Anomaly Detection for People with Visual Impairments Using an Egocentric 360-Degree Camera - Supplementary Metarials

Inpyo Song[1], Sanghyeon Lee[2], Minjun Joo[1], Jangwon Lee[1]
[1]Department of Immersive Media Engineering, Sungkyunkwan University
[2]School of Electronics and Information Engineering, Korea Aerospace University
{songinpyo, jmjs1526, leejang}@skku.edu, tkdgus4693@kau.kr

Dataset	Avg. frame count	Avg. abnormal frame count
UCF-Crime	7,247	603
XD-Violence	3,935	1,124
VIEW360 (Ours)	842	104

Table 1. Comparison of three weakly-supervised anomaly detection datasets by average number of frames.

1. Dataset Comparative Analysis

Figure 1 illustrates notable differences between our VIEW360 dataset and other widely used weakly-supervised anomaly detection datasets, UCF-Crime [3] and XD-Violence [6]. The primary differences between the datasets lie in the viewpoint and type of abnormal event. The majority of videos in the UCF-Crime and XD-Violence datasets are recorded using stationary cameras and sourced from diverse platforms like CCTV, movies, sports broadcasts, and video games. In contrast, our VIEW360 dataset leverages an egocentric 360-degree camera, offering a unique viewpoint for anomaly detection.

Moreover, the UCF-Crime and XD-violence datasets contain severe abnormal events, such as *arson*, *explosions*, *road accidents*, and *abuse*. In contrast, our VIEW360 dataset focuses primarily on the situations that visually impaired people often encounter in their everyday life, such as shoulder surfing (glance) and pickpocketing (stealing), which are relatively instantaneous. As shown in Table 1, our dataset also tends to feature shorter durations of abnormal events, reflecting our emphasis on quick, instantaneous abnormalities.

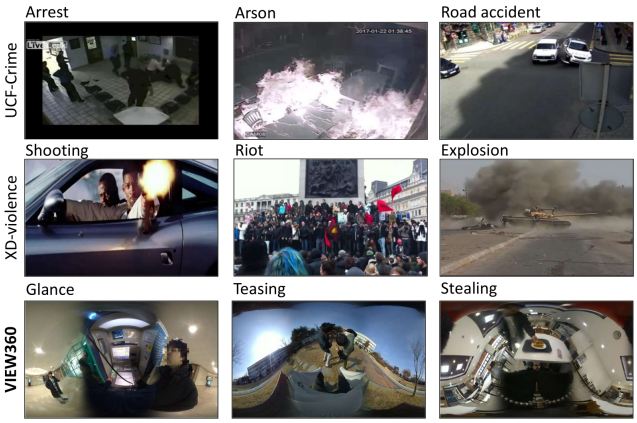

Figure 1. Sample images of the three datasets (UCF-Crime, XD-Violence and our VIEW360).

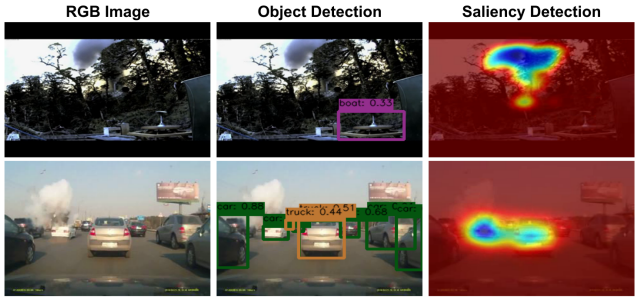

Figure 2. (**Top**) Object detection misses non-trained classes like fire smoke, while saliency detection successfully spots visually prominent areas. (**Bottom**) Object detection lacks relevance discernment for anomalies. Conversely, saliency detection highlights anomaly-involved regions by detecting active areas in the image.

2. Saliency-driven Image Masking

2.1. Saliency Detection vs. Object Detection

As discussed in the main paper's Table 4.b, Figure 2 demonstrates why saliency detection was chosen for image masking over object detection. Object detection, while effective for predefined classes, can miss critical objects not included in its training. Moreover, even when object detectors identify relevant classes, they often struggle to determine which detected objects are significant for anomaly detection. In contrast, saliency detection focuses on visually prominent areas, making it more adaptable and effective for anomaly detection. Its class-agnostic nature allows it to detect any significant regions, ensuring better identification of

Dataset	Saliency (%)		AD Perf. (AUC-ROC %)	
	Success	Failure	Success	Failure
VIEW360	98.54	1.46	85.89	85.53
UCF-Crime [25]	98.89	1.11	87.91	87.13

Table 2. This table shows the success and failure rates of saliency detection and the corresponding anomaly detection performance.

diverse anomalies.

