

## A Table of Contents

The supplementary material is organized as follows:

- Section B describes the construction process of the single-frame supervised VAD datasets.
- Section C provides additional dataset statistics.
- Section D provides annotation time estimation of different VAD paradigms.
- Section E provides detailed descriptions of the datasets.
- Section F outlines the baseline architecture.
- Section G presents the implementation details.
- Section H demonstrates additional experimental results.
- Section I provides further ablation studies.
- Section J provides hyperparameter analysis.
- Section K showcases additional qualitative results.
- Section L compares the difference between our method and related works.
- Section M discusses the limitation and further work.

## B Dataset Construction

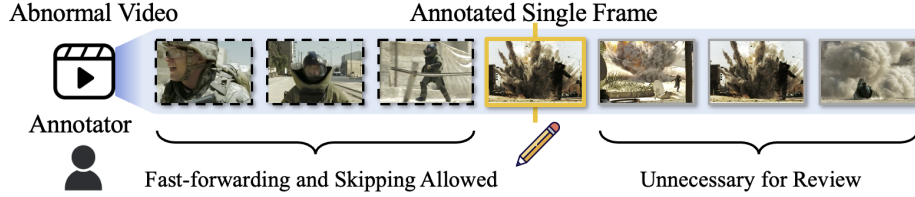


Figure 1: Illustration of single frame annotation procedure.

To adapt single-frame supervision for video anomaly detection (VAD), one of the primary challenges is the absence of appropriate datasets annotated with fine-grained frame-level labels. Most existing VAD benchmarks are formulated into semi-supervised and weakly-supervised settings, where only video-level ground-truth is provided, which falls short of the granularity required for frame-level supervision. Although Liu and Ma [12] offer frame-level annotations for the training set of UCF-Crime [21], the quality of annotation is sub-optimal, with omission of abnormal events and inexact localization of event boundaries, limiting the effectiveness for sampling frame-supervision from full annotations, as is commonly done in tasks, e.g., moment retrieval [6].

To address this limitation, we construct three high-quality, human-annotated Single-Frame supervised VAD (SF-VAD) datasets based on publicly available VAD benchmarks: ShanghaiTech Campus [13], UCF-Crime [21], and XD-Violence [28]. To maximize annotation efficiency while ensuring labeling accuracy, our SF-VAD datasets follow a practical single-frame annotation protocol that reflects how annotators behave in realistic labeling scenarios. Thereby, the constructed datasets not only enable the study of SF-VAD under realistic supervision constraints, but also reveal genuine human annotation preferences, offering valuable insights for developing methods that adapt to such real-world biases. Moreover, the protocol provides a scalable and cost-effective foundation for constructing large-scale SF-VAD benchmarks in the future.

Specifically, these SF-VAD datasets are annotated through a carefully designed crowdsourced annotation process where twelve human annotators participate. Before starting, annotators had to familiarize themselves with the definitions of various abnormal behaviors, such as abuse, riot, and shoplifting, and then pass a preliminary annotation test to ensure the annotation accuracy. Each annotator works independently, and cross-validation is conducted to ensure the consistency and quality of the annotations. As depicted in Fig. 1, to streamline the annotation process and ensure high accuracy, we provide annotators with the following guidelines: 1) Annotators are permitted to freely navigate the video timeline (e.g., via fast-forwarding or skipping) to identify potential abnormal events efficiently. 2) Annotators shall label exactly one frame per video, selected only when they are fully confident that the frame corresponds to an abnormal event. Once all individual annotations

are complete, a cross-verification process is performed to identify inconsistencies. Discrepancies between annotations are reviewed and corrected, ensuring the final annotations accurately reflect the frames where abnormal events occur.

## C Dataset Statistics

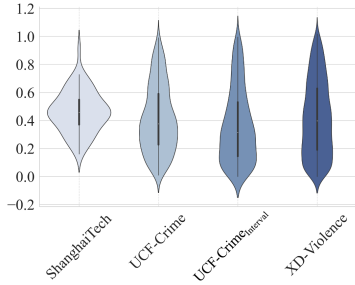


Figure 2: Violin plot of relative position of annotated frames. UCF-Crime<sub>Interval</sub> refers to relative position within abnormal events, while other entries indicate relative position w.r.t abnormal videos.

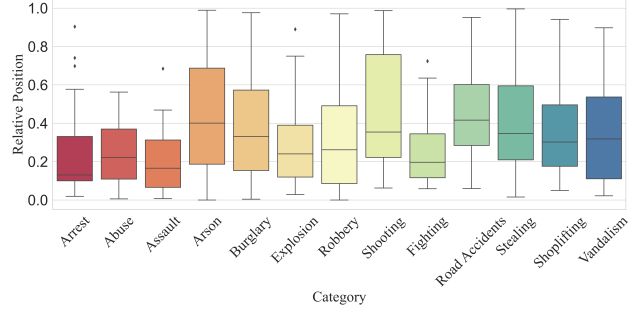


Figure 3: Box plot of the relative position of annotated single frames in UCF-Crime w.r.t different anomaly classes.

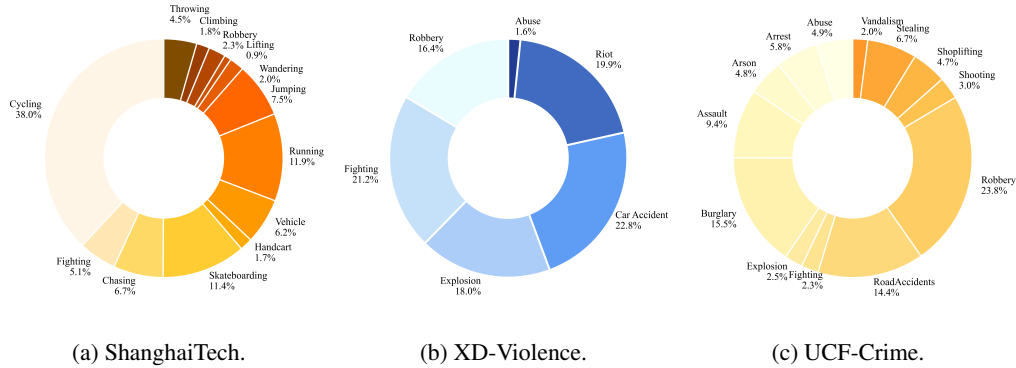


Figure 4: Proportion of total abnormal video duration accounted for each abnormal category across three VAD datasets.

