Advancing Video Anomaly Detection: A Concise Review and a New Dataset

Liyun Zhu[1*] **Lei Wang**[1,2] **Arjun Raj**[1] **Tom Gedeon**[3] **Chen Chen**[4]
[1]Australian National University, [2]Data61/CSIRO,
[3]Curtin University, [4]University of Central Florida
{liyun.zhu, lei.w, u7526852}@anu.edu.au,
tom.gedeon@curtin.edu.au, chen.chen@crcv.ucf.edu

Abstract

Video Anomaly Detection (VAD) finds widespread applications in security surveillance, traffic monitoring, industrial monitoring, and healthcare. Despite extensive research efforts, there remains a lack of concise reviews that provide insightful guidance for researchers. Such reviews would serve as quick references to grasp current challenges, research trends, and future directions. In this paper, we present such a review, examining models and datasets from various perspectives. We emphasize the critical relationship between model and dataset, where the quality and diversity of datasets profoundly influence model performance, and dataset development adapts to the evolving needs of emerging approaches. Our review identifies practical issues, including the absence of comprehensive datasets with diverse scenarios.[2] To address this, we introduce a new dataset, Multi-Scenario Anomaly Detection (MSAD), comprising 14 distinct scenarios captured from various camera views. Our dataset has diverse motion patterns and challenging variations, such as different lighting and weather conditions, providing a robust foundation for training superior models. We conduct an in-depth analysis of recent representative models using MSAD and highlight its potential in addressing the challenges of detecting anomalies across diverse and evolving surveillance scenarios. Our dataset is available here.

1 Introduction

Video Anomaly Detection (VAD) aims to automatically identify unusual occurrences in videos, enabling various applications in surveillance and monitoring [82]. Detecting anomalies is a challenging and complex task due to several factors: (i) There is no unified and clear definition of anomalies of interest. For example, the distinction between normal activities like walking on a sidewalk and abnormal activities like walking on a highway is context-dependent. (ii) The sporadic and rare occurrences of anomalies make the collection of well-curated datasets a demanding task, limiting the ability to learn anomalous patterns. Existing methods often treat VAD as either a one-class classification problem or an out-of-distribution detection task [4, 22, 36, 7, 24, 98, 20, 61, 27, 81, 50]. These approaches rely on training solely with normal samples and testing on both normal and ab-

[*]Liyun Zhu conducted this research under Lei Wang's supervision for a master's research project at ANU.

[2]In this paper, the term 'scenario' is used to signify different contexts such as retail, manufacturing, the education sector, smart cities, and others. In the context of different camera views, existing literature sometimes uses the term 'scene' to refer to the camera scene from a specific camera viewpoint.

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

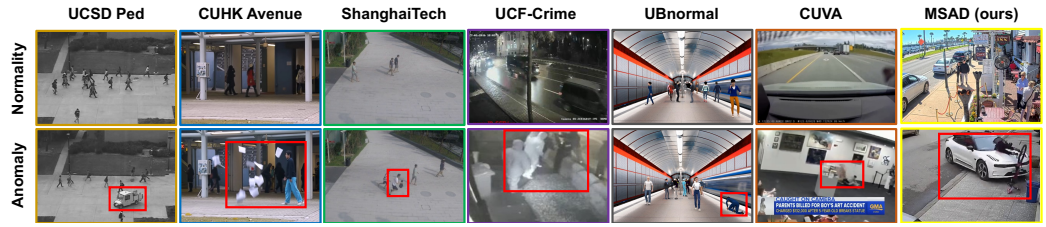


Figure 1: A comparison of existing datasets such as UCSD Ped, CUHK Avenue, ShanghaiTech, UCF-Crime, UBnormal and CUVA *vs*. our Multi-Scenario Anomaly Detection (MSAD) dataset.

normal samples, treating anomalies as outliers. Significant progress has been made in VAD during the past few years, thanks to benchmark datasets [35, 79, 37, 72, 14, 97], as well as synthetic data generation [2] that presents a range of anomalies. Fig. 1 shows some frames from different datasets.

Surveys. Advances in VAD have also led to some comprehensive surveys contributing significantly to the literature [71, 1, 69, 14, 28]. A survey on deep learning models for VAD can be found in [71]. A comprehensive study in [1] highlights challenging issues in VAD, such as varying environments, the complexity of human activities, the ambiguous nature of anomalies, and the absence of appropriate datasets. A systematic review in [69] discusses various potential challenges, opportunities, and provides guidance for future research. A recent work [14] offers a comprehensive benchmark for understanding VAD causation using Large Language Models (LLMs) for text annotations on human-related anomalies. While we focus on static cameras, a survey on VAD in dynamic scenes captured by moving cameras can be found in [28]. Although these surveys are comprehensive, they are not portable and lightweight. Our lightweight review offers several benefits: (i) it provides a quick reference for researchers and practitioners, making it easier to get up to speed on VAD without sifting through extensive details; (ii) it enhances accessibility for a broader audience, including newcomers to the field, by allowing a quick grasp of essential concepts; and (iii) it offers focused guidance by distilling critical information into clear, actionable insights.

Challenges. Various studies [103, 75] indicate that existing methods are often restricted to detecting only a handful of specific anomalies due to (i) a limited amount of videos and (ii) limited camera viewpoints, scenarios and anomaly types per dataset. These methods also frequently suffer from poor generalizability, necessitating retraining for each unique target camera viewpoint or new scenario, which subsequently increases computational costs [11]. Most existing methods [59, 15, 10, 97] are particularly vulnerable to various factors such as reflection, illumination changes, and complex background environments, leading to frequent false positives and negatives, thereby affecting the precision and reliability of detection. These challenges highlight the need for a high-quality, multi-scenario, and comprehensive dataset. However, existing benchmarks mostly focus on single-scenario (either single or multiple camera viewpoints), human-related anomalies. No existing works explore the multi-scenario generalization abilities of VAD from a practical perspective.

Motivations & contributions. Given the aforementioned challenges in VAD, along with practical problems often ignored by the community, we are motivated to compile this lightweight and insightful survey to bridge this gap and provide valuable insights to both interested readers and domain experts. Our survey begins with a discussion of existing VAD methods (Sec. 2.1), highlights the critical relationship between model and dataset development, as well as recent trends and potentials. Our concise review offers fruitful insights between models and datasets (Sec. 2.2), motivating us to contribute a new dataset (Sec. 3), Multi-Scenario Anomaly Detection (MSAD), to advance VAD. We conduct a deep analysis of recent representative methods using our dataset and demonstrate its potential for addressing the challenges of detecting anomalies across diverse and evolving surveillance scenarios (Sec. 4). Finally, we conclude our work (Sec. 5).

Table 1: Comparisons between our Multi-Scenario Anomaly Detection (MSAD) dataset and existing datasets. Our MSAD is the first large-scale, comprehensive benchmark for real-world multi-scenario video anomaly detection. It has 35 human-related anomalies (HRA) and 20 non-human-related anomalies (NHRA). Our dataset (i) comprises 14 distinct scenarios, including roads, malls, parks, sidewalks, and more (ii) incorporates various objects like pedestrians, cars, trunks, and trains, along with (iii) dynamic environmental factors such as different lighting and weather conditions.

Dataset	Year	Source	Domain	#Video	#HRA	#NHRA	#View	#Scenario	Modality	Resolution	Variations
Subway Entrance [3]	2008	Surveillance	Pedestrian	1	5	-	1	1	RGB	512×384	✗
Subway Exit [3]	2008	Surveillance	Pedestrian	1	3	-	1	1	RGB	512×384	✗
UMN [45]	2009	Surveillance	Behavior	5	1	-	3	1	RGB	320×240	✗
UCSD Ped1 [79]	2010	Surveillance	Pedestrian	70	5	-	1	1	RGB	238×158	✗
UCSD Ped2 [79]	2010	Surveillance	Pedestrian	28	5	-	1	1	RGB	238×158	✗
CUHK Avenue [35]	2013	Surveillance	Pedestrian	35	5	-	1	1	RGB	640×360	✗
ShanghaiTech [37]	2017	Surveillance	Pedestrian	437	13	-	13	1	RGB	856×480	✗
UCF-Crime [72]	2018	Online Surv.	Crime	1900	12	1	NA	NA	RGB	Multiple	✓
Street Scene [55]	2020	Surveillance	Traffic	81	17	-	1	1	RGB	1280×720	✗
IITB Corridor [63]	2020	Surveillance	Pedestrian	358	10	-	1	1	RGB	1920×1080	✗
XD-Violence [92]	2020	Films/Online	Violence	4754	5	1	NA	NA	RGB+Audio	640×360	✓
UBnormal [2]	2022	3D modeling	Pedestrian	543	20	2	29	8	RGB	1080×720	✓
NWPU Campus [6]	2023	Surveillance	Pedestrian	547	27	1	43	1	RGB	Multiple	✗
CUVA [14]	2024	News/Online	Multiple	1000	27	15	NA	NA	RGB+Text	Multiple	✓
MSAD (ours)	2024	Online Surv.	Multiple	720	**35**	**20**	**∼500**	**14**	RGB	Multiple	✓

2 A Concise Review

Our aim is to offer a lightweight reference for researchers and practitioners striving to advance VAD. Below we show the relationship between model and dataset development via a review on VAD methods. A review on VAD benchmark datasets is presented in Appendix G. Table 1 presents a comparison of existing datasets from various perspectives.

2.1 Dataset deficiencies and biases

From handcrafted to learned features. Early VAD methods use traditional techniques such as background subtraction, optical flow, and handcrafted feature extraction [44, 78, 65, 74], relying on appearance, motion, and texture to model motions and crowds [10, 96, 95, 5, 57]. However, these features are often insufficient due to the low resolution of benchmarks and limited training sources [3, 45, 79, 35], *e.g.*, Subway, UMN, UCSD Ped and CUHK Avenue, *etc*. Subsequent works [29, 32, 52, 89, 101] explore the combination of local/global features, spatial/temporal normalcy, *etc*. Nevertheless, these methods are ineffective when applied to different camera views and cannot adapt to unseen anomalies, when ShanghaiTech (13 camera views) is introduced. Consequently, emerging methods predominantly investigate the use of learned features, eliminating the need for handcrafted features and making them more adaptable to various camera viewpoints and new scenarios [36]. These methods mainly focus on creating new architectures or modules tailored to specific problems, and they can be broadly classified into four categories: reconstruction-based [4, 22, 27], prediction-based [33, 36, 20, 7, 61, 81], using classifiers [66, 47, 67, 68, 43, 27, 94], and scoring [30, 73, 26, 16, 38], ranging from simple architectures [82] to complex unified approaches [12].

Challenges of learned features. Most deep learning-based methods still carefully consider several aspects, such as visual appearance and motion, of human action [33, 20, 27, 86]. While these works highlight the importance of appearance and motion in detecting mostly human-related anomalies [20, 25, 27], non-human-related anomalies remain under-explored due to the lack of solid benchmarks, even when the UCF-Crime dataset (1 non-human-related anomaly) is introduced. Recent works [4] have delved into creating end-to-end deep learning approaches and unified architectures rather than using separate modules or components in a traditional pipeline, as end-to-end solutions are easily accessible, usable, and deployable. However, deep learning solutions require a significant amount of training data, posing a significant concern, especially considering older and smaller datasets such as

UCSD Ped [79] and CUHK Avenue [35]. Although efforts have been made in collecting large-scale datasets [72, 92] such as UCF-crime (video- and frame-level annotations for training and testing, respectively), an important issue lies in laborious video annotation, which is one of the main reasons why there haven't been as many large-scale VAD datasets published yet despite tons of data being publicly available on video sharing sites. A recent emerging few-shot learning framework [19] also encourages researchers to start looking at few-shot VAD models, due to (i) its fast adaptation to novel camera viewpoints/scenarios, (ii) relieving the training data hungry issue, and (iii) its huge potential in real-world applications. Since then, even smaller datasets have proven to be helpful [36, 39].

The beauty of few-shot learning. Few-shot learning aims to adapt quickly to a new task with only a few training samples. One of the most widely recognized methods in VAD is the Few-shot Scene-adaptive Anomaly Detection (FSAD) model [36]. This model uses the meta-learning framework [19] to train a model using video data collected from various camera views within the same scenario, such as ShanghaiTech. The model is then fine-tuned on a different camera viewpoint within the university site (*e.g.*, UCSD Ped and CUHK Avenue). Although the trained model can be adapted to novel viewpoints, its adaptability is still confined to a specific scenario, such as the university street example. The absence of a multi-scenario dataset hampers the widespread application of the rapid adaptation capabilities of VAD models. Recent few-shot VAD methods include [36, 91, 39, 84, 85].