Additionally, we provide more details of those experiments in the main paper's Table 4.b. We first detected objects in the videos with a threshold confidence score of 0.3. Then, we retained the detected bounding boxes' areas and masked the remaining areas. We replaced our Saliency-driven Image Masking with this object detector-based masking process.

2.2. Failure Case of Saliency Detection

We measured the impact of saliency detection on two evaluation datasets VIEW360 and UCF-Crime (Table 2). First, we manually annotated the ground truth (GT) bounding boxes of anomaly objects. Then, we applied the Saliency-driven Image Masking process. After this masking process, we measured the masked ratio of the GT bounding boxes. If the GT bounding box of anomaly objects is masked over the threshold, our masking process is considered a failure; otherwise, it is considered a success.

Initially, we set the threshold at 50% (i.e., does the GT bounding box remain over 50% after masking?). The success rate of the masking process was about 99.7%. This indicates that our masking process works quite well. However, for further analysis, we lowered the threshold to 30% (i.e., does the GT bounding box remain over 70% after masking?). Even with this stricter threshold, our masking process rarely failed, with failure rates of 1.46% in VIEW360 and 1.11% in UCF-Crime.

Moreover, when the masking process was imperfect, occasionally missing target objects, most failures were successfully compensated by neighboring frames. Therefore, the anomaly detection performance in failure cases was not significantly lower than in success cases. These results demonstrate that our masking process is quite robust.

2.3. Optimizing Saliency-driven Image Masking

In our research, we implemented a saliency-driven image masking approach that involved experimenting with different grid sizes and selecting the top-K salient regions. The objective was to find a balance that emphasized anomalies while retaining essential spatial information. Quantitative optimizing results are in main paper Table 5 about "Optimizing Image Masking". To further aid understanding and provide insights into our methodology, we present the variations in grid masking patterns that we explored in Figure 3.

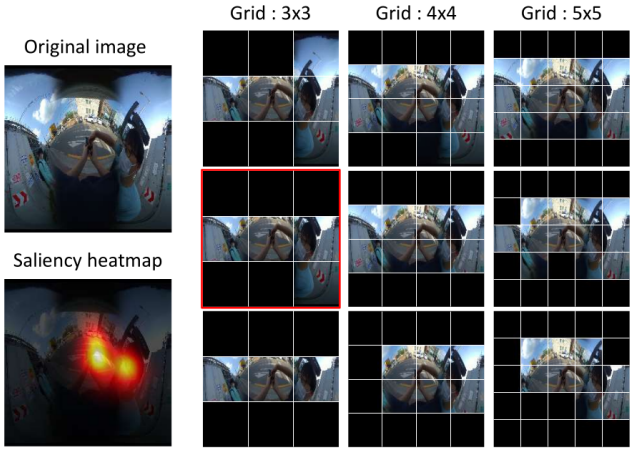

Figure 3. Variations of grid masking patterns evaluated in this research. The red box highlights the selected configuration of $n = 3$, $K = 4$ identified as the best balance for our approach.

Through extensive experimentation, we determined that a grid size of $n = 3$ with the top-4 ($K = 4$) salient regions provided the optimal solution.

3. FDPN: Architecture Details

In this section, we provide supplementary details regarding our architecture of the Frame and Direction Prediction Network (FDPN) employed in our study.

3.1. Prediction Subnetworks.

The Frame Prediction Subnetwork (FPS) and the Direction Prediction Subnetwork (DPS) are designed to enable effective information exchange across adjacent frames. Both FPS and DPS share a similar architectural design, inspired by PoolFormer [9], as shown in Figure 4. In this architecture, an average pooling operation (depicted in purple) with a kernel size of 3 is applied to the frame dimension. This operation serves to capture the temporal relationships between neighboring frames, providing insight into their sequential nature.

Following this, two 1D convolutions with a kernel size of 1 (illustrated in blue) are implemented to obtain an invariant representation across the channel dimensions. This design helps to extract essential features that are consistent throughout the channels. The final stage of the architecture consists of a 1D convolution with a kernel size of 5, paired with a Multi-Layer Perceptron (MLP) block (represented in green). This combination is responsible for calculating the anomaly score or predicting the direction of the anomaly, taking into account the relationships with surrounding frames.

