Text Prompt with Normality Guidance for Weakly Supervised Video Anomaly Detection

Supplementary Material

1. Network Structure Details

TCSAL. The TCSAL module consists of 4 transformer-encoder layers with 4 attention heads per layer, and each self-attention head of each layer is self-adaptively adjusting its attention span by a soft-masking function $\chi_z(h)$. The shape of the soft mask function $\chi_z(h)$ is shown in Fig. 1.
Classifier. The classifier adopts a simple structure that consists of a layer normalization layer, a linear layer, and a sigmoid layer.

2. Finetuning and Prompt Learning

The finetuning of the CLIP text encoder is performed together with the training of the NVP, PLG, TCSAL, and Classifier modules. During this process, the weights of both the CLIP image and text encoder are frozen, except for the last projection layer of the text encoder which is unfrozen for finetuning.

To set the optimal fine-tuning configuration, we perform finetuning experiments on the final projection layers of the CLIP image encoder and text encoder as shown in Tab. 1. We find that based on prompt learning, there is a large performance improvement after fine-tuning the CLIP text encoder alone, whereas if we finetune only the final projection layer of the CLIP image encoder or both the image encoder and the text encoder at the same time, the performance instead decreases in both cases. Thus our final choice is prompt learning + finetuning (text encoder). We think this is due to the relatively small video anomaly dataset causing overfitting of the CLIP image encoder, which affects the method performance. When finetuning the text encoder alone, the overfitting situation is mitigated because of the prompt learning assistance. Finetuning also facilitates domain adaptation, so this combination of prompt learning + finetuning (text encoder) performs optimally.

CLIP finetuning and prompt learning configurations	UCF (AUC)	XD (AP)
No finetune + prompt learning	86.45%	81.33%
Image encoder finetuning + prompt learning	84.23%	81.16%
Text&Image encoder finetuning + prompt learning	85.76%	82.12%
Text encoder finetuning + prompt learning	87.79%	83.68%

Table 1. The AUC and AP change of our method on the UCF-Crime and XD-Violence datasets with different finetuning and prompt learning configurations.

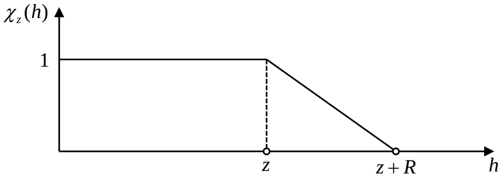


Figure 1. The shape of the soft mask function $\chi_z(h)$.

3. Training and Inference

Different from existing two-stage methods that separate pseudo-label generation and classifier self-training into two stages, in our approach, we synchronize pseudo-label generation and classifier training until both converge. This ensures that the updated pseudo-labels are used for supervised classifier training in real time, minimizing the interference of noisy labels on classifier training. After training under the supervision of the generated pseudo-labels, only the CLIP image encoder, the TCSAL, and the classifier are involved in the testing phase, where the video frame anomaly scores are predicted directly by the classifier.

4. Implementation Details.

Our method is implemented on a single NVIDIA RTX 3090 GPU using the Pytorch framework. We use Adam optimizer with a weight decay of 0.005. The batch size is set to 64, which contains 32 normal videos and 32 abnormal videos randomly sample from the training dataset. For the UCF-Crime dataset, the learning rate and total epoch are set to 0.001 and 50, respectively. For the XD-Violence dataset, the learning rate and the total epoch are set to 0.0001 and 20, respectively.

5. Impact of Normality Guidance Weight α.

The normality guidance weight α is used to control the degree of fusion of $\tilde{S}^{an}_{i,\,k}$ and $\tilde{S}^{aa}_{i,\,\tau}$ during pseudo-labels generation. In order to analyze the effect of α, we set different values of α for comparison experiments. As shown in Fig. 2, our method achieves optimal performance on both UCF-Crime and XD-Violence datasets when α is set to 0.2. It can be observed that as α gradually increases, the performance of our method gradually decreases, we consider that it is because too large α instead affects the alignment of the real anomaly event description text and the anomaly frames, and $\alpha = 0.2$ is the best trade-off.