Why self-supervised and weakly-supervised? Traditional supervised learning methods require labeled data, which is often scarce or expensive to obtain for anomalies, even though normal video data are easy to obtain. This is where self-supervised and weakly supervised methods come into play. Anomaly samples are difficult to obtain, and it is challenging to define all types of anomalies (Table 1 shows earlier datasets have very limited anomaly types); therefore, state-of-the-art VAD methods are rarely trained using fully supervised approaches [60]. Self-supervised methods are proposed based on the assumption that any pattern that deviates from the learned normal patterns will be considered anomalies. Representative methods include reconstruction-based [22, 39, 50, 98], prediction-based [33, 36, 20, 7, 61, 81], and distance-based [24, 27, 56] solutions. While previous studies have emphasized the importance of memorizing normal patterns [4, 22, 27], numerous studies [72, 13, 102] have indicated that the assumption underlying the self-supervised paradigm is not always valid: (i) It is impractical to obtain all normal types from diverse scenarios with varied distributions (*e.g.*, crowded streets *vs*. empty parking lots). (ii) The boundary between normal and abnormal behavior is often ambiguous; even the same abnormal behavior may lead to different detection results under different scenarios. Therefore, weakly supervised VAD has emerged, training on both normal and abnormal samples using video-level annotations, *e.g.*, on UCF-Crime. This approach avoids the issue of frame-level or pixel-level annotations, which are time-consuming and labor-intensive; however, video-level annotations only indicate that the video contains anomalies, leaving the exact start and end of the anomaly unclear. Existing methods in this category often use pre-trained models, such as TSN [87], C3D [77], I3D [9], SwinTransformer [34], *etc*., as an encoder to extract features.

Expanding modalities for human-related anomaly detection. Human-related anomaly detection solutions are not limited to the use of RGB videos. To efficiently extract motions, optical flow, precomputed on RGB videos, has been widely used as temporal or motion information for VAD. Due to the heavy computational cost of optical flow and the redundant nature of RGB videos, researchers have started to explore alternative data modalities for efficient feature extraction [88]. Human pose estimation frameworks, like OpenPose [8], have made human skeleton sequences readily available from RGB videos. This not only introduces a new data modality but also addresses privacy concerns in human-related anomaly detection. The lightweight nature of skeletons and informative spatio-temporal sequences have fostered many skeletal anomaly detection solutions [46, 25]. For example, [46] uses 2D human skeleton motions to detect anomalous human behavior in surveillance footage. Recently, LLMs and pretrained video caption models provide rich descriptions and prompts for video contents to assist VAD tasks [14, 40, 100]. Since then, multi-modal datasets have begun to emerge, with representative examples such as CUVA (RGB videos with text descriptions) in 2024.

Advancements in multi-modal fusion. A growing number of methods have begun integrating multi-modal information in recent years. Relying solely on optical flows [33, 83] or skeletons [24]

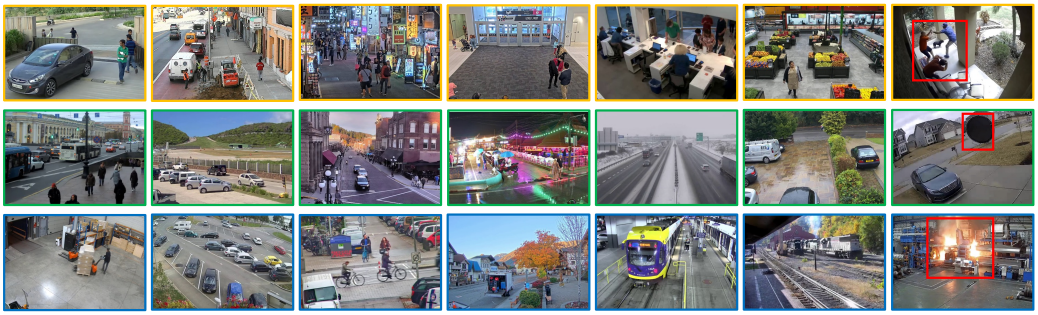


Figure 2: Our MSAD includes a diverse range of scenarios, both indoor and outdoor, featuring various objects, *e.g.*, pedestrians, cars, trains, *etc*. The first row shows different real-world common motions, while the second row demonstrates variations in weather and lighting conditions. The third row displays different moving objects. The last column shows human- and non-human-related anomalies.

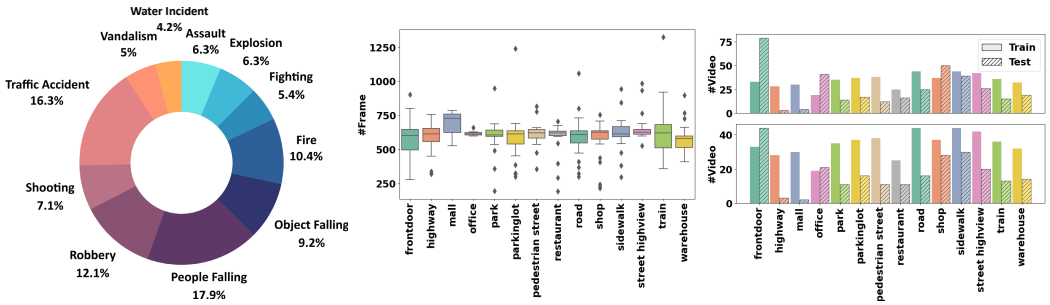
may not accurately detect anomalies for two main reasons: (i) motions may not reflect relevant anomalies, and (ii) non-human-related anomalies may be treated as background. To address these challenges, several methods have emerged to comprehensively understand anomalies. These include the development of an audio-visual violence dataset and a new fusion method [92], as well as the exploration of semantic information in VAD [12, 53, 93, 99]. One notable approach is a two-branch setup [93] (visual and language-visual branches) that utilizes a pre-trained visual-language model (*e.g.*, CLIP [54]). Additionally, the use of knowledge-based prompts as semantic information [53] and the fusion of text features extracted from a pre-trained video caption model with visual features [12] have been explored. However, compared to single-modality methods, the improvement from these fusion approaches remains somewhat limited, indicating the need for further refinement [62, 12].

2.2 Discussion on model and dataset evolution

Context-awareness. Detecting individual objects or actions without considering context can introduce bias or errors, given that anomalies are defined based on their contextual relevance. For example, the NWPU dataset introduced in [6] is designed to detect scene-dependent anomalies and aims to predict these anomalies in advance, thereby enabling early warnings [17]. The robustness of VAD to various environmental variations such as lighting, weather, and road conditions becomes increasingly crucial for real-world applications. However, most popular benchmark datasets such as UCSD Ped, ShanghaiTech, and CUHK Avenue only simulate anomalies using humans and do not consider diverse and complex environmental conditions. This limitation makes it difficult for well-trained models to effectively detect real-world anomalies. Our dataset accounts for these environment variations, making it better aligned with real-world applications.

Generalizability. Existing algorithms often restrict themselves to detecting a handful of specific anomaly types. These models frequently suffer from poor generalizability, necessitating retraining for each specific target scenario or even a particular camera viewpoint, which subsequently increases computational cost. Additionally, many techniques are particularly vulnerable to various external factors such as illumination changes and complex backgrounds (*e.g.*, tree swaying and rainy), resulting in frequent false positives or negatives and consequently lowering the precision and reliability of VAD. Collecting and labeling anomalies remains challenging due to their rarity and diversity. While synthetic anomaly data can be generated using game engines across multiple scenarios with precise pixel-level annotations, high-quality real data is still required. Along with the datasets, establishing relevant benchmark evaluation metrics, such as more comprehensive evaluations, allows researchers to compare different methods and drive the development of better-performing algorithms. This motivates us to collect a new comprehensive benchmark dataset with thoughtfully designed evaluation protocols.

Adaptability and reliability. VAD faces the challenge of evolving definitions of anomalies over time, even for a specific camera viewpoint. For example, anomalies during the daytime and nighttime



(a) The proportions of anomalies. (b) Variations in frame numbers. (c) Distributions of train/test splits.

Figure 3: The statistics of our MSAD dataset include: (a) a breakdown of main anomaly types and their respective percentages, (b) a boxplot illustrating frame number variations across scenarios in MSAD training set, and (c) the distributions of train/test splits across scenarios for two evaluation protocols (see Sec. 3 evaluation protocols): (*top* plot) generalizability and adaptability, and (*bottom* plot) practical applicability and effectiveness.

could differ, as could anomalies during workdays and weekends, necessitating adaptable and robust detection systems. These systems must continuously adapt to new video signals to maintain long-term reliability and accuracy, whether in single-scenario multiple viewpoints or multi-scenario settings. The robustness of VAD systems is significantly affected by the quality and size of the datasets used in training. High-quality, expansive datasets, combined with advanced algorithms, are essential to address the evolving nature of anomalies in VAD. Our experiments show the potential of our newly introduced dataset in tackling the challenges of detecting anomalies across diverse and dynamic surveillance scenarios. Notably, our dataset includes longer videos that capture the long-term evolution of video signals as well as anomalies.

Interpretability and privacy concerns. Below we list several limitations we observe during our review. First, only a few datasets in VAD contain multimodal data, which limits the development of relevant multimodal methods. XD-Violence introduced audio information and demonstrated the positive impact of audio-visual fusion. Therefore, exploring multimodal datasets and better fusion strategies that maximize the benefits of each modality could be a promising research direction. We plan to extend our dataset to include more modalities, such as audio and video descriptions. Second, most VAD models adopt a data-driven, end-to-end pipeline. Although effective, the learned features are often not interpretable, which may hinder deployment in real-world applications due to security and safety concerns. Developing explainable and interpretable models provides insights into the underlying reasoning for detected anomalies, making the automated system easier for end users to accept. Lastly, it is crucial to address privacy and ethical concerns associated with VAD, particularly regarding data collection in public spaces for surveillance. Balancing VAD performance while removing personally identifiable information, such as faces, poses, or gaits of individuals, and license plates of vehicles, would be an interesting area of exploration. We leave these for future work. More discussions are provided in Appendix E.

3 A Multi-Scenario Dataset

Unlike many datasets with fixed camera viewpoints and limited scenarios, our Multi-Scenario Anomaly Detection (MSAD) dataset boasts a broader range of scenarios and camera viewpoints (refer to Table 1 for a comparison). The process for collecting our dataset is detailed in Appendix A.

Viewpoints *vs*. scenarios. Traditional datasets commonly define a scene as the perspective captured by a camera, often referred to as a camera viewpoint. For instance, ShanghaiTech [37] comprises 13 scenes, representing videos recorded from 13 distinct camera viewpoints at a university. However, relying solely on the 'scene' or 'camera view' concept is insufficient for robust anomaly detection. Models trained on multi-view videos per scenario face limitations and struggle to adapt to new

scenarios, particularly when confronted with novel camera viewpoints. To overcome this limitation and enhance multi-scenario anomaly detection, we introduce the 'scenario' concept to describe different environments. Our dataset encompasses 14 distinct scenarios, including front door areas, highways, malls, offices, parks, parking lots, pedestrian streets, restaurants, roads, shops, sidewalks, overhead street views, trains, and warehouses. Fig. 2 showcases some video frames from these diverse scenarios. As depicted in the figure, our dataset features more realistic scenarios compared to existing benchmarks. It covers a broad spectrum of objects and motions, along with multiple variations in the environment, such as changes in lighting, diverse weather conditions, and more.

Human-*vs*. non-human-related anomalies. Previous studies have primarily characterized anomalies as human-related behaviors, encompassing activities like running, fighting, and throwing objects. This emphasis on human-related anomalies stems from their greater prevalence in real-world scenarios. However, enumerating all types of anomalies in diverse real-world contexts poses a significant challenge. To address this, our dataset is further categorized into two principal subsets: (i) Human-related: This subset features scenarios where only human subjects engage in activities, facilitating human-related anomaly detection. For instance, scenarios involve people interacting with various objects like balls or engaging with vehicles such as cars and trains. (ii) Non-human-related: This subset includes scenarios related to industrial automation or smart manufacturing, denoting the use of control systems for operating equipment in factories and industrial settings with minimal human intervention. This subset is designed for detecting anomalies, *e.g*., water leaks, fires, *etc*.