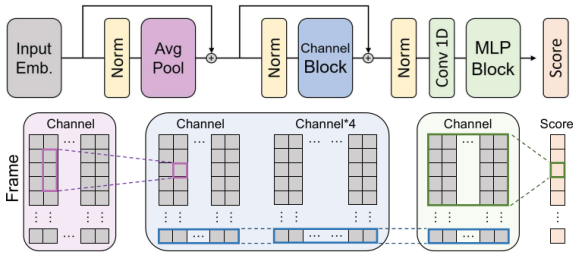

Figure 4. Our FDPN comprises two subnetworks: Frame Prediction Subnetwork (FPS) and Direction Prediction Subnetwork (DPS), each tailored to fulfill our two objectives: (1) detecting suspicious or abnormal activities in a 360-degree video stream and (2) determining their directional orientation. Both subnetworks share a PoolFormer-based architecture.

Dataset	VIEW360	UCF-Crime
RTFM	83.92	84.03
FDPN (w/ RTFM)	**86.00** (+2.08)	84.41 (+0.38)
MGFN	80.43	86.98
FDPN (w/ MGFN)	83.52 (+3.09)	**88.03** (+1.05)

Table 3. Performance evaluation of the FDPN architecture utilizing different Snippet Network configurations (RTFM and MGFN), as assessed on the VIEW360 and UCF-Crime datasets. The numbers represent AUC-ROC(%) values.

3.2. Choice of Snippet Network

In our methodology, the choice of the Snippet Network plays a crucial role, as it is tailored to suit the unique characteristics of the datasets under consideration. We thus conducted a series of experiments using different Snippet Networks on the two datasets we employed: VIEW360 and UCF-Crime [3]. The results of these experiments, presented in Table 3, guided our selection of the Snippet Network. The table compares the performance of two Snippet Networks, RTFM [4] and MGFN [2], when integrated with our Frame and Direction Prediction Network (FDPN). This comparison not only illustrates the effectiveness of each combination in anomaly detection but also highlights the enhancements achieved through FDPN. By enabling frame-level prediction, FDPN leads to more nuanced anomaly detection and demonstrates adaptability and effectiveness across different configurations and datasets. Other methods that used language-image pre-training model (CLIP) like VadCLIP [7] and TPWNG [8] was also considered. Althogh they well performs, lacks of video information because they depends on CLIP visual features. To intergrate our frame-level prediction mechanism, we found that built on methods which use video features more appropriate for our idea.

4. Challenges in Comprehensive Comparison

In the main paper's Tables 2 and 3, we aimed to provide a comprehensive comparison between several state-of-the-art methods. For the VIEW360 dataset results in main paper's Table 2, we were unable to include PE-MIL [1] and TPWNG [8] as their codes have not been published yet. Regarding the Shanghaitech dataset results in main paper's Table 3, VadCLIP [7] is a multi-modal method that requires anomaly class names for its operation. However, the Shanghaitech dataset lacks these specific annotations, making it impossible to apply VadCLIP accurately. Additionally, as previously mentioned, the code for TPWNG is not available. While there are a few methods we could not include due to these constraints, we have strived to ensure a thorough and fair comparison using the resources and information currently accessible.

5. More Qualitative Results

The Figure 5, 6 show the qualitative results of our proposed method compared to several state-of-the-art methods (RTFM [4], S3R [5], and MGFN [2]) on the VIEW360 and UCF-Crime [3] datasets. Each figure consists of two graphs, with each graph representing an individual video. These graphs display the anomaly scores, with light blue boxes in the background marking the ground-truth anomalous frames. Images corresponding to specific frames in the video are placed at the top of each graph, providing visual context.

Please refer to the following page for Figure 5 and 6 to see the detailed qualitative comparisons.