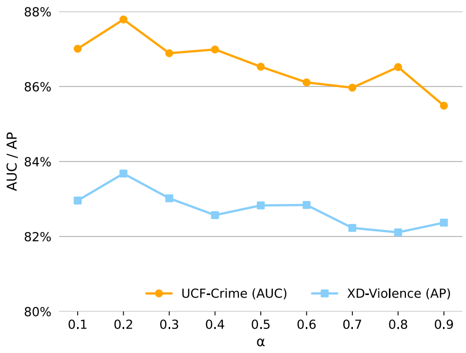


Figure 2. The AUC and AP change of our method on the UCF-Crime and XD-Violence datasets with different normality guidance weight α.

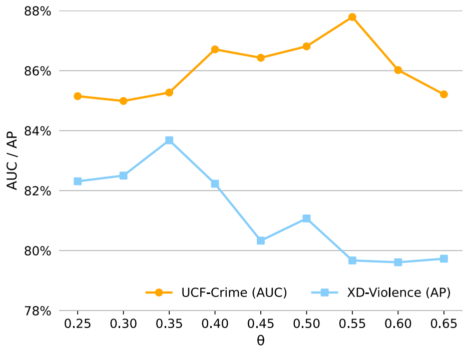


Figure 3. The AUC and AP change of our method on the UCF-Crime and XD-Violence datasets with different pseudo-label generation threshold θ.

6. Impact of Pseudo-label Generation Threshold θ.

To analyze the impact of different pseudo-label generation thresholds θ on the performance of our method, we set up a series of different thresholds θ to perform comparative experiments. As shown in Fig. 3, the two datasets have different sensitivities to the threshold θ. When θ is set to 0.55 and 0.35, our method achieves the optimal performance on UCF-Crime and XD-Violence datasets, respectively.

7. Impact of Context Length l in Learnable Prompt.

To investigate the optimal length of learnable textual prompts, we conduct comparative experiments with the context length l being set to 4, 8, 16, and 32, respectively. As shown in Tab. 2, both datasets achieve the best performance with the context length l set to 8, and slightly lower performance with a length of 16. However, when the context length l is set to 4 or 32, the performance of our method suffers a large degradation. We conjecture that the reason for this result is that too short a context length leads to textual prompts that do not fully characterize the video frame events, leading to model underfitting. Conversely, too long context length may lead to model overfitting.

l	UCF-Crime (AUC)	XD-Violence (AP)
4	82.26%	77.45%
8	**87.79%**	**83.68%**
16	87.24%	82.99%
32	85.23%	81.78%

Table 2. The AUC and AP of our method on the UCF-Crime and XD-Violence datasets with different context lengths l.