Dataset statistics. Figure 3 provides statistics for our MSAD dataset. Our dataset features a wide range of anomalies, including 35 human-related anomalies such as people falling down, school fights, street fights, facility vandalism, and shop robberies, as well as 20 non-human-related anomalies like water leaks, floods, factory fires, trees falling down, and office fires. Figure 3a illustrates the proportions of main anomalies in the dataset. Note that we group anomalies such as street fights and school fights into the main category "Fighting" for better visualization purposes. We provide all detailed anomaly types in Appendix B. Our dataset contains 720 videos (447,236 frames in total), with an average video length of 621.16 frames, and some videos extending up to 6,026 frames. The frames are extracted from the original videos at a rate of 30 FPS.

Evaluation protocols. To ensure a fair comparison between different algorithms, we provide frame-level annotations during testing stage. More detailed information can be found in Appendix B. We design two evaluation protocols for using our MSAD dataset:

i. Train on 360 normal videos from 14 scenarios and test on the remaining 120 normal videos and 240 abnormal videos. Figure 3b shows a boxplot of frame number variations across scenarios in training set. This protocol is suitable for evaluating self-supervised methods.

ii. Train on 360 normal and 120 abnormal videos, and test on 120 normal and 120 abnormal videos. During training, we only provide video-level annotations. This protocol is suitable for evaluating weakly-supervised methods trained with our video-level annotations.

We also apply Protocol i to both Few-shot Scene-adaptive Anomaly Detection (FSAD)[36] and our Scenario Adaptive Anomaly Detection (SA^2D). A detailed description of our SA^2D is provided in Appendix D. Our method focuses on evaluating how MSAD can contribute to fast adaptation not only to new camera viewpoints but also to new scenarios in VAD.

The design of these two evaluation protocols aims to facilitate a comprehensive evaluation by considering different models and training schemes. Such a design has not been considered before. Our dataset usage and maintenance are detailed in Appendix C. Figure 3c (*top* plot) shows the train/test split distributions for Protocol i, while (*bottom* plot) shows the distributions for Protocol ii.

4 Experiment

In this section, we design two main sets of experiments to explore the capabilities of our MSAD dataset. For generalizability and adaptability evaluation, we use few-shot methods to conduct

Table 2: Experimental results on single-scenario and cross-scenario evaluations. On ShanghaiTech (ShT), only 7 views are used during training and the rest views are individually used for testing. The notation ShT-$v*$ denotes the use of different camera views.

Test view	Train	FSAD [36]		Train	FSAD [36]		**SA^{2}D (ours)**	
		Micro	Macro		Micro	Macro	Micro	Macro
ShT-$v1$	ShT (7 views)	61.36	55.34	**MSAD**	63.74	62.92	**68.96**	**77.89**
ShT-$v3$		26.51	26.58		64.39	62.56	**67.59**	**73.43**
ShT-$v5$		53.40	53.32		55.04	54.63	**55.74**	**54.02**
ShT-$v6$		78.36	78.27		70.26	71.02	**75.47**	**72.35**
ShT-$v8$		50.02	52.54		59.97	57.45	**60.85**	**61.52**

Table 3: Evaluations on cross-scenario setups. We use FSAD [36] and SA^{2}D (ours) for training on ShanghaiTech (ShT) and MSAD, respectively.

Train	Test	AUC	
		Micro	Macro
ShT	UCSD Ped2	57.38	58.36
	CUHK Avenue	69.98	78.32
	MSAD	**63.92**	**64.92**
MSAD	UCSD Ped2	**70.35**	**65.74**
	CUHK Avenue	**79.57**	**84.49**
	MSAD	**69.96**	**69.60**

both cross-view (single-scenario) and cross-scenario evaluations under Protocol i, which helps assess the model's ability to adapt to new viewpoints and scenarios with limited training data. For practical considerations and popularity, we select some representative weakly supervised methods for evaluation using Protocol ii. This protocol evaluates the practical applicability and effectiveness of these methods using our dataset, reflecting common practices in the field. All experiments are conducted using the National Computational Infrastructure (NCI) Gadi, with one V100 GPU allocated for each experiment. Additional evaluations can be found in Appendix F.

4.1 Generalizability and adaptability

Setup. We first show the superior performance of our scenario-adaptive model by comparing SA^{2}D with the FSAD model [36] on CUHK Avenue. To demonstrate the enhanced anomaly detection performance of our MSAD dataset, we train two adaptive models: one using ShanghaiTech, and the other leveraging our MSAD. For all experiments, we maintain $N = 7$ and $K = 10$ to ensure a fair comparison. Our training process spans over 1500 epochs. Throughout all evaluations, we consider both Micro and Macro AUC scores. We conduct two sets of experiments: (1) In the *single-scenario / cross-view evaluation*, the model is trained and tested on the same scenario. To facilitate a more comprehensive comparison with the view-adaptive anomaly detection model [36], we partition the ShanghaiTech dataset into 7 scenes (views), specifically, scene 2, 4, 7, 9, 10, 11, and 12 for training as in [36]. Subsequently, the remaining 5 camera views are individually used for testing purposes.[3] (2) In the *cross-scenario evaluation*, the model is trained on one scenario and tested on a completely different one. For model training, we use either ShanghaiTech or our MSAD.

Single-scenario evaluation. As illustrated in Table 2, our model is trained on seven camera views of the ShanghaiTech dataset and tested on the remaining views, such as $v1$, $v3$, and so forth. The performance in cross-view evaluation falls short compared to our SA^{2}D. Notably, our SA^{2}D, trained on our MSAD and tested on ShanghaiTech, exhibits significantly superior performance compared to the view-adaptive model. This disparity in performance may stem from: (i) the constraint of limited training views, preventing the model from effectively adapting to novel viewpoints, (ii) the camera view being tested on at ShanghaiTech is significantly different, almost equivalent to a novel scenario concept. Therefore, the view-adaptive model is unable to adapt to such novel scenarios.

Cross-scenario evaluation. Our model, trained on MSAD, demonstrates superior performance compared to the view-adaptive model (see Table 3). For instance, on CUHK Avenue, our model outperforms the view-adaptive model by 9.6% and 6.2% for Micro and Macro evaluation metrics, respectively. Furthermore, our SA^{2}D, trained and tested on MSAD, consistently achieves excellent performance. It's important to note that our MSAD test set is distinct from the MSAD training set, covering a diverse range of novel scenarios. These results underscore the robustness of our model in cross-scenario evaluations. However, we also observe a performance drop on ShanghaiTech ($v6$). The decline in performance is attributed to the dataset deviating from real-life scenarios, as it categorizes biking and driving as anomalies, a classification that does not align with reality.

[3]There are no test videos for scene 13, so we only consider 5 views during testing.

Table 4: Comparison of six methods with varying backbones on UCF-Crime, ShanghaiTech, and our MSAD dataset using three popular backbones: C3D, I3D, and SwinTransformer (SwinT).

	Venue	UCF-Crime			ShanghaiTech			**MSAD**	
		C3D	I3D	SwinT	C3D	I3D	SwinT	I3D	SwinT
MIST [18]	CVPR 2021	81.40	82.30	-	93.13	94.83	-	-	-
RTFM [76]	ICCV 2021	83.28	83.14	83.31	91.51	97.94	96.76	86.65	85.67
MSL [31]	AAAI 2022	82.85	85.30	85.62	94.23	95.45	97.32	-	-
UR-DMU [102]	AAAI 2023	82.65	86.19	83.74	94.67	96.15	95.71	85.02	72.36
MGFN [13]	AAAI 2023	82.37	83.44	84.30	90.82	93.97	93.58	84.96	78.94
TEVAD [12]	CVPRW 2023	83.39	84.54	84.65	92.05	98.10	97.63	86.82	83.60

Insights on single- and cross-scenario evaluations. Based on the aforementioned experiments, we can deduce that a model trained on intricate real-world scenarios exhibits superior generalization. This stems from the fact that real-world models are frequently influenced by the surrounding environment, encompassing elements like fluctuating traffic patterns, dynamic electronic displays, and the movement of trees in the wind. The model must discern the nuances of anomaly detection within a dynamic environment and comprehend the dynamics of objects and/or performing subjects within it. Our MSAD dataset provides a comprehensive representation of real-world scenarios.

4.2 Practical applicability and effectiveness

Setup. We select six recent representative weakly-supervised methods for evaluation: MIST [18], RTFM [76], MSL [31], UR-DMU [102], MGFN [13], and TEVAD [12]. Using Protocol ii, we choose two typical datasets, UCF-Crime and ShanghaiTech, in addition to our MSAD dataset. To maintain consistency and fairness across experiments, we use three popular backbones as video encoders for each method: C3D (2048-dim), I3D (2048-dim, 10-crop), and SwinTransformer (1024-dim). We rigorously maintain original parameters to reproduce reported results from previous studies, and conduct our own evaluations. We use the Micro AUC metric for evaluation.

Evaluation on methods. Table 4 shows the results. As shown in the table, there is no single best performer across all three datasets. However, on ShanghaiTech with I3D as the backbone, TEVAD emerges as the top performer. This demonstrates the efficacy of using text descriptions in video anomaly detection. The second-best performer is RTFM, followed by UR-DMU.

Evaluation on backbones. For MSL, we observe that SwinTransformer outperforms I3D, and I3D outperforms C3D. Specifically, on UCF-Crime, SwinTransformer as a backbone performs better than using I3D. However, on our MSAD dataset, I3D performs better than SwinTransformer. The potential reason for this behavior is that UCF-Crime generally contains longer videos, and SwinTransformer is good at capturing long-term motions.

The role of MSAD. We noticed that our dataset is sensitive to the choice of backbones, particularly for methods like UR-DMU and MGFN. Overall, using I3D as the backbone yields better results compared to using SwinTransformer. RTFM is robust to the choice of backbones, with the performance gap between I3D and SwinTransformer being within 1%. Our MSAD dataset can be used for model backbone selection or for exploring more powerful and practical methods for real-world applications.

5 Conclusion

In this paper, we provide a concise review and identify several practical issues in video anomaly detection, particularly the lack of comprehensive datasets with diverse scenarios. To address these challenges, we introduce the Multi-Scenario Anomaly Detection (MSAD) dataset. This dataset includes diverse motion patterns and challenging variations, providing a robust foundation for developing superior models. Alongside the dataset, we present our SA^{2}D model, which uses a few-shot learning framework to efficiently adapt to new concepts and scenarios. Experimental results demonstrate the model's robustness, excelling not only in new camera views of the same scenario but

also in novel scenarios. Our contributions offer valuable resources and insights to advance the field of video anomaly detection, addressing current challenges and setting the stage for future research directions.

Acknowledgments and Disclosure of Funding

Liyun Zhu conducted this research under the supervision of Lei Wang for master's final year research project at ANU. Liyun Zhu and Arjun Raj are recipients of research sponsorship from Active Intelligence Australia Pty Ltd in Perth, Western Australia, which includes The Active Intelligence Research Challenge Award. This work was also supported by the NCI Adapter Scheme Q4 2023 and the NCI National AI Flagship Merit Allocation Scheme, with computational resources provided by NCI Australia, an NCRIS-enabled capability supported by the Australian Government.

References

[1] Z. K. Abbas and A. A. Al-Ani. A comprehensive review for video anomaly detection on videos. In *2022 International Conference on Computer Science and Software Engineering (CSASE)*, pages 1–1, 2022.

[2] A. Acsintoae, A. Florescu, M.-I. Georgescu, T. Mare, P. Sumedrea, R. T. Ionescu, F. S. Khan, and M. Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20143–20153, 2022.

[3] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, 30(3):555–560, 2008.

[4] M. Astrid, M. Z. Zaheer, and S.-I. Lee. Synthetic temporal anomaly guided end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 207–214, 2021.

[5] W. Bao, Q. Yu, and Y. Kong. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2682–2690, 2020.

[6] C. Cao, Y. Lu, P. Wang, and Y. Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20392–20401, 2023.

[7] C. Cao, Y. Lu, and Y. Zhang. Context recovery and knowledge retrieval: A novel two-stream framework for video anomaly detection. *IEEE Transactions on Image Processing*, 2024.

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[9] J. Carreira and A. 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.