This section provides a more detailed analysis of the SF-VAD dataset statistics, offering insights into the characteristics of the annotated frames. First, we illustrated the relative position of annotated frames within abnormal intervals and videos in Fig. 2. The distribution of annotated frames within the ShanghaiTech dataset exhibits a near-Gaussian distribution, with its peak centered around the middle of the video. This suggests that abnormal videos in ShanghaiTech tend to comprise clear pre-event, abnormal event, and post-event stages, with the abnormal events typically unfolding near the temporal center of the videos. For both UCF-Crime and XD-Violence datasets, the annotated frames are predominantly concentrated towards the earlier segments of the videos. This bias implies that initial frames in these datasets often contain critical cues indicative of an impending or ongoing anomaly, which also potentially leads to significant reductions in annotation time.

Furthermore, we visualize the relative position of annotated frames within anomalous events for various anomaly classes in the UCF-Crime dataset, as depicted in Fig. 3. Generally, the majority of annotated frames across all anomaly classes are indeed skewed towards the beginning of the video. Notably, for classes such as 'Abuse,' 'Assault,' and 'Fighting,' which typically involve rapid and drastic movements, the variance in the relative position of annotated frames is remarkably small. This concentrated annotation suggests that the critical distinguishing features for these events are often visually salient and emerge early in the temporal sequence. This observation also substantiates

64 SF-VAD’s efficiency, as it can direct annotators to these crucial early frames, thereby streamlining  
65 the annotation process without full video review.

66 Beyond the temporal distribution, we also analyze the proportion of total abnormal video duration  
67 accounted for by each abnormal category within the training sets of each dataset. As shown in Fig. 4,  
68 for UCF-Crime, certain anomaly classes, e.g., ‘Vandalism’ and ‘Shooting’, constitute a relatively  
69 minor proportion of the overall training data. Despite this limited representation, our SF-VAD method  
70 achieves refined detection results for these underrepresented classes, as shown in Sec. 4.4. This  
71 remarkable performance on ‘trivial’ or low-shot classes underscores the effectiveness of SF-VAD in  
72 providing fine-grained guidance to highlight subtle anomalies from the context. By leveraging the  
73 limited yet informative cues, SF-VAD demonstrates its capability to learn robust representations even  
74 from sparse data, which is a significant advantage in real-world anomaly detection scenarios where  
75 certain anomalies are inherently rare.

## 76 D Annotation Time Estimation

77 In practice, data annotation is a highly intricate process that encompasses not only the explicit  
78 time required for watching videos and assigning labels, but also a significant amount of additional  
79 effort that is often overlooked. This includes reviewing and replaying video segments to identify  
80 specific frames, verifying the temporal boundaries of anomalous events, rechecking annotations  
81 for consistency, conducting cross-validation, resolving annotation conflicts, and training annotators.  
82 Given the diverse and layered nature of these activities, accurately measuring the true annotation  
83 time becomes exceedingly difficult. Therefore, in this work, we estimate the annotation cost using a  
84 theoretical lower bound based on a set of practical assumptions. The annotation time versus detection  
85 performance is depicted in Fig. 1c in the main paper.

86 **Fully-supervised VAD** utilizes frame-level labels, which requires annotators to watch all videos  
87 from beginning to end at least once. Accordingly, the theoretical lower bound of annotation time  
88 is equivalent to the total duration of the dataset. In practice, however, the actual annotation cost is  
89 significantly higher due to the exhaustive temporal localization of abnormal event boundaries, which  
90 often necessitates frequent playback, meticulous inspection, and multiple rounds of verification to  
91 ensure temporal accuracy and consistency.

92 **Semi-supervised VAD** leverage normal videos only, however, the annotators need to watch the entire  
93 video snippets to make sure that the videos do not contain anomalies of any form. Therefore, the  
94 lower bound of annotation time equals the total duration of the normal videos in the dataset.

95 **Weakly-supervised VAD** uses video-level binary labels. For videos in the test set, the estimated  
96 annotation time is equivalent to the total duration of the test videos. For normal videos in the train set,  
97 the estimated annotation time equals the total duration as well, since the annotators need to watch the  
98 entire video to make sure it is a normal one. For abnormal videos in the training set, the estimated  
99 annotation time is estimated as the sum of time an annotator spends observing an abnormal frame  
100 within an abnormal video.

101 **Single-Frame supervised VAD** leverages single-frame annotation. Assuming that we elaborately  
102 devise an annotation platform, that enables the annotators to label the abnormal frame as soon as they  
103 identify one and let annotation proceed, the low bound annotation time is equal to weakly-supervised  
104 VAD. Notably, in piratical scenarios, the annotation time of single frame supervised VAD is slightly  
105 larger than weakly-supervised VAD, since single frame annotations involve playback from short  
106 anomalies and extra cross validation time to handle the conflict of annotations.

## 107 E Dataset Description

108 In this work, we construct SF-VAD benchmarks based on three widely-applied VAD datasets,  
109 ShanghaiTech Campus [13], UCF-Crime [21], XD-Violence [28], which cover broad range of  
110 abnormal behaviors, scene types, lengths and frequencies of abnormal events, and varying camera  
111 perspectives, as depicted in Tab. 1. The examples of annotated abnormal frames across various  
112 anomaly classes is depicted in Fig. 5.

113 **ShanghaiTech Campus** [13] comprises 437 videos from 13 fixed-view campus surveillance cameras.  
114 The abnormal types are cycling, chasing, cart, fighting, skateboarding, vehicle, running, jumping,



Figure 5: Illustrative examples of annotated abnormal frames across various anomaly classes.

wandering, lifting, robbery, climbing over, throwing. The background of the frames is rather steady and contains less noise, which highlights the behaviors within the frames.

**UCF-Crime** [21] comprises 1900 videos collected from a variety of sources including videos from surveillance cameras and social media with a total duration of 128 hours. The dataset covers 13 real-world anomalies of crimes including abuse, arrest, arson, assault, burglary, explosion, fighting, road accident, shooting, shoplifting, stealing, vandalism and robbery. The representations of the anomalies are varied and differentiated which increases the challenge of the detection by requiring a more thorough understanding of the anomaly semantics.

**XD-Violence** [28] is the largest and most challenging multi-modal VAD dataset containing 4754 untrimmed videos with a total duration of 217 hours. The dataset contains videos from various sources such as movies, social media, car cameras, surveillance, and games where exist extensive artistic expressions such as changing perspective, view zooming, dynamic lighting, and rapid camera movements. The above characteristics of the datasets draw non-negligible difficulty to anomaly detection models. It covers anomalies of 7 types including abuse, car accidents, explosions, fighting, riots, robbery and shooting.

