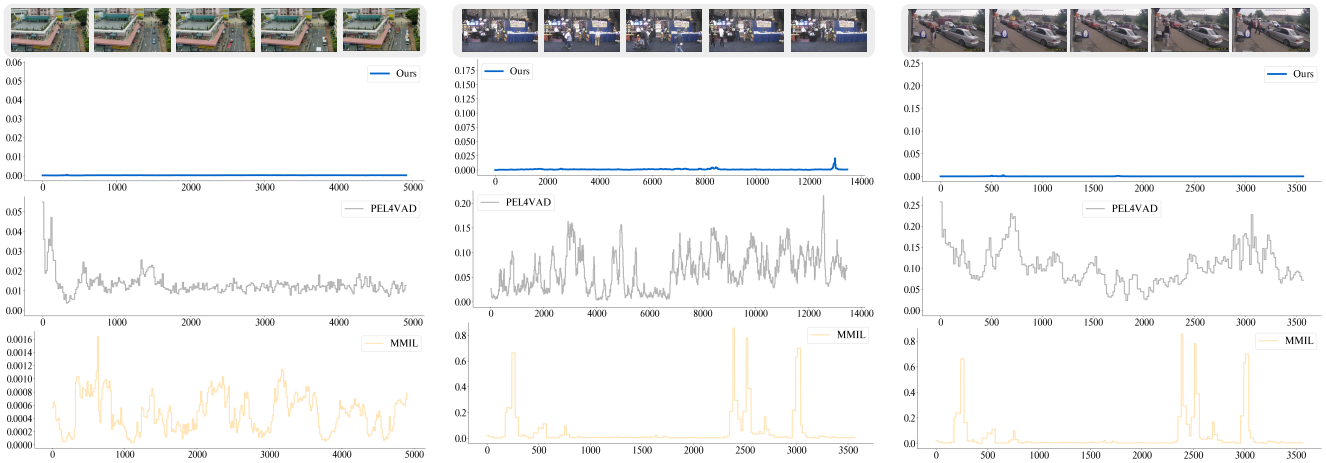
In Fig. 1, we visualize the detection results of SOTA methods in test set on XD-Violence [11]. The Y-axis represents anomaly scores, while the X-axis represents the frame number of videos. The pink area refers to the regions where anomalies take place. The blue lines are the predictions of our method; the gray lines are the predictions of SOTA method PEL4VAD [7]; the orange lines indicate the anomaly scores predicted by MMIL [9]. The above frames are snapshots from videos and the red ranges indicate the abnormal sections. When dealing with intricate scenarios with noisy backgrounds and ambiguous abnormal event boundaries, other methods generate noisy predictions with false alarms and enlarged boundaries. While, as shown in Fig. 1b, Fig. 1e and Fig. 1i, our method can not only locate anomalies accurately, but can detect the subtle normal intervals and generate clear event boundaries as well. As illustrated in Fig. 1c, Fig. 1d and Fig. 1g, our method presents the capability to detect various intricate abnormal events. Fig. 1a, Fig. 1f and Fig. 1h further demonstrate the ability to handle long-term videos and detect the abnormal events and event boundaries precisely.

In Fig. 2, we visualize the anomaly scores in test set of UCF-Crime [9]. As demonstrated in Fig. 2a and Fig. 2c, our model can detect subtle abnormal events in long-term videos. This demonstrates the capability of our method to model complex abnormal patterns in noisy background with hard temporal relationship. As Fig. 2b, Fig. 2d, Fig. 2e and Fig. 2f present, our method shows superior performance against the baseline. This shows the effectiveness of proposed abnormal-aware prompt learning which can facilitate



(a) (b) (c)

Figure 3. Anomaly scores of normal videos on XD-Violence. The Y-axis represents anomaly scores (1 for abnormal and 0 for normal), while the X-axis represents the frame number of videos. The blue lines are the predictions of our method; the gray lines are the predictions of PEL4VAD [7]; the orange lines are predictions of MMIL [9]. The above frames are snapshots from normal videos.