[10] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun. Anticipating accidents in dashcam videos. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part IV 13*, pages 136–153. Springer, 2017.

[11] M.-N. Chapel and T. Bouwmans. Moving objects detection with a moving camera: A comprehensive review. *Computer science review*, 38:100310, 2020.

[12] W. Chen, K. T. Ma, Z. J. Yew, M. Hur, and D. A.-A. Khoo. Tevad: Improved video anomaly detection with captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5548–5558, 2023.

[13] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, and Y.-C. Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37(1), pages 387–395, 2023.

[14] H. Du, S. Zhang, B. Xie, G. Nan, J. Zhang, J. Xu, H. Liu, S. Leng, J. Liu, H. Fan, et al. Uncovering what, why and how: A comprehensive benchmark for causation understanding of video anomaly. *arXiv preprint arXiv:2405.00181*, 2024.

[15] G. Erez, R. S. Weber, and O. Freifeld. A deep moving-camera background model. In *European Conference on Computer Vision*, pages 177–194. Springer, 2022.

[16] Y. Fan, G. Wen, D. Li, S. Qiu, M. D. Levine, and F. Xiao. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. *Computer Vision and Image Understanding*, 195:102920, 2020.

[17] J. Fang, J. Qiao, J. Xue, and Z. Li. Vision-based traffic accident detection and anticipation: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[18] J.-C. Feng, F.-T. Hong, and W.-S. Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14009–14018, 2021.

[19] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[20] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12742–12752, 2021.

[21] M. I. Georgescu, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4505–4523, 2021.

[22] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[24] O. Hirschorn and S. Avidan. Normalizing flows for human pose anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13545–13554, 2023.

[25] O. Hirschorn and S. Avidan. Normalizing flows for human pose anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13545–13554, 2023.

[26] X. Hu, S. Hu, Y. Huang, H. Zhang, and H. Wu. Video anomaly detection using deep incremental slow feature analysis network. *IET Computer Vision*, 10(4):258–267, 2016.

[27] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2019.

[28] R. Jiao, Y. Wan, F. Poiesi, and Y. Wang. Survey on video anomaly detection in dynamic scenes with moving cameras. *Artificial Intelligence Review*, 56(Suppl 3):3515–3570, 2023.

[29] P. Kaur, M. Gangadharappa, and S. Gautam. An overview of anomaly detection in video surveillance. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 607–614. IEEE, 2018.

[30] F. Landi, C. G. Snoek, and R. Cucchiara. Anomaly locality in video surveillance. *arXiv preprint arXiv:1901.10364*, 2019.

[31] S. Li, F. Liu, and L. Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1395–1403, 2022.

[32] X. Li and Z.-m. Cai. Anomaly detection techniques in surveillance videos. In *2016 9th International congress on image and signal processing, BioMedical engineering and informatics (CISP-BMEI)*, pages 54–59. IEEE, 2016.

[33] W. Liu, W. Luo, D. Lian, and S. Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018.

[34] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[35] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.

[36] Y. Lu, F. Yu, M. K. K. Reddy, and Y. Wang. Few-shot scene-adaptive anomaly detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 125–141. Springer, 2020.

[37] W. Luo, W. Liu, and S. Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, pages 341–349, 2017.

[38] W. Luo, W. Liu, D. Lian, J. Tang, L. Duan, X. Peng, and S. Gao. Video anomaly detection with sparse coding inspired deep neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1070–1084, 2019.

[39] H. Lv, C. Chen, Z. Cui, C. Xu, Y. Li, and J. Yang. Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15425–15434, June 2021.

[40] H. Lv and Q. Sun. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*, 2024.

[41] A. V. Malawade, S.-Y. Yu, B. Hsu, D. Muthirayan, P. P. Khargonekar, and M. A. Al Faruque. Spatiotemporal scene-graph embedding for autonomous vehicle collision prediction. *IEEE Internet of Things Journal*, 9(12):9379–9388, 2022.

[42] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[43] J. R. Medel and A. Savakis. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv:1612.00390*, 2016.

[44] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. *IEEE Transactions on pattern analysis and machine intelligence*, 23(8):873–889, 2001.

[45] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942, 2009.

[46] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11988–11996, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society.

[47] M. G. Narasimhan. Dynamic video anomaly detection and localization using sparse denoising autoencoders. *Multimedia Tools and Applications*, 77:13173–13195, 2018.

[48] R. Nayak, U. C. Pati, and S. K. Das. A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, 106:104078, 2021.

[49] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

[50] H. Park, J. Noh, and B. Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14372–14381, 2020.

[51] D. Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991.

[52] O. P. Popoola and K. Wang. Video-based abnormal human behavior recognition—a review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):865–878, 2012.

[53] Y. Pu, X. Wu, and S. Wang. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *arXiv preprint arXiv:2306.14451*, 2023.

[54] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[55] B. Ramachandra and M. Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2569–2578, 2020.

[56] B. Ramachandra, M. Jones, and R. Vatsavai. Learning a distance function with a siamese network to localize anomalies in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2598–2607, 2020.

[57] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6262–6271, 2019.

[58] T. Reiss and Y. Hoshen. Attribute-based representations for accurate and interpretable video anomaly detection. *arXiv preprint arXiv:2212.00789*, 2022.

[59] F. Rezazadegan, S. Shirazi, B. Upcrofit, and M. Milford. Action recognition: From static datasets to moving robots. In *2017 IEEE International conference on robotics and automation (ICRA)*, pages 3185–3191. IEEE, 2017.

[60] N.-C. Ristea, F.-A. Croitoru, R. T. Ionescu, M. Popescu, F. S. Khan, and M. Shah. Self-distilled masked auto-encoders are efficient video anomaly detectors. *arXiv preprint arXiv:2306.12041*, 2023.

[61] N.-C. Ristea, N. Madan, R. T. Ionescu, K. Nasrollahi, F. S. Khan, T. B. Moeslund, and M. Shah. Self-supervised predictive convolutional attentive block for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13576–13586, 2022.

[62] L. Riz, A. Caraffa, M. Bortolon, M. L. Mekhalfi, D. Boscaini, A. Moura, J. Antunes, A. Dias, H. Silva, A. Leonidou, et al. The monet dataset: Multimodal drone thermal dataset recorded in rural scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2545–2553, 2023.

[63] R. Rodrigues, N. Bhargava, R. Velmurugan, and S. Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[64] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[65] P. Rota, N. Conci, N. Sebe, and J. M. Rehg. Real-life violent social interaction detection. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 3456–3460, 2015.

[66] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing*, 26(4):1992–2004, 2017.

[67] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Computer Vision and Image Understanding*, 172:88–97, 2018.

[68] B. Sabzalian, H. Marvi, and A. Ahmadyfard. Deep and sparse features for anomaly detection and localization in video. In *2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, pages 173–178. IEEE, 2019.

[69] Y. A. Samaila, P. Sebastian, N. S. S. Singh, A. N. Shuaibu, S. S. A. Ali, T. I. Amosa, G. E. Mustafa Abro, and I. Shuaibu. Video anomaly detection: A systematic review of issues and prospects. *Neurocomputing*, 591:127726, 2024.

[70] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[71] J. J. P. Suarez and P. C. N. J. au2. A survey on deep learning techniques for video anomaly detection, 2020.

[72] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.

[73] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.

[74] L. A. Thomaz, E. Jardim, A. F. da Silva, E. A. da Silva, S. L. Netto, and H. Krim. Anomaly detection in moving-camera video sequences using principal subspace analysis. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 65(3):1003–1015, 2017.

[75] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang. Eliminating background-bias for robust person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[76] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4975–4986, October 2021.

[77] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[78] J. A. Uribe, L. Fonseca, and J. Vargas. Video based system for railroad collision warning. In *2012 IEEE International Carnahan Conference on Security Technology (ICCST)*, pages 280–285. IEEE, 2012.

[79] M. Vijay, W.-X. LI, B. Viral, and V. Nuno. Anomaly detection in crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1975–1981, 2010.

[80] H. Vu, D. Phung, T. D. Nguyen, A. Trevors, and S. Venkatesh. Energy-based models for video anomaly detection. *arXiv preprint arXiv:1708.05211*, 2017.

[81] G. Wang, Y. Wang, J. Qin, D. Zhang, X. Bao, and D. Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *European Conference on Computer Vision*, pages 494–511. Springer, 2022.

[82] L. Wang, D. Q. Huynh, and M. R. Mansour. Loss switching fusion with similarity search for video classification. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 974–978. IEEE, 2019.

[83] L. Wang and P. Koniusz. Flow dynamics correction for action recognition. *ICASSP*, 2024.

[84] L. Wang, J. Liu, and P. Koniusz. 3D skeleton-based few-shot action recognition with JEANIE is not so naïve. *arXiv preprint arXiv: 2112.12668*, 2021.

[85] L. Wang, J. Liu, L. Zheng, T. Gedeon, and P. Koniusz. Meet jeanie: a similarity measure for 3d skeleton sequences via temporal-viewpoint alignment. *International Journal of Computer Vision*, pages 1–32, 2024.

[86] L. Wang, K. Sun, and P. Koniusz. High-order tensor pooling with attention for action recognition. *ICASSP*, 2024.

[87] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[88] L. Wang, X. Yuan, T. Gedeon, and L. Zheng. Taylor videos for action recognition. In *Forty-first International Conference on Machine Learning*, 2024.

[89] S. Wang, E. Zhu, J. Yin, and F. Porikli. Video anomaly detection and localization by local motion based joint video representation and ocelm. *Neurocomputing*, 277:161–175, 2018.

[90] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[91] Z. Wang, Y. Zhou, R. Wang, T.-Y. Lin, A. Shah, and S. N. Lim. Few-shot fast-adaptive anomaly detection. *Advances in Neural Information Processing Systems*, 35:4957–4970, 2022.

[92] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer, 2020.

[93] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, and Y. Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. *arXiv preprint arXiv:2308.11681*, 2023.

[94] K. Xu, T. Sun, and X. Jiang. Video anomaly detection and localization based on an adaptive intra-frame classification network. *IEEE Transactions on Multimedia*, 22(2):394–406, 2019.

[95] Y. Yao, X. Wang, M. Xu, Z. Pu, Y. Wang, E. Atkins, and D. J. Crandall. Dota: Unsupervised detection of traffic anomaly in driving videos. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):444–459, 2022.

[96] Y. Yao, M. Xu, C. Choi, D. J. Crandall, E. M. Atkins, and B. Dariush. Egocentric vision-based future vehicle localization for intelligent driving assistance systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9711–9717. IEEE, 2019.

[97] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins. Unsupervised traffic accident detection in first-person videos. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 273–280. IEEE, 2019.

[98] J. Yu, Y. Lee, K. C. Yow, M. Jeon, and W. Pedrycz. Abnormal event detection and localization via adversarial event prediction. *IEEE transactions on neural networks and learning systems*, 33(8):3572–3586, 2021.

[99] T. Yuan, X. Zhang, K. Liu, B. Liu, C. Chen, J. Jin, and Z. Jiao. Towards surveillance video-and-language understanding: New dataset, baselines, and challenges, 2023.

[100] L. Zanella, W. Menapace, M. Mancini, Y. Wang, and E. Ricci. Harnessing large language models for training-free video anomaly detection. *arXiv preprint arXiv:2404.01014*, 2024.

[101] Y. Zhang, H. Lu, L. Zhang, and X. Ruan. Combining motion and appearance cues for anomaly detection. *Pattern Recognition*, 51:443–452, 2016.

[102] H. Zhou, J. Yu, and W. Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37(3), pages 3769–3777, 2023.

[103] J. T. Zhou, L. Zhang, Z. Fang, J. Du, X. Peng, and Y. Xiao. Attention-driven loss for anomaly detection in video surveillance. *IEEE transactions on circuits and systems for video technology*, 30(12):4639–4647, 2019.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes]
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

i. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

ii. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

iii. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

iv. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]

v. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Dataset collection process

We have obtained permission from YouTube, Itemfix, and Bilibili to use their video data for non-commercial purposes. To facilitate the data collection, we use a script to automatically download relevant videos from these platforms and trim them into short clips. We also release these scripts along with our dataset for interested researchers.