Table 1: Comparison of video anomaly detection datasets

Dataset	Domin	#Videos	#Train Abn.	#Train Nor.	#Test Abn.	#Test Nor.	#Abn. Types	Resolution
ShanghaiTech [13]	Campus	437	63	175	44	155	13	856×480
UCF-Crime [21]	Crime	1900	810	800	140	150	13	Multiple
XD-Violence [28]	Violence	4754	1905	2049	500	300	7	640×360

## F Baseline

The architecture of the overall framework is depicted in Fig. 2 in the main paper. Concretely, given an untrimmed video, pertained feature encoders [18, 2] are employed to obtain multi-modal features. Subsequently, the features are passed through the Transformer-based Temporal Modeling (TTM) module and detector to predict frame-level anomaly scores. Building upon the temporal features, multi-layer convolutional networks are employed as evidence encoder and anomaly detector to predict evidence for relevance estimation and the final anomaly scores.

Considering the trade-off of computational overhead and detection performance, the input videos are split into 16-frame non-overlapping clips. Pre-trained frozen encoders are utilized to extract embedding features, formulating clip feature sequences. Embedding features are denoted as  $\mathbf{x} \in \mathbb{R}^{N \times D_m}$  where  $N$  equals the number of the clips and  $D_m$  is the dimension of the features.

Owing to resounding success in natural language processing areas, Transformer [25] has been verified as a highly effective architecture for capturing global dependencies. And it has been successfully employed in temporal modeling [32, 1]. Therefore, we apply TTM module, following [17], to capture multi-scale temporal cues for evidence and anomaly score prediction, as depicted in Fig. 6.



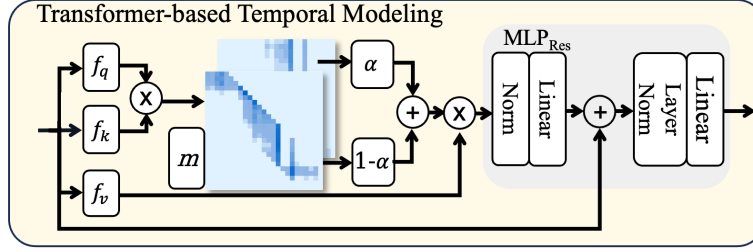


Figure 6: Architecture of Transformer-based temporal modeling module.

145 First, the attention mechanism’s similarity matrix  $\mathbf{m} \in \mathbb{R}^{N \times N}$  is computed with dynamic position  
 146 encoding  $\mathcal{E} \in \mathbb{R}^{N \times N}$  added to incorporate temporal position prior:

$$\begin{aligned} \mathbf{m} &= f_q(\mathbf{x}) \cdot f_k(\mathbf{x})^\top + \mathcal{E} \\ \mathcal{E}_{j,k} &= \exp(-|\gamma(j-k)^2 + \beta|) \end{aligned} \quad (1)$$

147 where  $f(\cdot)$  refers to linear layers and  $j, k \in [1, N]$  indicate index of clips.  $\gamma$  and  $\beta$  represent learnable  
 148 weight and bias. Then, global attention feature  $\mathbf{f} \in \mathbb{R}^{N \times D_h}$  is computed based on the similarity  
 149 matrix and the linear projection of  $\mathbf{x}$ . The process can be denoted as follows:

$$\mathbf{f} = \text{softmax}\left(\frac{\mathbf{m}}{\sqrt{D_h}}\right) \cdot f_v(\mathbf{x}), \quad (2)$$

150 where  $D_h$  indicates the hidden dimension. To highlight short-range temporal attention of events and  
 151 solve long-range noise, the similarity matrix is masked by a sliding window. The process can be  
 152 denoted as:

$$\widetilde{\mathbf{m}}_{ij} = \begin{cases} \mathbf{m}_{ij}, & j \in [\max(0, i - \lfloor \frac{w}{2} \rfloor), \min(i + \lfloor \frac{w}{2} \rfloor, N)] \\ -\infty, & \text{otherwise} \end{cases} \quad (3)$$

153 where  $w$  refers to the window size and  $\widetilde{\mathbf{m}}$  indicates local similarity matrix. Correspondingly, local  
 154 attention feature  $\widetilde{\mathbf{f}} \in \mathbb{R}^{L \times D_h}$  is computed by Eq. 2. Then, global and local features are fused by gate  
 155 weight  $\alpha$ . Subsequently, a residual connection is utilized followed by layer normalization to derive  
 156 temporal feature  $\mathbf{f}^t \in \mathbb{R}^{L \times D_m}$ , which can be formulated as:

$$\begin{aligned} \mathbf{f}^t &= f_o\left(\text{Norm}\left(\alpha \cdot \mathbf{f} + (1 - \alpha) \cdot \widetilde{\mathbf{f}}\right)\right) \\ \mathbf{z} &= \text{LayerNorm}(\mathbf{x} + \mathbf{f}^t) \end{aligned} \quad (4)$$

157 where  $\text{Norm}(\cdot)$  denotes a composite of power normalization [34] and L2 normalization. Eventually,  
 158 TTM acquires multi-scale temporal feature  $\mathbf{z} \in \mathbb{R}^{N \times D_m}$ . Eventually, multi-layer convolutional  
 159 networks are employed as evidence encoder and anomaly detector to predict evidence  $\mathbf{e} \in \mathbb{R}^{N \times 2}$  for  
 160 relevance estimation, which can be denoted as:

$$\begin{aligned} \text{MLP} &= \text{Dropout}(\text{GELU}(\text{Conv}(\cdot))) \\ \mathbf{e} &= \text{LeakyReLU}(f_t(\text{MLP}(\text{MLP}(\mathbf{z})))) \end{aligned} \quad (5)$$

161 where  $\text{Conv}(\cdot)$  refers to one-dimension convolution followed by GELU [9] and  $f_t(\cdot)$  represents  
 162 causal convolutional layer. LeakyReLU corresponds to the activation function [14]. Similarly, the  
 163 final anomaly scores  $\hat{\mathbf{y}} \in \mathbb{R}^N$  can be predicted as:

$$\hat{\mathbf{y}} = \sigma(f_t(\text{MLP}(\text{MLP}(\mathbf{z})))) \quad (6)$$

164 where  $\sigma$  indicates the sigmoid activation function.