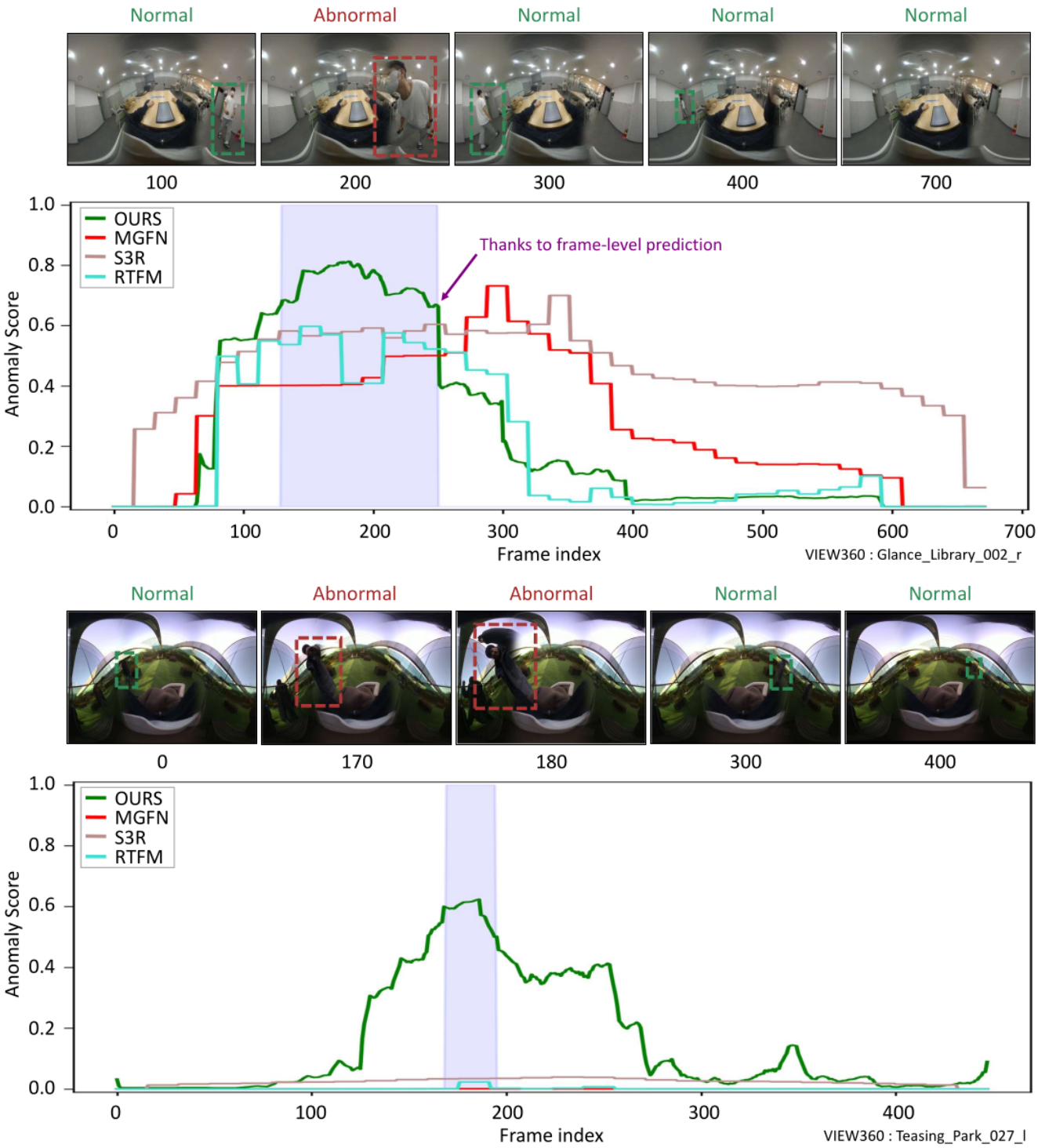

Figure 5. The results presented above highlight the effectiveness of the proposed approach in detecting anomalies in VIEW360 videos, even in the presence of distortion due to the nature of 360-degree images. Traditional snippet-level methods often struggle with varying information in each frame caused by distortion, but our frame-level prediction approach can overcome this challenge and detect anomalies despite distortion. It is worth noting that in the second result, where the distortion and anomaly occur for a very short duration, only the frame-level prediction approach was able to detect the anomalies, further emphasizing its effectiveness in detecting instantaneous anomalies.