Real-world scenarios present greater challenges due to diverse weather and lighting conditions, as well as unpredictable objects that can impact detection, such as dynamic electronic advertising screens or moving car headlights during nighttime in video surveillance. Our dataset considers these variations. Our test set covers a broad spectrum of abnormal events observed in diverse scenarios, encompassing both human-related and non-human-related anomalies, such as fire incidents, fighting, shooting, traffic accidents, falls, and other anomalies. Our test set exclusively contains video frame-level annotations. We apply a rigorous filtering process to ensure the quality and appropriateness of the collected videos. The following criteria are used to remove unsuitable videos:

i. Quality considerations:
   (a) Low resolution: videos that do not meet our resolution standards are excluded.
   (b) Grayscale videos: only color videos are retained to ensure richness in visual information.
   (c) Moving camera views: videos with unstable or moving camera perspectives are excluded to maintain consistency in viewpoint, as we focus on static camera video anomaly detection (VAD).

ii. Content considerations:
   (a) Excessive text overlays: videos with many text overlays that obscure the visual content are removed.
   (b) Potentially invasive content: videos that potentially invade privacy, such as those showing identifiable faces, are excluded.
   (c) Violent content: videos that are overly violent, depict violence against children, or contain graphic content are removed to ensure ethical standards.
   (d) Political content: videos containing political content are excluded to avoid any potential biases or controversies.

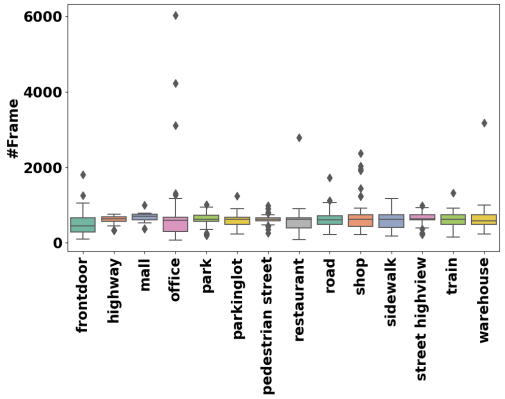
This thorough process ensures that our dataset is of high quality, ethically sound, and suitable for research purposes in VAD.

B Further dataset details

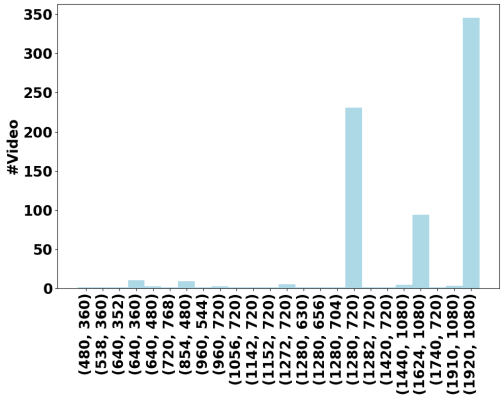
Anomaly types. Table 5 shows the details of the main anomalies, detailed anomalies per main anomaly type, and their relevant domains. As shown in the table, our dataset encompasses a broad range of both human-related and non-human-related anomalies, spanning multiple domains. A boxplot showing frame number variations across scenarios in the whole MSAD dataset is provided in Fig. 4a.

Resolution. All the videos we've collected are high resolution, such as 1920×1080 (see Fig. 4b). Each video is captured from a fixed camera view with a standard frame rate of 30. Our dataset offers several advantages: (i) Most existing cameras in industry surveillance feature high resolution, thanks to advanced camera technologies. (ii) High-resolution videos provide more detailed information for anomaly detection tasks, particularly when the performing subject is distant from the camera, or the object being interacted with is relatively small. (iii) Existing anomaly detection models trained on low-resolution videos often suffer from poor detection performance, especially in the presence of tiny or fine-grained motion patterns. Therefore, our dataset is better suited for training a more effective anomaly detection model.

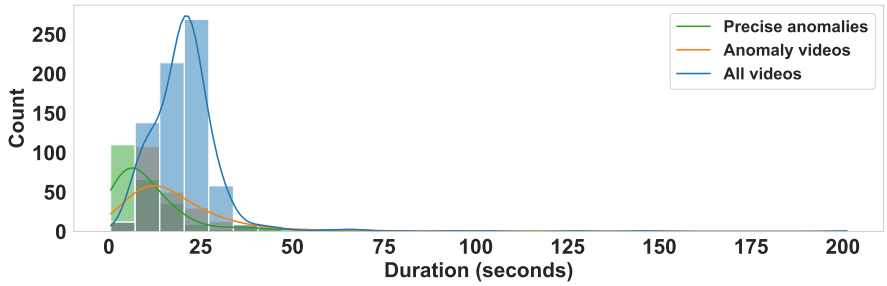
Video/anomaly duration. As shown in figure 4c, in our dataset, the majority of all videos (blue) have durations between 10 and 30 seconds, with a peak around 20 seconds. There is a sharp decline in the



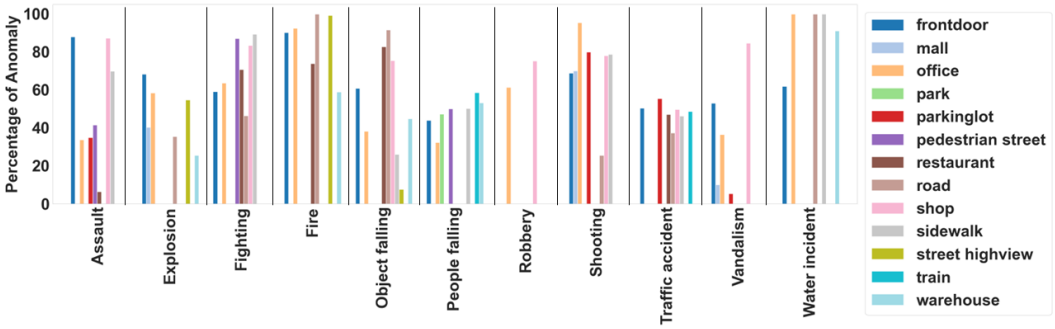
(a) Frame number variations.



(b) Distribution of resolutions.



(c) Distribution of video durations. The histogram displays video durations for actual anomaly video length (precise anomalies in green), anomaly videos (orange), and all videos (blue).



(d) The bar chart illustrates the percentage of anomalies in the anomaly videos grouped by anomaly types and scenarios. The black lines separate different types of anomalies.

Figure 4: Distributions of (a) frame number variations, (b) video resolutions, (c) video durations, and (d) detailed distributions per anomaly type per scenario in the entire MSAD dataset are presented. Our main paper presents the distribution of frame number variations in the training set (see Fig. 3b).

Table 5: Details of our MSAD dataset. Our dataset has a broad range of both human-related and non-human-related anomalies, spanning multiple domains.

	Main Anomaly Types	Detailed Anomaly Types	Domain
Human-related	Assault	Assault on street	Crime
		Assault in office	Crime
		Other assault	Crime
	Fighting	Fighting on street	Violence
		Fighting in a restaurant	Violence
		Fighting in a shop	Violence
		Fighting in front of a door	Violence
		Fighting indoors	Violence
		Other fighting	Violence
	People Falling	People falling to ground	Pedestrian
		People falling into pool	Pedestrian
		People falling from high places	Pedestrian
		People falling into subway	Pedestrian
		Other people falling	Pedestrian
	Robbery	Shop robbery	Crime
		Office robbery	Crime
		Theft	Crime
		Car theft	Crime
		Other robbery	Crime
	Shooting	Shooting on the road	Crime
		Shooting indoors	Crime
		Holding a gun	Crime
		Other shooting	Crime
	Traffic Accident	Car falling	Traffic
		Car crash	Traffic
		Speeding	Traffic
		Car rushing into building	Traffic
		Car crash with people	Traffic
		Car crash with object	Traffic
		Car crash with train	Traffic
		Motorcycle crash	Traffic
		Other traffic accident	Traffic
	Vandalism	Vandalizing glass	Violence
		Vandalizing door	Violence
		Other vandalism	Violence
Non-human-related	Explosion	Street explosion	Emergency
		Firework explosion	Emergency
		Factory explosion	Emergency
		Indoor explosion	Emergency
		Other explosion	Emergency
	Fire	Smoke	Emergency
		Factory fire	Emergency
		Building on fire	Emergency
		Bush fire	Emergency
		Other fire	Emergency
	Object Falling	Strong wind	Natural hazard
		Object falling in home	Emergency
		Tree falling	Emergency
		Large objects falling	Emergency
		Glass falling	Emergency
		Other objects falling	Emergency
	Water Incident	Flood	Natural hazard
		Water leakage	Emergency
		Heavy rain	Natural hazard
		Other water incidents	Emergency

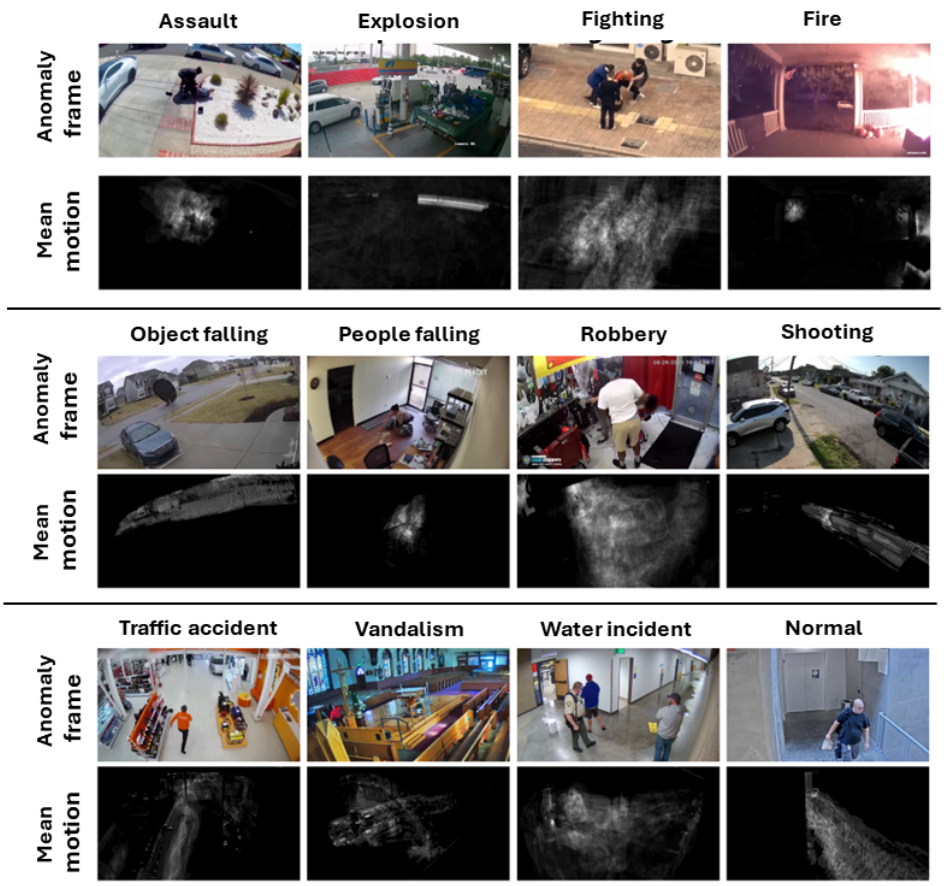


Figure 5: Visualizations of anomaly frames and their corresponding mean motion maps. The mean motion map is calculated by averaging the frame difference maps obtained from pairs of consecutive frames.

count of videos longer than 30 seconds. While the overall distribution (blue) and anomaly videos (orange) follow a similar trend, precise anomalies (green) show distinctive peaks at shorter durations, suggesting that anomalies are very rare and brief. This indicates that our dataset effectively captures essential and engaging moments within concise time frames, making it highly valuable for quick insights and analyses. The distribution highlights the efficiency and focus of the captured events.

Scenario and anomaly details. Fig. 4d shows more detailed distributions per anomaly type per scenario. It should be noted that not all anomaly types have bar lines because there are no scenarios associated with them. The plot shows a varying set of anomaly video percentages. In our dataset, we have endeavored to give a comprehensive collection of videos with varying anomaly lengths in them.

Motion visualisation. In figure 5, we present visualizations of some anomaly frames from our MSAD dataset, along with their corresponding mean motion maps. These maps are generated by averaging the frame difference maps from consecutive frames. The intensity and distribution of motion vary across different types of anomalies. For instance, continuous anomalies like fighting and robbery exhibit a widespread motion pattern, while abrupt incidents such as robbery and object falling display more concentrated motion. The maps also reveal specific directional movements, such as the downward motion evident in the object falling and people falling anomalies. Our dataset is designed to include a diverse array of videos, capturing a wide range of these attributes.