(a) (b) (c)

Figure 4. Anomaly scores of normal videos on UCF-Crime. The Y-axis represents anomaly scores(1 for abnormal and 0 for normal), while the X-axis represents the frame number of videos. The blue lines are the predictions of our method; the gray lines are the predictions of PEL4VAD [7]; the orange lines are predictions of MMIL [9]. The above frames are snapshots from normal videos.

the model learning various and complex abnormal patterns.

In Fig. 3, we present the prediction results of normal videos on XD-Violence. The normal videos in Fig. 3a, Fig. 3b and Fig. 3c are hard cases with intensive camera movement and dramatic scene changes, which are features that coupled with abnormal patterns. These characters of videos confuse the models [7, 9] to generate fuzzy predictions and aggravate false alarms. With the enhancement of normal context prompts, our method can accurately predict the anomaly scores for normal videos as demonstrated in Fig. 3a, Fig. 3b and Fig. 3c.

In Fig. 4, we show the prediction results for normal videos on UCF-Crime [9] dataset. The normal videos depicted in Fig. 4a, Fig. 4b, and Fig. 4c represent challenging cases characterized by intense camera movements and dramatic scene changes. These particular characteristics, which are often associated with abnormal videos, tend to confuse the models [7, 9] and regard the normal videos as abnormal ones. However, through the integration of normal context prompts, our method achieves precise anomaly score predictions for normal videos, as evidenced by the results shown in Fig. 4a, Fig. 4b, and Fig. 4c. The results demonstrate that the proposed normal context prompt can increase the discriminability of the ambiguous and fuzzy

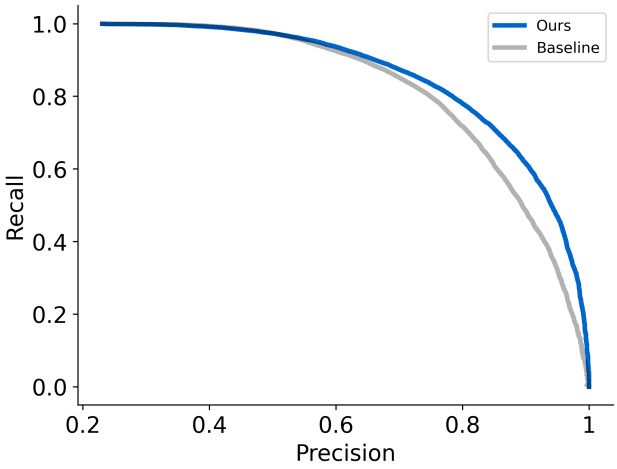


Figure 5. Precision-recall curves of our method and baseline model on XD-Violence. The blue line is the curve of our method, the gray line is the curve of baseline method without abnormal-aware prompt learning and normal-context prompt.

features, decreasing the rate of false alarm and improving the accuracy of anomaly detection.

E. Precision-Recall Curves

Fig. 5 shows the precision-recall curves of our method and baseline model on XD-Violence. The blue line is the curve of our method, the gray line is the curve of baseline method without abnormal-aware prompt learning and normal-context prompts. The AP score of our method is 88.21% and the AP score of the baseline method is 80.82%. As illustrated in Fig. 5, the performance is considerably improved when the proposed modules are added to the baseline, and there is a noticeable performance gap between the blue and gray lines. This demonstrates that the proposed abnormal-ware prompt learning can facilitate detecting various and intricate abnormal events, while simultaneously generating clear event boundaries, leading to a decrement in false alarm.

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. 1

[2] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 1

[3] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 1

[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[6] W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[7] Yujiang Pu, Xiaoyu Wu, and Shengjin Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *arXiv preprint arXiv:2306.14451*, 2023. 1, 2, 3, 4

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[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*, pages 6479–6488, 2018. 1, 2, 3, 4

[10] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30:3513–3527, 2021. 1

[11] 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*, page 322–339. 2020. 1, 3