## 165 G Implementation Details

166 **Feature Extraction.** To extract video features, we follow existing methods [27, 17, 28]. We apply  
 167 the I3D [2] video encoder that is pre-trained on Kinetics [10] dataset, to acquire video motion  
 168 features. I3D processes each video frame and aggregates temporal context over a sequence of frames,

enabling it to extract rich, motion-aware features from the video. Video features are extracted from *global\_pool* layer from the I3D encoder which is 1024 dimensions. To acquire video appearance features, we utilize CLIP [18](ViT-B/16) image encoder. CLIP extracts visual semantic features for each frame that generally focus on the overall appearance. The acquired appearance features contain 512 dimensions. For the trade-off of detection performance and computational overhead, each video is split into 16-frame non-overlapping clips. Notably, we employ a crop augmentation strategy to enhance the generalization ability. For UCF-Crime and ShanghaiTech datasets, we apply a ten-crop augmentation strategy, which includes crops from the center, four corners, and their mirrored counterparts. For XD-Violence dataset, we employ a five-crop augmentation strategy, which includes crops from the center and four corners.

**Hyperparameter.** The hidden dimension  $D_h$  of transformer-based temporal modeling module is set to 128. The initial gate weight  $\alpha$  of transformer-based temporal modeling module is set to 0.5. The window size  $w$  is set to 5, 9, and 9 for ShanghaiTech, UCF-Crime, and XD-Violence, respectively. The kernel size and stride of the one-dimensional convolutional layer  $f_t$  are set to 3 and 1, respectively. In abnormal event mining algorithm, the threshold  $\theta_1$  that filters the total variance of similarity is set to 0.1. The threshold  $\theta_2$  that controls the prominence of similarity of key frames is set to 0.96. The threshold  $\theta_3$  that controls the gap of abnormal events is set to 0.2.

**Training Details.** All experiments are conducted on a single NVIDIA RTX 3090 GPU using PyTorch. During training, the model parameters are initialized by Xavier initialization. The batch size is set to 128. The learning rate is  $5 \times 10^{-4}$  initially and controlled by a cosine decay strategy. The parameters are optimized using Adam optimizer. The number of training epochs is set to 50. For the balance between computational overhead and detection performance, the maximum sampling sequence length is set to 200 during the training phase.

## H Further Evaluation

We further evaluate our FPL framework and baseline models [24, 4, 37] under single-frame supervision and weak supervision. The results are depicted in Tab. 2. With the same feature setting, single-frame supervision enables RTFM [24], MGFN [4], and UR-DMU [37] to outperform their weakly-supervised counterparts across all datasets. The superior performance demonstrates the effectiveness of single-frame supervision, which provides precise anomaly cues that guide more accurate abnormal patterns modeling while reducing the impact of noisy or irrelevant frames. In addition, we further evaluate our baseline model under full supervision. Remarkably, our baseline model trained with single-frame supervision yields better results than with full supervision. We observe that the frame-level annotations [12] suffer from the omission of some abnormal events and inconsistent event boundary labeling, leading to a hardship to learn a coherent decision boundary by such inaccurate and strict constraints. In contrast, our single-frame supervision paradigm does not depend on the comprehensiveness of annotations or the precise localization of event boundaries. Instead, empowered by the proposed FPL framework, it progressively generalizes sparse, frame-level signals to entire abnormal intervals. This allows our method to achieve significantly superior performance while requiring substantially less annotation effort, highlighting its annotation efficiency and effectiveness. By leveraging single-frame supervision and our proposed FPL framework, we not only generalize precise supervision signals to broader abnormal intervals, but also effectively decouple the modeling of normal patterns, thereby achieving optimal detection performance.

## I Further Ablation Studies

To further validate the effectiveness of our proposed framework, we conduct extensive ablation studies on the UCF-Crime dataset. Tab. 3 presents the contributions of the two key procedures within the abnormal event mining algorithm, key frame selection and interval mining, as well as the normal behavior decoupling strategy, which separates the pre-event and post-event context. We observe that incorporating key frame selection alone brings performance improvement in terms of AUC, demonstrating the advantage of extracting similar key frames for more comprehensive anomaly pattern modeling. Moreover, interval mining further boosts the performance, as it enriches the temporal context around anomalies. When both components are activated, the performance reaches 90.23% in terms of AUC, showing their complementary effects.

Table 2: Performance comparison with state-of-the-art methods

Supervision	Methods	Text	Feature	XD(%)	SH(%)	UCF(%)
Fully-Supervised	ARG <sub>MM</sub> '19 [12]	-	NLN	-	-	82.0
	Our Baseline	-	I3D RGB	-	-	85.52
Semi-Supervised	SVM Baseline	-	I3D+VGGish	50.78	-	-
	SCR <sub>MM</sub> '20 [22]	-	-	-	74.70	72.7
	Conv-AE <sub>CVPR</sub> '16 [8]	-	I3D+VGGish	30.77	-	50.60
	LANP <sub>ECCV</sub> '24 [20]	-	I3D RGB	-	88.32	80.02
	MGE <sub>net</sub> <sub>MM</sub> '24 [30]	-	Video Swin	-	86.9	-
	AED-MAE <sub>CVPR</sub> '24 [19]	-	-	-	79.1	-
	MULDE [15] <sub>CVPR</sub> '24	-	Hiera-L	-	81.3	78.50
Weakly-Supervised	MIL-Rank [21] <sub>CVPR</sub> '18	-	C3D RGB	73.20	86.30	75.41
	CA-VAD <sub>TMM</sub> '21 [3]	-	I3D RGB	76.90	92.25	84.62
	RTFM <sub>ICCV</sub> '21 [24]	-	I3D RGB	77.81	97.21	84.30
	CRFD <sub>TIP</sub> '21 [27]	-	I3D RGB	75.90	97.48	84.89
	MSL <sub>AAAI</sub> '22 [11]	-	VideoSwin	78.59	97.32	85.62
	S3R <sub>ECCV</sub> '22 [26]	-	I3D RGB	80.26	97.48	85.99
	CMA-LA <sub>ICCECE</sub> '22 [16]	-	I3D+VGGish	83.54	-	-
	MACIL-SD <sub>MM</sub> '22 [33]	-	I3D+VGGish	83.40	-	-
	MGFN <sub>AAAI</sub> '23 [4]	-	VideoSwin	80.11	-	86.67
	UR-DMU <sub>AAAI</sub> '23 [37]	-	I3D RGB	81.66	-	<b>86.97</b>
	CU-Net <sub>CVPR</sub> '23 [35]	-	I3D+VGGish	81.43	-	86.22
	CoMo <sub>CVPR</sub> '23 [5]	-	I3D RGB	81.30	<b>97.60</b>	86.10
	PEL4VAD <sub>TIP</sub> '24 [17]	✓	I3D RGB	85.59	98.14	86.36
	VadCLIP <sub>AAAI</sub> '24 [29]	✓	CLIP	84.51	-	88.02
	HLGAtt <sub>CVPR</sub> '24 [7]	-	I3D+VGGish	<b>86.34</b>	-	-
	TPWNG <sub>CVPR</sub> '24 [31]	✓	CLIP	83.68	-	87.79
	RTFM <sub>ICCV</sub> '21 [24]	-	I3D RGB	77.37	94.32	82.80
	MGFN <sub>AAAI</sub> '23 [4]	-	I3D RGB	76.10	88.67	83.21
	UR-DMU <sub>AAAI</sub> '23 [37]	-	I3D RGB	82.58	90.51	86.38
Frame-Supervised	RTFM <sub>ICCV</sub> '21 [24]	-	I3D RGB	82.31	97.69	85.60
	MGFN <sub>ICCV</sub> '21 [4]	-	I3D RGB	81.27	94.52	85.23
	UR-DMU <sub>ICCV</sub> '21 [37]	-	I3D RGB	86.30	95.38	88.17
	<b>Ours</b>	-	I3D RGB	88.12	<b>98.41(+0.81)</b>	89.86
	<b>Ours</b>	-	I3D+CLIP	<b>89.56(+3.22)</b>	98.32	<b>90.23(+3.26)</b>