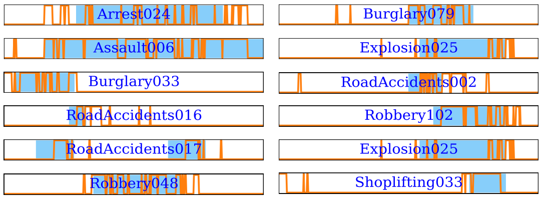


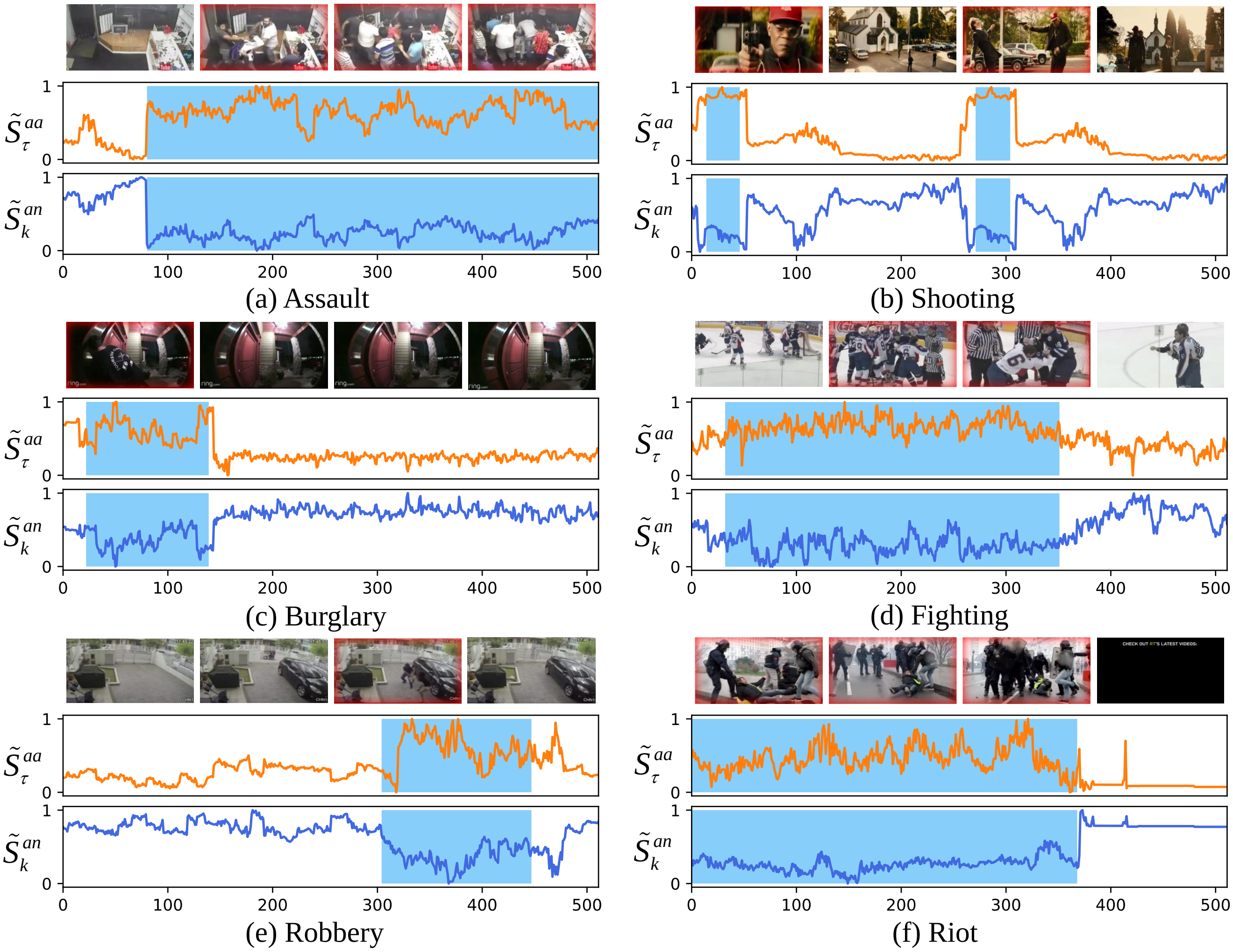
Figure 4. Visualization of pseudo-labels of some video clips on the UCF-Crime dataset.

8. Visualization of Pseudo-labels.

We visualize part of the pseudo-labels (UCF-Crime) in Fig. 4. The generated pseudo-labels (orange solid line) approximate the ground-truth (shades of blue) well in most cases, which indicates the effectiveness of the generated pseudo-labels.

9. Visualization of Match Similarities.

To more intuitively show that our constructed framework can facilitate the CLIP model to perform domain adaptation for matching video event text descriptions and corresponding video frames, we visualize the $\tilde{S}_{\tau}^{aa}$ and $\tilde{S}_{k}^{an}$, i.e., the match similarities of real abnormal event description text and normal event description text with corresponding abnormal videos, on the UCF-Crime and XD-Violence datasets, respectively. We can observe from Fig. 5 that the distributions of $\tilde{S}_{\tau}^{aa}$ and $\tilde{S}_{k}^{an}$ are contradictory which can align anomalous video frames and normal video frames, respectively. This shows the effectiveness of our designed distributional inconsistency loss L_{dil}. In addition, we can notice from Fig. 5 (a) and (f) that there are fluctuations in the alignment of the real abnormal event description text and the corresponding abnormal video frames in these two samples, while the normal event description text has a more accurate alignment, in which case our proposed normal guidance mechanism can assist $\tilde{S}_{\tau}^{aa}$ to better align the abnormal video frames.



(a) Assault

(b) Shooting

(c) Burglary

(d) Fighting

(e) Robbery

(f) Riot

Figure 5. Visualization of match similarities between video event description text and video frames for several anomaly samples on the UCF-Crime and XD-Violence test datasets. The light blue range represents abnormal ground truth.