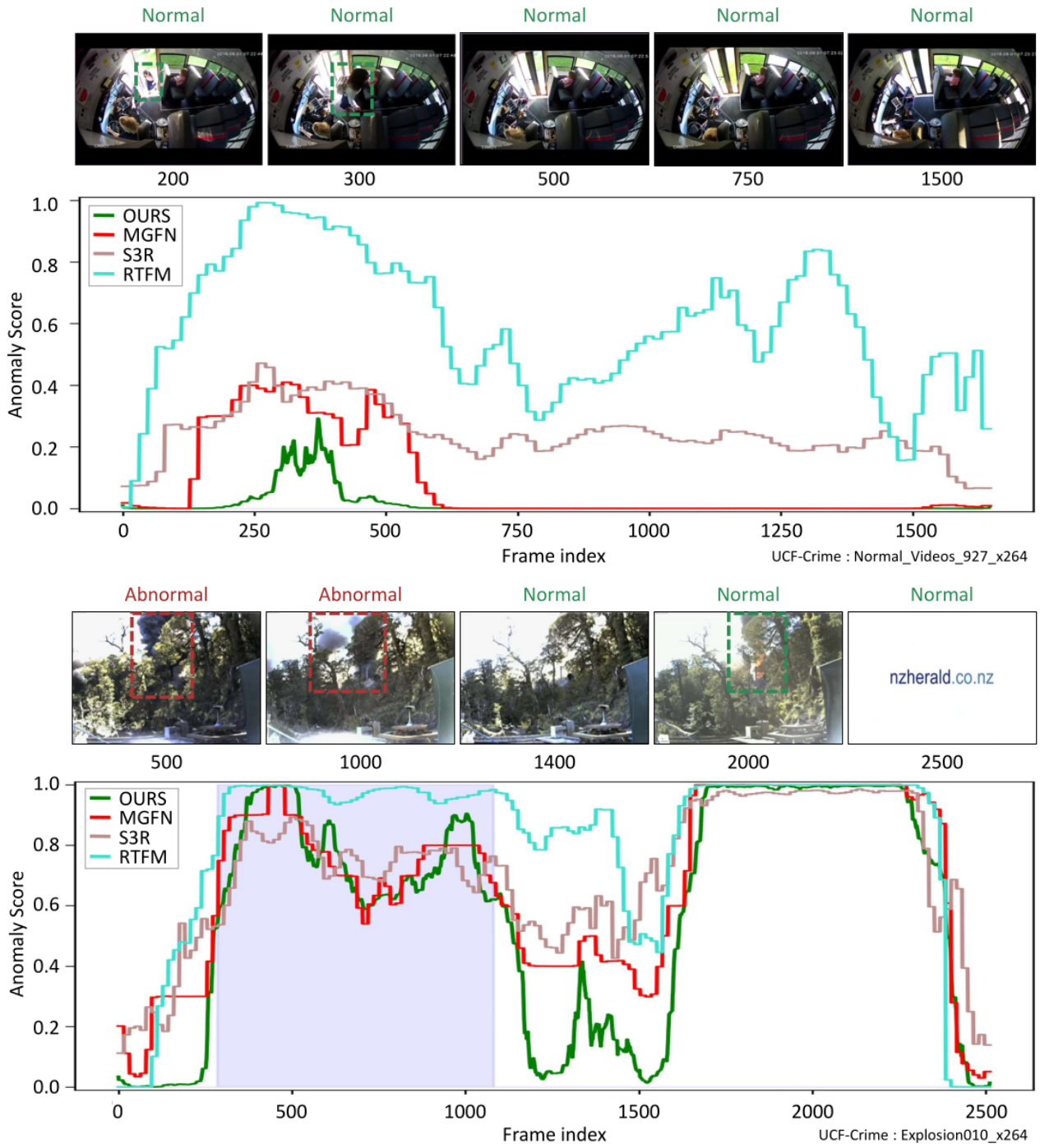

Figure 6. These two graphs show the experiment results of our approach compared to several state-of-the-art models for normal and abnormal situations on the UCF-Crime dataset. The top graph displays the anomaly scores of each method when they observe a normal event, where someone is riding a bus. Many of the SOTA approaches incorrectly judge it as an abnormal event, likely due to the sudden appearance of the person in the image frame. However, our method is able to correctly distinguish it thanks to the frame-level prediction. The bottom graph shows the scores when they observe a more complex video. This video contains two anomalous events: an explosion (abnormal), a pause (normal), and a fire (abnormal), although the ground truth labels in the video only contain one label for the explosion event. As you can see in the graph, only our approach can properly distinguish between those three situations over time thanks again to our frame-level prediction.

References

[1] Junxi Chen, Liang Li, Li Su, Zheng-jun Zha, et al. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In *CVPR*, 2024. 3

[2] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *AAAI*, 2023. 3

[3] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, 2018. 1, 3

[4] 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. In *ICCV*, 2021. 3

[5] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *ECCV*. Springer, 2022. 3

[6] 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 *ECCV*. Springer, 2020. 1

[7] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *AAAI*, volume 38, 2024. 3

[8] Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *CVPR*, 2024. 3

[9] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *CVPR*, 2022. 2