“\*” denote these methods are reproduced by the official codes on weakly-supervised and frame-supervised setting, respectively.

In addition to abnormal event mining, we investigate the effect of our normal decoupling strategy, which explicitly separates normal behavior into pre-event and post-event contexts. As shown in Tab. 3, using neither pre- nor post-event modeling results in 88.63% AUC on UCF. Introducing pre-event normal context decoupling alone yields a notable gain, indicating that learning normal patterns preceding anomalies helps suppress false positives. Alternatively, post-event normal context decoupling also provides performance improvement, suggesting that post-anomaly context also carries normal cues. The combination of both achieves the highest performance. The results indicate that decoupling normal context in both phases helps suppress pseudo anomalies and reveal more discriminative features, thereby enabling more precise anomaly detection.

In Tab. 4, we study the impact of different supervision paradigms. While the baseline under complete weak supervision only achieves 83.67% in terms of AUC. Gradually increasing the ratio of single-frame supervised training video leads to substantial performance improvements. The hybrid setting with 50% weakly-supervised and 50% single-frame annotations achieves 87.79% AUC. Remarkably, the fully single-frame supervised version reaches 90.23% AUC, demonstrating that concise but precise single-frame supervision is highly effective for anomaly localization. These results suggest that, with comparable annotation cost to weak supervision, single-frame supervision offers a more cost-effective solution by providing fine-grained anomaly cues that substantially improve anomaly localization performance.

Table 3: Ablation study of the key procedures within the abnormal event mining algorithm and normal decoupling strategy.

Abnormal Event Mining		UCF	Normal Decoupling		UCF
Key Frame	Interval Mining		Pre-event	Post-event	
-	-	85.13	-	-	88.63
✓	-	86.69	✓	-	89.83
-	✓	88.82	-	✓	89.05
✓	✓	90.23	✓	✓	90.23

Table 4: Ablation study of the ratio of training data under different supervision paradigms.

Paradigm	Weakly-supervised	Single-frame supervised	UCF
Weakly-supervised	100%	0%	83.67
Hybrid	75%	25%	85.36
	50%	50%	87.79
	25%	75%	88.51
Single-frame supervised	0%	100 %	90.23

## J Hyperparameter Analysis

**Effect of Threshold  $\theta_2$ .** In UCF-Crime and XD-Violence, we conduct a hyperparameter analysis to investigate the effect of the abnormal event mining threshold  $\theta_2$ , which controls the required prominence of feature similarity among selected key frames. A larger  $\theta_2$  enforces stricter similarity constraints, leading to the selection of more confidently abnormal frames. Conversely, a smaller  $\theta_2$  allows for more diverse but potentially noisier frames to be included. As shown in Fig. 7a, performance initially improves as  $\theta_2$  increases, benefiting from more precise supervision signals. However, overly large values of  $\theta_2$  may result in overly conservative frame selection, missing important abnormal cues and leading to performance degradation. Empirically,  $\theta_2 = 0.95$  achieves the best performance, striking a good balance between precision and coverage in selected key frames.

**Effect of Threshold  $\theta_3$ .** We further analyze the effect of threshold  $\theta_3$ , which controls the minimum temporal distance between selected abnormal key frames, thereby encouraging diversity among discovered abnormal events. A larger  $\theta_3$  enforces broader temporal separation, promoting exploration of distinct abnormal segments. As shown in Fig. 7b, moderate values of  $\theta_3$  improve performance by preventing supervision collapse into a single event, while overly large values may overlook densely occurring anomalies. As the results show, setting  $\theta_3 = 0.2$  yields the best performance, effectively balancing redundancy reduction and anomaly coverage.

## K Qualitative Results

To illustrate the effectiveness of our method, we further visualize the anomaly scores of some hard cases with background interference, noisy scenes, subtle abnormal behaviours, and varied anomaly durations, compared with MGFN [4], UR-DMU [37], and PEL4VAD [17].

Fig. 8a and Fig. 8b visualize the detection results on videos with anomalous events occurring at different temporal scales. Our dynamic anomaly event mining algorithm effectively captures anomaly patterns across varying durations by jointly leveraging anomaly relevance and feature similarity. As a result, it achieves robust detection performance across diverse temporal scopes and produces clear and well-aligned event boundaries. Fig. 8c and Fig. 8d present detection results on grayscale videos, where the anomalous behaviors are visually subtle and corrupted by significant noise. In such challenging settings, our model still accurately identifies the anomaly duration, attributed to the precise guidance provided by single-frame supervision. Unlike weakly-supervised approaches that rely on coarse temporal labels, the fine-grained supervision facilitates robust learning of discriminative features.

Compared with MGFN, UR-DMU, and PEL4VAD, our method demonstrates more precise temporal localization, effectively capturing the onset and end of temporal episodic anomalies, such as car accidents in Fig. 9a and shootings in Fig. 9b. These events typically occur and vanish rapidly, making them challenging to detect with weak supervision. Our method successfully localizes them without

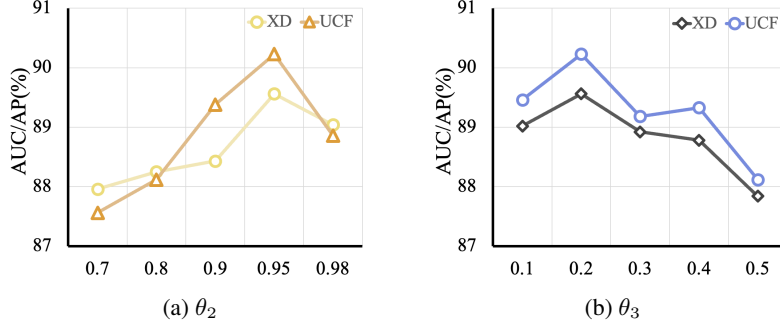


Figure 7: Hyperparameter analysis of  $\theta_2$  and  $\theta_3$  in abnormal event mining algorithm in XD-Violence and UCF-Crime.