Evaluation metrics. Existing anomaly detection datasets are annotated in different ways, including pixel-level, frame-level and video-level annotations. Different annotation methods correspond to different evaluation metrics. For example, frame-level annotation often uses area under the curve

(AUC) as the evaluation metric. As discussed in [21], there are two kinds of AUC score, namely Micro-AUC and Macro-AUC. The former concatenates the frames from all the videos and computes the overall AUC value, and the latter computes the AUC value for each video and averages it.

For pixel-level annotations, Ramachandra et al. [55] introduced two evaluation metrics: the Region-Based Detection Criterion (RBDC) and the Track-Based Detection Criterion (TBDC). These metrics are designed to prioritize the false positive rate on both temporal and spatial dimensions. This consideration stems from the fact that anomalies in videos often extend across multiple frames. Hence, detecting anomalies in any segment and reducing the false detection rate holds significance for VAD systems.

To ensure a fair comparison between different algorithms, we use frame-level annotations for test. To be more precise, when the frame contains abnormal regions, we consider the frame as anomaly to obtain frame-level label. It is worth noting that many methods only use frame-level AUC as the evaluation metric. However, besides detecting anomalous events, it is crucial to avoid misclassifying normal events as anomalies as much as possible in practical applications. Therefore, in order to accurately evaluate the model's performance, we consider both frame-level AUC and false positive rate.

C Dataset usage and maintenance

VAD datasets serve multiple purposes within the research community. Primarily, they are used for training and evaluating machine learning models. By providing labeled examples of normal and anomalous events, these datasets enable the development of effective anomaly detection algorithms. Researchers also use these datasets to benchmark the performance of different algorithms, facilitating consistent and fair comparisons across various methods. Additionally, the inclusion of diverse scenarios allows for cross-scenario testing, assessing the generalizability of anomaly detection models across different environmental conditions. Furthermore, these datasets are valuable for feature extraction and analysis, providing insights into the nature of anomalies and enhancing the interpretability of detection models.

The use of this dataset is governed by the following terms and conditions:

i. Without the expressed permission from our research group members, any of the following will be considered illegal: redistribution, derivation or generation of a new dataset from this dataset, and commercial usage of any of these videos in any way or form, either partially or in its entirety.

ii. For the sake of privacy, images of all subjects in any of these videos are only allowed for demonstration in academic publications and presentations.

This dataset is released for academic research only and is free to researchers from educational or research institutes for non-commercial purposes.

Ensuring the reliability and relevance of VAD datasets requires ongoing maintenance efforts. We will expand our dataset by adding more scenarios and anomaly types to cover more domains. We will continue collecting high-quality videos to advance static camera VAD. The limitation of our work is presented in Sec. H.

Our dataset is publicly available here. We will create a new website and provide long-term support for researchers interested in using our dataset.

D Scenario-adaptive anomaly detection framework

Below we present our Scenario-Adaptive Anomaly Detection (SA^{2}D) model. First, we introduce the notations.

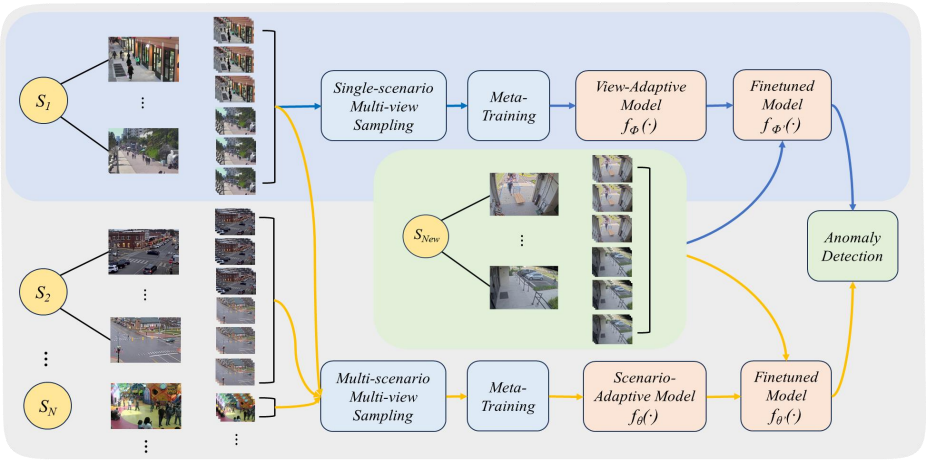


Figure 6: A comparison between the existing few-shot scene-adaptive (view-adaptive) anomaly detection model (depicted by the light blue block) and our proposed Scenario-Adaptive Anomaly Detection (SA^{2}D) model (illustrated by the light gray block). On the left-hand side of the figure, the first column, $\{S_1, S_2, \cdots, S_N\}$, represents different scenarios. The second column signifies various camera viewpoints, while the third column indicates the videos captured under each camera viewpoint. In contrast to the existing view-adaptive model, SA^{2}D extends its capabilities by incorporating a few-shot multi-scenario multi-view learning framework. The blue arrow illustrates the workflow of existing models, while the orange arrows show the workflow of our model.

Notations: $\mathcal{I}_K$ represents the index set $1, 2, \cdots, K$. Calligraphic mathcal fonts denote tensors (*e.g.*, $\mathbf{V}$), capitalized bold symbols are matrices (*e.g.*, $\boldsymbol{F}$), lowercase bold symbols denote vectors (*e.g.*, $\boldsymbol{x}$), and regular fonts indicate scalars (*e.g.*, x). Concatenation of n scalars is denoted as $[x]_{i\in\mathcal{I}_n}^{\oplus}$.

Few-shot multi-scenario multi-view learning. Using a few-shot learning framework for anomaly detection is not an entirely new concept. One of the most widely recognized methods in this domain is the Few-shot Scene-adaptive Anomaly Detection (FSAD) model [36]. This model leverages the meta-learning framework [19] to train a model using video data collected from various camera views within the same scenario[4], such as a university street. The model is then fine-tuned on a different camera viewpoint within the university site. Although the trained model can be adapted to novel viewpoints, still its adaptability is confined to the university scenario. To overcome this limitation, we introduce the Scenario-Adaptive Anomaly Detection (SA^{2}D) method. Diverging from existing few-shot anomaly detection models [36], our approach: (i) builds on and extends one-scenario multi-view to multi-scenario multi-view learning problem, and (ii) broadens the scope of test cases from multiple camera views to encompass novel scenarios, as well as multiple camera views per scenario. Our model stands out as a more advanced and versatile version in comparison to existing anomaly detection models. Fig. 6 shows a comparison between the existing few-shot anomaly detection model and our SA^{2}D model.

Our model operates at the frame level, recognizing the significance of early anomaly detection, where swift action is imperative to prevent or mitigate potential issues. We opt for a future frame prediction model, employing a T-frame video $\mathbf{V} = [\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_T] \in \mathbb{R}^{H\times W\times T}$. For the sake of simplicity, we omit the three color channels. A temporal sliding window of size T' with a step size of 1 is used to select video sub-sequences (termed video temporal blocks) $\mathbf{V}_i = [\boldsymbol{F}_i, \boldsymbol{F}_{i+1}, \cdots, \boldsymbol{F}_{i+T'-1}] \in \mathbb{R}^{H\times W\times T'}$, where i represents the ith temporal block. For simplicity, we omit i in the subsequent discussion, *e.g.*, $\mathbf{V} = [\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_{T'}] \in \mathbb{R}^{H\times W\times T'}$ denotes a temporal block. We consider the first $T'-1$ frames as the input $\mathbf{X} = [\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_{T'-1}] \in \mathbb{R}^{H\times W\times (T'-1)}$ to future frame prediction model, and the last

[4]In this paper, the term 'scenario' is employed to signify different contexts such as retail, manufacturing, the education sector, smart cities, and others. In the context of different camera views, existing literature sometimes uses the term 'scene' to refer to the camera scene from a specific camera viewpoint.

frame as the output $\boldsymbol{Y} = \boldsymbol{F}_{T'} \in \mathbb{R}^{H \times W}$, forming an input/output pair $(\mathbf{X}, \boldsymbol{Y})$. Our future frame prediction model is represented as $f_\theta : \mathbf{X} \rightarrow \boldsymbol{Y}$.

Sampling strategy. In the context of meta-learning, we begin by sampling a set of N scenarios $\{S_1, S_2, \cdots, S_N\}$. Under each scenario, we randomly select one camera view from a set of M camera viewpoints $\{V_1, V_2, \cdots, V_M\}$. Under the chosen camera view, we sample K video temporal blocks to formulate an N-way K-shot learning problem. This sampling strategy enables the creation of a corresponding task $\mathcal{T}_{n,m} = \{\mathcal{D}^{\text{tr}}_{n,m}, \mathcal{D}^{\text{val}}_{n,m}\}$ per training episode, where $n \in \mathcal{I}_N$, $m \in \mathcal{I}_M$, and $\mathcal{D}^{\text{tr}}_{n,m} = \{(\mathbf{X}_1, \boldsymbol{Y}_1), (\mathbf{X}_2, \boldsymbol{Y}_2), \cdots, (\mathbf{X}_K, \boldsymbol{Y}_K)\}$. These K pairs of $\mathcal{D}^{\text{tr}}_{n,m}$ are utilized to train the future frame prediction model f_θ. Additionally, we sample a subset of input/output pairs to form the validation set $\mathcal{D}^{\text{val}}_{n,m}$, excluding those pairs in $\mathcal{D}^{\text{tr}}_{n,m}$. For the backbone selection, we adopt U-Net [64] for future frame prediction, incorporating a ConvLSTM block [70] for sequential modeling to retain long-term temporal information. Following [33] and [36, 91], we incorporate GAN [23] architecture for video frame reconstruction. It is important to note that the backbone architectures are not the primary focus of our work, and in theory, any anomaly detection network can serve as the backbone architecture. Given that each training episode randomly selects the camera viewpoint per scenario, and the scenarios are also randomly selected, our training scheme can be viewed as a few-shot multi-scenario multi-view learning.

Training. Given a pretrained anomaly detection model f_θ, following [19, 36], we define a task $\mathcal{T}_{n,m}$ by establishing a loss function on the training set $\mathcal{D}^{\text{tr}}_{n,m}$ of this task:

$$\mathcal{L}_{\mathcal{T}_{n,m}}(f_\theta; \mathcal{D}^{\text{tr}}_{n,m}) = \sum_{(\mathbf{X}_{n,m}, \boldsymbol{Y}_{n,m}) \in \mathcal{D}^{\text{tr}}_{n,m}} L(f_\theta(\mathbf{X}_{n,m}), \boldsymbol{Y}_{n,m}), \tag{1}$$

Here, $L(f_\theta(\mathbf{X}_{n,m}), \boldsymbol{Y}_{n,m})$ computes the difference between the predicted frame $f_\theta(\mathbf{X}_{n,m})$ and the ground truth frame $\boldsymbol{Y}_{n,m}$. Following [36], we define $L(\cdot)$ as the summation of the least absolute deviation (L_1 loss) [51], multi-scale structural similarity measurement [90], and the gradient difference loss [42]. The updated model parameters θ' are adapted to the task $\mathcal{T}_{n,m}$. On the validation set $\mathcal{D}^{\text{val}}_{n,m}$, we measure the performance of $f_{\theta'}$ as:

$$\mathcal{L}_{\mathcal{T}_{n,m}}(f_{\theta'}; \mathcal{D}^{\text{val}}_{n,m}) = \sum_{(\mathbf{X}_{n,m}, \boldsymbol{Y}_{n,m}) \in \mathcal{D}^{\text{val}}_{n,m}} L(f_{\theta'}(\mathbf{X}_{n,m}), \boldsymbol{Y}_{n,m}), \tag{2}$$

We formally define the objective of meta-learning as:

$$\min_\theta \sum_{\substack{n \in \mathcal{I}_N, \\ m \in \mathcal{I}_M}} \mathcal{L}_{\mathcal{T}_{n,m}}(f_{\theta'}; \mathcal{D}^{\text{val}}_{n,m}). \tag{3}$$

where N and M denotes respectively the number of scenarios and camera views. Note that Eq. (3) sums over all tasks during meta-training. In practice, we sample a mini-batch of tasks per iteration.