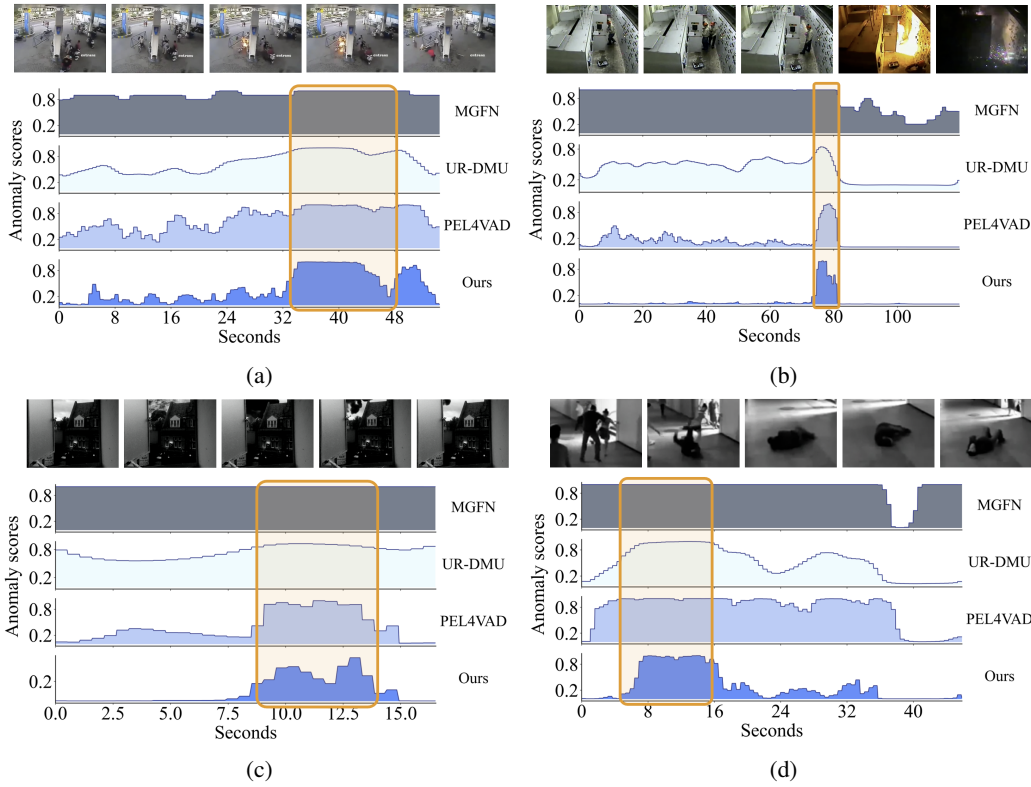


Figure 8: Visualization of anomaly scores in the UCF-Crime dataset. The Y-axis displays the anomaly scores, with 1 indicating abnormal and 0 indicating normal, while the X-axis shows the duration of the videos. The orange-shaded regions highlight the frames where anomalies occur. The frames above are snapshots from the videos. From top to bottom, the anomaly scores are generated by MGFN [4], UR-DMU [37], PEL4VAD [17], and Ours, respectively.

273 triggering excessive false alarms, benefiting from the proposed pre-event normal decoupling strategy,  
 274 which disentangles the contextual patterns preceding abnormal events. This decoupling enables the  
 275 model to distinguish normal fluctuations from truly anomalous changes. In Fig. 9c, we observe a  
 276 well-localized prediction for a sparse anomaly, alongside effective suppression of pseudo anomalies  
 277 in unrelated regions. Fig. 9d and Fig. 9e show ideal detection results on longer anomalous intervals,  
 278 while Fig. 9f demonstrates the ability to detect short anomalies embedded within long abnormal  
 279 periods. These results highlight our model’s ability to decouple fine-scale anomalies from extended  
 280 contextual sequences, significantly reducing false alarms.



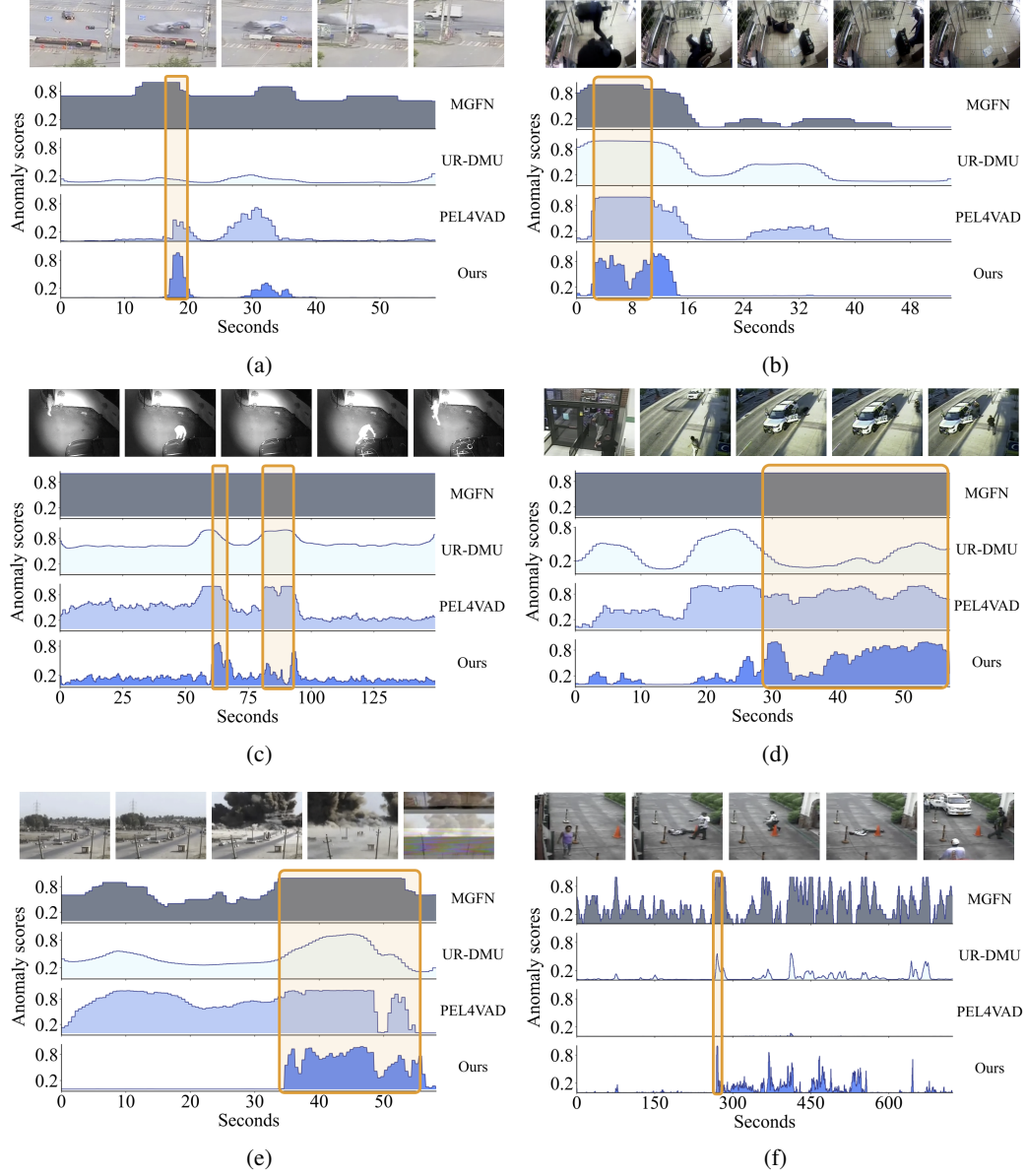


Figure 9: Visualization of anomaly scores in the UCF-Crime dataset. The Y-axis displays the anomaly scores, with 1 indicating abnormal and 0 indicating normal, while the X-axis shows the duration of the videos. The orange-shaded regions highlight the frames where anomalies occur. The frames above are snapshots from the videos. From top to bottom, the anomaly scores are generated by MGFN [4], UR-DMU [37], PEL4VAD [17], and Ours, respectively.

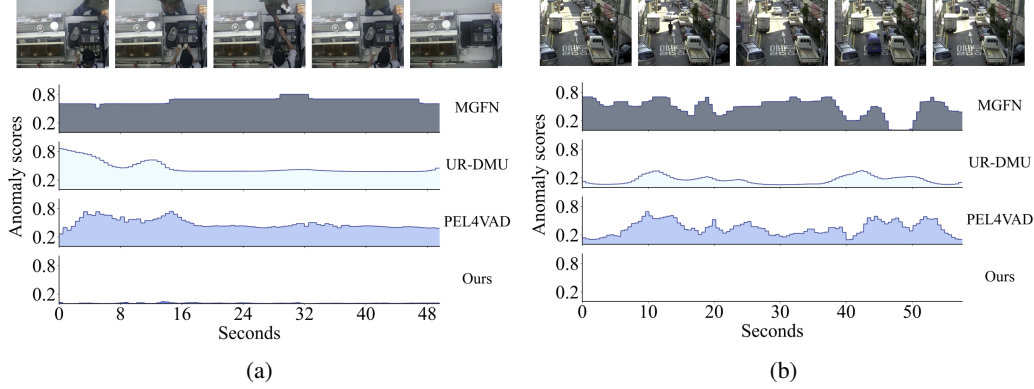


Figure 10: Anomaly scores of normal videos. Smaller anomaly scores indicate fewer false alarms and demonstrate a more reliable detection result. From top to bottom, the anomaly scores are generated by MGFN [4], UR-DMU [37], PEL4VAD [17], and Ours, respectively.

Fig. 10 shows detection results on normal videos with high visual similarity to anomalous cases, such as cashier scenes (visually similar to robberies) and traffic scenarios (resembling accidents). Our method yields nearly flat anomaly scores across the entire video, indicating strong confidence in normality. This performance benefits from the proposed pattern decoupling strategy, which explicitly separates abnormal patterns from high-frequency but non-anomalous behaviors. Unlike prior methods that often confuse visually similar contexts, our model learns semantically meaningful representations that generalize well to hard negatives, enabling accurate rejection of false positives in visually ambiguous settings.

## L Comparison with Related Methods

This section discusses the difference between our method with related works, including the supervision paradigm [36] and methods [38, 23].

Recently, Zhang et al. [36] study glance annotation in VAD, leveraging a frame annotation per abnormal event. Since multiple abnormal events may be involved in an abnormal video, such glance annotation typically requires *multiple* frame annotations per abnormal video, which imposes a high demand on the comprehensiveness of the labeling. On the one hand, this labeling process is more labor-intensive, as a full video review is necessary to ensure the completeness of the annotations. On the other hand, glance annotation requires precise temporal localization of abnormal events, as annotators must label each frame within distinct abnormal events, which necessitates validating both the onset and the conclusion of these events. As a result, the theoretic low bound of annotation time of glance supervision is close to that of fully supervision. In contrast, as depicted in Sec. C, our single-frame supervised paradigm increases annotation efficiency dramatically compared to glance annotation, as full video review and exhaustive temporal localization are not required in SF-VAD. As a consequence, the theoretic annotation time of single frame supervision is closed to weak supervision.

From a methodological perspective, glanceVAD [36] integrates UR-DMU [37] framework with temporal Gaussian splatting to identify static abnormal intervals, where the variance of Gaussian distribution is static as hyperparameter setting. In contrast, our Frame-guided Progressive Learning (FPL) takes anomaly relevance and feature similarity into consideration to dynamically prob the abnormal event intervals in a reliable way. In addition, FPL decouples normal context in abnormal videos to suppress false alarms, while significantly reducing the annotation burden.

Previous works [38, 23] employ evidential learning to solve VAD problems as well. Zhu et al. [38] integrate evidential learning to select reliable snippets by evidential learning to solve open-set VAD problems. Sun et al. [23] capture the deviation of normal samples as anomalies by evidential learning in semi-supervised VAD paradigm. Fundamentally, our FPL differs from previous methods in the following aspects. First, we leverage evidential learning to estimate anomaly relevance, where only annotated frame is involved to ensure a noise-free anomaly evidence learning process, replacing top-k sampling procedure that introduces noise and destabilizes the training. Second,

we leverage Beta distribution in evidential learning instead of Dirichlet distribution for VAD, as a binary classification problem. Third, to encourage relevance learning and predominant evidence, we incorporate regularization term  $\mathcal{L}^{KL}$ . As a result, we realize a reliable anomaly relevance estimation by evidential learning.