Inference. During the meta-testing stage, when provided with a video from a specific camera view of a new scenario or a novel camera viewpoint of a known scenario S_{new}, we employ the first several frames of the video in S_{new} for adaptation (using Eq. (1)) and then utilize the rest of the frames for testing.

Scoring. For anomaly detection, we calculate the disparity between the predicted frame $\hat{\boldsymbol{F}}_t$ and the ground truth frame $\boldsymbol{F}_t$, $t \in \mathcal{I}_T$. Following [33], we use the Peak Signal-to-Noise Ratio (PSNR) as a metric to evaluate the quality of the predicted frame:

$$\text{PSNR}(\boldsymbol{F}_t, \hat{\boldsymbol{F}}_t) = 10 \log_{10} \frac{(\max \boldsymbol{F}_t)^2}{\frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} (\boldsymbol{F}_t[i,j] - \hat{\boldsymbol{F}}_t[i,j])^2}, \tag{4}$$

here, $\max \boldsymbol{F}_t$ is the maximum possible pixel value of the ground truth frame, $[i, j]$ are the spatial location of the video frame, H and W are the height and width of the frame respectively. In general, higher PSNR values indicate better generated quality. The PSNR value of a video frame can be used to assess the consistency of the frames, with higher PSNR values indicating normal events as

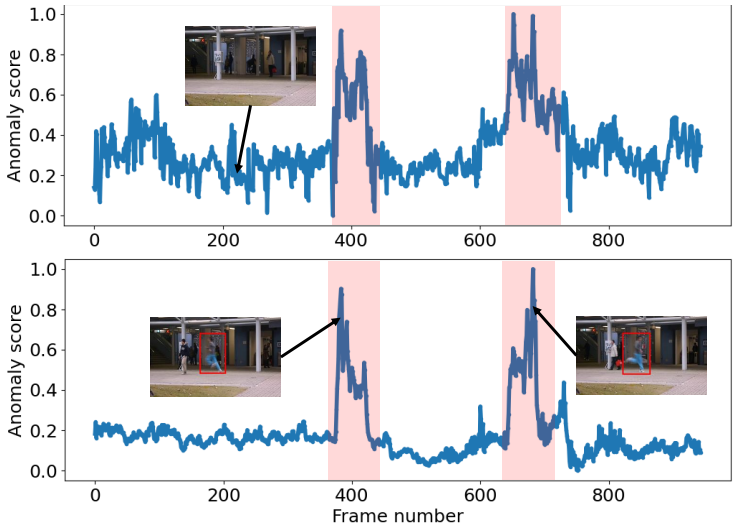


Figure 7: Frame-level anomaly scores (normalised between 0 and 1) are depicted for a test video from CUHK Avenue. The light red color block highlights the time period when the anomaly occurs. The top row illustrates the curve generated by the view-adaptive model, while the bottom row displays our scenario-adaptive model. Notably, our SA^{2}D exhibits considerably smoother curves compared to the few-shot scene-adaptive model. This observation suggests that our model, built on our MSAD, demonstrates superior anomaly detection performance.

the frame is well predicted. Following [42], we normalize the PSNR scores of a T-frame video for anomaly scoring:

$$s_t = \frac{\text{PSNR}(\boldsymbol{F}_t, \hat{\boldsymbol{F}}_t) - \min[\text{PSNR}(\boldsymbol{F}_\tau, \hat{\boldsymbol{F}}_\tau)]^{\oplus}_{\tau \in \mathcal{I}_T}}{\max[\text{PSNR}(\boldsymbol{F}_\tau, \hat{\boldsymbol{F}}_\tau)]^{\oplus}_{\tau \in \mathcal{I}_T} - \min[\text{PSNR}(\boldsymbol{F}_\tau, \hat{\boldsymbol{F}}_\tau)]^{\oplus}_{\tau \in \mathcal{I}_T}}, \tag{5}$$

where s_t varies within the range of 0 to 1. The normalized PSNR value can be used to assess the abnormality of a specific frame. A predefined threshold, *e.g.*, 0.8, can be set to determine whether a specific frame, such as $\boldsymbol{F}_t$, is considered an anomaly or not, by comparing it with s_t.

E Further discussions

Advances in feature learning. Multiple Instance Learning (MIL) framework is introduced in [72]. They treat normal and abnormal videos as bags, short clips of each video as instances, and use a ranking loss to distinguish between the highest-scoring abnormal and normal instances. Researchers [18] notice that directly using pre-trained models as feature extractors may not be suitable for surveillance as they are pre-trained on action recognition datasets. Moreover, the snippet with the highest anomaly score introduced in [72] may not come from an abnormal snippet within an abnormal video [76]. Their advanced work proposes a theoretical model to better differentiate between the top-k snippet feature magnitudes of abnormal and normal snippets. However, relying solely on feature magnitude to distinguish between normal and abnormal snippets can be ambiguous in certain situations [13]. To address this issue, a magnitude contrastive loss is introduced in [13] which enhances the similarity of feature magnitudes for videos within the same category and increases the separability between normal and abnormal videos. In comparison to earlier works [76, 13], Zhou et al. [102] emphasize the shortcomings of learning discriminative features for anomalous instances while neglecting the implication of normal data. Inspired by the methods with memory banks [4, 22, 27], they propose dual memory units for recording both normal and abnormal patterns.

Enhancing anomaly understanding. Most existing methods primarily focus on detection accuracy [49], while only a few works address interpretable anomaly detection [14, 58]. Since anomalous behaviors are scene-dependent [6], it is essential for detection methods to understand the underlying

Table 6: A comparison of recent state-of-the-art self-supervised and weakly-supervised methods on five benchmarks: UCSD Ped2 (Ped2), CUHK Avenue (CUHK), ShanghaiTech (ShT), UCF-Crime (UCF), and XD-Violence (XD). For Ped2, CUHK, ShT, and UCF, frame-level Micro-AUC (%) is used for evaluation, while for XD, average precision (%) is used.

Paradigm	Year	Method	Description	Modality	Dataset				
					Ped2	CUHK	ShT	UCF	XD
Self-sup.	2018	Liu et al. [33]	Future frame prediction	RGB+opt.	95.4	85.1	72.8	-	-
	2019	Ionescu et al. [27]	Object-centric encoder	RGB	97.8	90.4	84.9	-	-
	2019	MemAE [22]	Memory module	RGB	94.1	83.3	71.2	-	-
	2020	FSAD [36]	Few-shot scene adaptive	RGB	96.2	85.8	77.9	-	-
	2021	Georgescu et al. [21]	Background agnostic method	RGB	98.7	92.3	82.7	-	-
	2021	MPN [39]	Meta prototype unit	RGB	96.9	89.5	73.8	-	-
	2021	AEP [98]	Adversarial learning	RGB	97.3	90.2	-	-	-
	2022	Wang et al. [81]	Spatio-temporal jigsaw puzzles	RGB	99.0	92.2	84.3	-	-
	2023	STG-NF [25]	Normalizing flow	Skeleton	-	-	85.9	-	-
	2023	Ristea et al. [60]	Lightweight masked encoder	RGB	95.4	91.3	79.1	-	-
Weakly-sup.	2018	Sultani et al. [72]	Multiple instance learning	RGB	-	-	-	75.4	-
	2020	Wu et al. [92]	Multimodal fusion	RGB+audio	-	-	-	-	78.6
	2021	MIST [18]	Task-specific encoder	RGB	-	-	94.8	82.3	-
	2021	RTFM [76]	Feature magnitude learning	RGB	98.6	-	97.2	84.3	77.8
	2022	MSL [31]	Multi-sequence learning	RGB	-	-	97.3	85.6	78.6
	2022	UBnormal [2]	Synthesis dataset	RGB	-	93.0	83.7	-	-
	2023	MGFN [13]	Glance-and-focus	RGB	-	-	-	87.0	80.1
	2023	UR-DMU [102]	Dual-memory units	RGB	-	-	-	87.0	81.7
	2023	VAD-CLIP [93]	Language-visual model	RGB+text	-	-	-	88.0	84.5
	2023	TEVAD [12]	Multimodal fusion	RGB+text	98.7	-	98.1	84.9	79.8
	2024	Pu et al. [53]	Prompt-enhanced learning	RGB+text	-	-	98.1	86.8	85.6

rules and constraints of the scene in order to comprehend the causes and consequences of anomalies. Although some existing works attempt to integrate visual and semantic information to understand anomalies [12, 53, 93, 99], the guidance provided by semantic information for anomaly detection is still insufficient. Therefore, continuing to explore new methods, *e.g.*, Large Language Models (LLMs) [14, 40, 100]), that enable models to better understand the semantic information of anomalies is crucial for advancing the field.

Anomaly anticipation. Most existing methods prioritize detecting anomalies rather than predicting them. Recently, Cao et al. [6] introduce the concept of video anomaly anticipation, which aims to predict anomalies in advance to enable early warnings [17]. This proactive approach leverages trends or clues in the events to foresee potential incidents, allowing timely intervention to prevent accidents. For example, early detection of smoke before a fire ignites, masked suspects wielding knives, or individuals about to fall can mitigate risks and prevent anomalies from happening. This forward-looking strategy not only improves anomaly detection systems but also stimulates research and in predictive analytics and real-time intervention methods.

Lightweight models. Despite the advent of powerful models in recent years, most existing methods consume substantial computational resources and runtime for extracting video features [48] or using an object detector [27, 20], creating a bottleneck for deployment in real-world scenarios [80, 48]. While a few studies have focused on model light-weighting [60], the trade-off between accuracy and model efficiency remains a significant challenge. To address these issues, several approaches can be considered. On one hand, data-efficiency methods such as data distillation can be employed to reduce redundant video data, thereby decreasing the computational cost [48]. On the other hand, improving the model architecture to learn more representative features or using knowledge distillation during deployment can enhance the efficiency. Furthermore, future research should consider incorporating the running time as an evaluation metric to better access the practical utility of these models in real-world scenarios.

With vision-language models. Exploring advanced deep learning and generative models to better capture the complexity of real-world scenarios is of vital importance. In particular, using pretrained

large-scale vision-language models (VLMs) allows for a better understanding of anomalies in videos and produces corresponding descriptions. Notably, researchers have achieved state-of-the-art performance on multiple VAD tasks by using CLIP [54] to extract deep features, followed by a lightweight trainable network. Modeling both visual and language information for multimodal fusion, contextual understanding, handling complex scenarios, and providing explanatory descriptions would be an interesting future direction.

Quantitative evaluations. Table 6 presents a comparison of recent self-supervised and weakly-supervised methods on five popular video anomaly detection datasets. We observe that UCSD Ped2 (Ped2), CUHK Avenue (CUHK), and ShanghaiTech (ShT) are widely used for self-supervised learning methods, whereas ShanghaiTech (ShT), UCF-Crime (UCF), and XD-Violence (XD) are commonly employed for weakly-supervised learning approaches. Additionally, we note that weakly-supervised methods generally outperform self-supervised methods, especially on more complex datasets such as ShanghaiTech (ShT). These findings suggest that incorporating additional weak supervision can significantly enhance the detection performance of VAD models, particularly in challenging environments.

F Further evaluations

Scenario-adaptive outperforms view-adaptive. We commence by comparing our SA^{2}D with the view-adaptive model (few-shot scene-adaptive [36]), utilizing test videos from CUHK Avenue for evaluation. In Fig. 7, we present the visualization of frame-level anomaly scores for both models. The illustration indicates that our SA^{2}D (Fig. 7 *bottom*) produces significantly smoother curves compared to the view-adaptive model (Fig. 7 *top*). This observation underscores our model's enhanced adaptability to new scenarios, affirming the superior performance of our scenario-adaptive approach over the view-adaptive model.