## M Limitation and Future Work

In the current approach, the extension from single-frame to multiple anomaly events relies primarily on feature similarity for anomaly detection. While this method shows promising results in the context of the experiments conducted, it places significant demands on the discriminative power of the features. As the complexity of the scenarios increases, the challenge lies in extracting more distinctive features that can effectively differentiate between various anomaly events. Moreover, the ability to reliably explore dynamic, continuous multiple anomalies over time remains an open issue. The model’s current formulation may not fully capture the temporal dependencies and interrelations between anomalous segments. Therefore, future work will focus on enhancing feature extraction techniques and developing more robust dynamic strategies to improve the model’s capability in detecting multiple anomalies in a continuous sequence.

In addition, current inexact supervision primarily focuses on the temporal dimension, leveraging frame-level annotations for anomaly detection. However, the spatial aspect remains relatively unexplored. A promising direction for future work is to extend this supervision to the spatial domain, where incorporating point-level supervision could highlight the anomalous objects or regions within each frame. By doing so, the model could achieve more accurate spatiotemporal anomaly localization, identifying both the occurrence and the specific spatial location of the anomaly. This would allow for a more granular understanding of abnormal events, further enhancing the model’s capability to detect and localize anomalies across both space and time. Therefore, future research will explore methods to integrate spatial cues into the existing framework to improve the robustness and precision of anomaly detection.

## References

- [1] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. 2021. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 565–574.
- [2] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.
- [3] Shuning Chang, Yanchao Li, Shengmei Shen, Jiashi Feng, and Zhiying Zhou. 2021. Contrastive attention for video anomaly detection. *IEEE Transactions on Multimedia* 24 (2021), 4067–4076.
- [4] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. 2023. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 387–395.
- [5] MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun Lee. 2023. Look Around for Anomalies: Weakly-Supervised Anomaly Detection via Context-Motion Relational Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12137–12146.
- [6] Ran Cui, Tianwen Qian, Pai Peng, Elena Daskalaki, Jingjing Chen, Xiaowei Guo, Huyang Sun, and Yu-Gang Jiang. 2022. Video moment retrieval from text queries via single frame annotation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1033–1043.
- [7] Ayush Ghadiya, Purbayan Kar, Vishal Chudasama, and Pankaj Wasnik. 2024. Cross-Modal Fusion and Attention Mechanism for Weakly Supervised Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1965–1974.

- [8] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 733–742.
- [9] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).
- [10] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).
- [11] Shuo Li, Fang Liu, and Licheng Jiao. 2022. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1395–1403.
- [12] Kun Liu and Huadong Ma. 2019. Exploring background-bias for anomaly detection in surveillance videos. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1490–1499.
- [13] W. Liu, D. Lian W. Luo, and S. Gao. 2018. Future Frame Prediction for Anomaly Detection – A New Baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [14] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, Vol. 30. Atlanta, GA, 3.
- [15] Jakub Micorek, Horst Possegger, Dominik Narnhofer, Horst Bischof, and Mateusz Kozinski. 2024. MULDE: Multiscale Log-Density Estimation via Denoising Score Matching for Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18868–18877.
- [16] Yujiang Pu and Xiaoyu Wu. 2022. Audio-guided attention network for weakly supervised violence detection. In *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*. IEEE, 219–223.
- [17] Yujiang Pu, Xiaoyu Wu, Lulu Yang, and Shengjin Wang. 2024. Learning Prompt-Enhanced Context Features for Weakly-Supervised Video Anomaly Detection. *IEEE Transactions on Image Processing* 33 (2024), 4923–4936.
- [18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [19] Nicolae-C Ristea, Florinel-Alin Croitoru, Radu Tudor Ionescu, Marius Popescu, Fahad Shahbaz Khan, Mubarak Shah, et al. 2024. Self-distilled masked auto-encoders are efficient video anomaly detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15984–15995.
- [20] Haoyue Shi, Le Wang, Sanping Zhou, Gang Hua, and Wei Tang. 2024. Learning Anomalies with Normality Prior for Unsupervised Video Anomaly Detection. In *European Conference on Computer Vision*. Springer, 163–180.
- [21] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6479–6488.
- [22] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. 2020. Scene-aware context reasoning for unsupervised abnormal event detection in videos. In *Proceedings of the 28th ACM international conference on multimedia*. 184–192.
- [23] Che Sun, Yunde Jia, and Yuwei Wu. 2022. Evidential reasoning for video anomaly detection. In *Proceedings of the 30th ACM International Conference on Multimedia*. 2106–2114.

- [24] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4975–4986.
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [26] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. 2022. Self-supervised sparse representation for video anomaly detection. In *European Conference on Computer Vision*. 729–745.
- [27] Peng Wu and Jing Liu. 2021. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing* 30 (2021), 3513–3527.
- [28] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference*. 322–339.
- [29] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. 2024. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 6074–6082.
- [30] Guoqing Yang, Zhiming Luo, Jianzhe Gao, Yingxin Lai, Kun Yang, Yifan He, and Shaozi Li. 2024. A Multilevel Guidance-Exploration Network and Behavior-Scene Matching Method for Human Behavior Anomaly Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 5865–5873.
- [31] Zhiwei Yang, Jing Liu, and Peng Wu. 2024. Text Prompt with Normality Guidance for Weakly Supervised Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18899–18908.
- [32] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. 2020. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII* 16. 507–523.
- [33] Jiashuo Yu, Jinyu Liu, Ying Cheng, Rui Feng, and Yuejie Zhang. 2022. Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection. In *Proceedings of the 30th ACM International Conference on Multimedia*. 6278–6287.
- [34] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 1821–1830.
- [35] Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. 2023. Exploiting Completeness and Uncertainty of Pseudo Labels for Weakly Supervised Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16271–16280.
- [36] Huaxin Zhang, Xiang Wang, Xiaohao Xu, Xiaonan Huang, Chuchu Han, Yuehuan Wang, Changxin Gao, Shanjun Zhang, and Nong Sang. 2024. GlanceVAD: Exploring Glance Supervision for Label-efficient Video Anomaly Detection. *arXiv preprint arXiv:2403.06154* (2024).
- [37] Hang Zhou, Junqing Yu, and Wei Yang. 2023. Dual Memory Units with Uncertainty Regulation for Weakly Supervised Video Anomaly Detection. *arXiv preprint arXiv:2302.05160* (2023).
- [38] Yuansheng Zhu, Wentao Bao, and Qi Yu. 2022. Towards open set video anomaly detection. In *European Conference on Computer Vision*. Springer, 395–412.