Performance comparison by anomaly type. Table 7 presents the frame-level Micro AUC and Average Precision (AP) performance, in percentage, for three weakly-supervised methods (RTFM [76], MGFN [13], and UR-DMU [102]) pretrained on either UCF-Crime or our MSAD training set. These methods are applied to our MSAD test set across various anomaly types (a total of 11 main anomaly types, as shown in Fig. 3a). We use I3D as the backbone for all methods in these experiments. We observe that methods pretrained on our MSAD dataset generally achieve better performance on anomalies such as fighting, fire, object falling, shooting, traffic accidents, and water incidents. Overall, using our dataset leads to better performance (see the overall performance column) compared to using UCF-Crime. This indicates that existing datasets do not focus sufficiently on non-human-related anomalies, and our dataset helps bridge this gap. However, for some anomaly types, such as explosions, people falling, and vandalism, using the UCF-Crime dataset achieves slightly better performance. The potential reasons for this are: (i) the I3D backbone encoder's difficulty in capturing sudden motions, as it was originally designed for action recognition, and (ii) limitations of our dataset, which suggests the need for further collection and augmentation of video samples to improve model performance. We also note that there is no single best performer for all anomaly types, indicating that our dataset can support the VAD community in exploring and developing more robust anomaly detection algorithms.

Performance comparison by scenario. We now perform evaluations on scenario concepts, and the results are summarized in Table 8. As shown in the table, some scenarios, such as highways and trains, achieve quite low performance. The possible reasons for this are: (i) limited video data available for these scenarios, and (ii) these scenarios are challenging due to complex motions involving varying directions, speeds, different objects, human movements, *etc*. On average, the models pretrained on our dataset achieve better performance compared to those pretrained on UCF-Crime. This demonstrates that our dataset covers a diverse range of scenarios and could potentially facilitate multi-scenario anomaly detection, especially when the definition of anomalies is uncertain.

Cross-dataset evaluation. In real-world VAD systems, the model's ability to generalize to unseen scenarios is crucial. To evaluate this capability, we implement a zero-shot performance evaluation.

Table 7: Performance evaluations by anomaly type (a total of 11 main anomaly types) on our MSAD test set are conducted. We use frame-level Micro AUC (%) and Average Precision (AP, in %) as evaluation metrics for models pretrained on either UCF-Crime or our MSAD. We use I3D as the backbone for all methods. The best training scheme for each method is highlighted in bold.

Training set	Method	Assault		Explosion		Fighting		Fire	
		AUC	AP	AUC	AP	AUC	AP	AUC	AP
UCF-Crime	RTFM [76]	60.6	62.2	**69.3**	**79.0**	68.5	80.7	36.0	64.5
	MGFN [13]	**60.5**	**62.0**	65.5	**74.3**	53.6	63.9	21.6	55.0
	UR-DMU [102]	**59.4**	60.5	**69.3**	**82.0**	71.2	85.2	36.2	66.5
MSAD	RTFM [76]	**68.1**	**67.3**	46.8	60.4	**89.6**	**93.0**	**61.3**	**81.2**
	MGFN [13]	59.7	59.0	**64.5**	71.9	**89.4**	**93.5**	**86.0**	**93.0**
	UR-DMU [102]	56.9	**64.5**	67.9	74.5	**83.9**	**90.4**	**61.2**	**82.9**

Training set	Method	Object Falling		People Falling		Robbery		Shooting	
		AUC	AP	AUC	AP	AUC	AP	AUC	AP
UCF-Crime	RTFM [76]	82.0	88.8	**69.5**	**63.0**	**76.8**	**90.6**	59.7	65.7
	MGFN [13]	65.5	73.1	**57.2**	**59.5**	72.0	**89.1**	42.1	57.6
	UR-DMU [102]	72.4	76.5	**69.3**	**57.6**	**69.7**	**81.5**	59.9	73.8
MSAD	RTFM [76]	**94.7**	**96.7**	56.5	50.4	65.7	81.2	**78.2**	**84.7**
	MGFN [13]	**90.9**	**94.8**	52.7	47.8	**73.9**	86.7	**86.8**	**88.5**
	UR-DMU [102]	**92.1**	**95.8**	42.5	43.7	63.5	79.3	**81.4**	**87.8**

Training set	Method	Traffic Accident		Vandalism		Water Incident		**Overall**	
		AUC	AP	AUC	AP	AUC	AP	AUC	AP
UCF-Crime	RTFM [76]	55.6	45.1	**86.0**	**85.2**	93.5	98.5	71.9	47.4
	MGFN [13]	52.6	45.3	80.7	**81.4**	41.0	81.7	61.8	31.2
	UR-DMU [102]	53.0	47.9	**91.6**	**89.7**	64.6	91.3	74.3	53.4
MSAD	RTFM [76]	**62.2**	**51.8**	85.2	76.1	**96.3**	**99.1**	**86.7**	**66.3**
	MGFN [13]	**68.6**	**54.5**	**82.4**	80.1	**85.5**	**97.0**	**85.0**	**63.5**
	UR-DMU [102]	**62.0**	**55.6**	84.7	77.0	**98.5**	**99.5**	**85.0**	**68.3**

Specifically, we select four state-of-the-art weakly supervised methods: RTFM [76], UR-DMU [102], MGFN [13], and TEVAD [12], and train them on multi-scenario datasets, either UCF-Crime or our MSAD, using Protocol ii. Subsequently, we evaluate their performance on single-scenario datasets, such as ShanghaiTech, CUHK Avenue, and UCSD Ped2, without further training or fine-tuning. This approach allows us to test how well these pre-trained models adapt to new, previously unseen scenarios, which is essential for practical applications.

Following [76], we select the top three snippets with the largest feature magnitudes from both normal and abnormal videos to train a snippet classifier, using I3D as the backbone. The evaluation is carried out using the Micro AUC metric, with results recorded in Table 9.

As shown in the table, models pre-trained on our MSAD dataset generally have better zero-shot performance on the CUHK and UCSD Ped2 datasets. However, on ShanghaiTech, the performance is not as good. This issue may be due to the categorization of biking and driving as anomalies, which does not align with reality.

G A review of advances in datasets

From single-view to multi-view. Existing datasets fall into two main categories: single-view, such as UCSD Ped [79] and CUHK Avenue [35], and multi-view like ShanghaiTech [37] and UBnormal [2]. UCSD Ped [79] and CUHK Avenue [35] primarily originate from campus surveillance videos. ShanghaiTech [37], a pioneering multi-view dataset, has served as a benchmark for various methods. However, akin to other university datasets, it lacks object diversity and deviates from real-world scenarios regarding environment variations. UCF-Crime [72] was the first real-world dataset with multiple views. Despite encompassing diverse anomaly types, including abuse, explosions, stealing,

Table 8: Performance evaluations by scenario (a total of 14 scenarios) on our MSAD test set are conducted. We use frame-level Micro AUC (%) and Average Precision (AP, in %) as evaluation metrics for models pretrained on either UCF-Crime or our MSAD. We use I3D as the backbone for all methods. The best training scheme for each method is highlighted in bold.

Training set	Method	Frontdoor		Highway		Mall		Office	
		AUC	AP	AUC	AP	AUC	AP	AUC	AP
UCF-Crime	RTFM [76]	80.8	80.1	37.1	1.4	86.0	**87.1**	68.5	63.2
	MGFN [13]	68.4	70.2	36.3	1.4	**79.6**	**80.4**	64.5	60.2
	UR-DMU [102]	84.7	82.6	18.9	1.1	83.1	80.6	66.6	57.6
MSAD	RTFM [76]	**84.1**	**81.1**	**63.7**	**4.1**	**87.2**	72.2	**78.1**	**68.8**
	MGFN [13]	**86.4**	**85.1**	**79.7**	**4.1**	65.3	56.6	**75.1**	**62.4**
	UR-DMU [102]	**84.8**	**82.8**	**31.5**	**1.3**	**91.0**	**83.8**	**77.8**	**67.3**

Training set	Method	Park		Parkinglot		Pedestrian st.		Restaurant	
		AUC	AP	AUC	AP	AUC	AP	AUC	AP
UCF-Crime	RTFM [76]	**75.3**	23.7	66.7	16.7	84.1	**67.6**	66.5	56.5
	MGFN [13]	55.3	7.9	59.5	12.3	74.4	11.2	47.3	32.4
	UR-DMU [102]	**91.6**	34.8	62.2	17.6	58.5	6.1	75.7	74.4
MSAD	RTFM [76]	69.0	**25.6**	**74.4**	**35.9**	**97.4**	50.6	**96.1**	**91.9**
	MGFN [13]	**77.9**	**38.3**	**68.1**	**14.5**	**88.0**	**20.4**	**95.8**	**91.8**
	UR-DMU [102]	87.8	**36.2**	**91.4**	**53.9**	**81.9**	**11.5**	**93.1**	**87.4**

Training set	Method	Road		Shop		Sidewalk		Street highview	
		AUC	AP	AUC	AP	AUC	AP	AUC	AP
UCF-Crime	RTFM [76]	**82.9**	**47.1**	**85.1**	68.5	**89.1**	**66.1**	**82.6**	**35.9**
	MGFN [13]	54.4	18.3	69.4	60.4	47.4	26.4	37.2	8.3
	UR-DMU [102]	49.5	26.6	78.8	**66.5**	68.0	55.9	62.0	23.0
MSAD	RTFM [76]	54.0	16.8	80.6	**77.3**	52.5	17.1	43.3	12.3
	MGFN [13]	**77.9**	**49.7**	**84.9**	**77.2**	**85.5**	**62.3**	**87.6**	**40.7**
	UR-DMU [102]	**83.0**	**64.4**	**81.3**	64.5	**86.5**	**64.1**	**85.0**	**37.7**

Training set	Method	Train		Warehouse		**Overall**	
		AUC	AP	AUC	AP	AUC	AP
UCF-Crime	RTFM [76]	52.2	**5.0**	**82.3**	**52.8**	71.9	47.4
	MGFN [13]	39.8	2.1	55.4	18.3	61.8	31.2
	UR-DMU [102]	51.3	2.6	**86.9**	54.0	74.3	53.4
MSAD	RTFM [76]	**66.9**	3.9	69.5	37.4	**86.7**	**66.3**
	MGFN [13]	**53.0**	**3.1**	**72.3**	**30.9**	**85.0**	**63.5**
	UR-DMU [102]	**59.0**	**3.1**	81.2	**59.1**	**85.0**	**68.3**

Table 9: Comparison of cross-dataset results using four recent anomaly detection models with the I3D backbone. UCF, ShT, CUHK, and Ped2 denote UCF-Crime, ShanghaiTech, CUHK Avenue, and UCSD Ped2, respectively. Improvements from using models pre-trained on the MSAD dataset are highlighted in red, while performance drops are indicated in blue.

Method	UCF→ShT	UCF→CUHK	UCF→Ped2	**MSAD**→ShT	**MSAD**→CUHK	**MSAD**→Ped2
RTFM [76]	42.62	50.76	60.03	39.59 (↓3.03%)	63.23 (↑12.47%)	57.97 (↑2.06%)
UR-DMU [102]	46.69	45.67	62.90	35.05 (↓11.64%)	58.86 (↑13.19%)	66.84 (↑3.94%)
MGFN [13]	37.58	44.48	51.75	48.10 (↑10.52%)	56.66 (↑12.18%)	62.09 (↑10.34%)
TEVAD [12]	59.34	43.39	36.96	45.27 (↓14.07%)	64.82 (↑21.43%)	62.56 (↑25.60%)

and fighting, its quality is suboptimal due to monochrome video footage, low resolution, moving cameras, and redundancy. Still, most existing datasets focus on human-related anomalies.

From single-scenario to multi-scenario. Single-scenario datasets [6, 35, 79, 37, 63] are typically sourced from specific locations with one or multiple camera viewpoints, such as universities or urban streets, capturing routine activities and occasional anomalies. Early datasets, such as UCSD Ped [79] and CUHK Avenue [35], were collected from a single camera viewpoint at a university. Although single-scenario datasets are valuable, they lack diversity in scenarios, each with distinct backgrounds and anomalies. For instance, a parking lot typically has less movement compared to a crowded street, and a shopping mall is busier than a university avenue. Training a model exclusively on a single scenario limits its ability to generalize, as it only learns the specific features of that scenario and performs poorly when applied to new scenes [36]. To address the limited real-world diversity, multi-scenario datasets, such as UCF-Crime [72] and CUVA [14], have been collected from online video platforms (*e.g.*, YouTube), offering a wide range of scenarios. These variations in scenarios, with diverse weather and lighting conditions, greatly enhance the robustness and adaptability of anomaly detection models for real-world applications. UCF-Crime [72] has become the largest and most popular benchmark for weakly-supervised methods. The CUVA [14] dataset is another large-scale benchmark focused on the causation of video anomalies and is the first to include high-quality text descriptions of video content. This is designed to aid video large language models (VLMs) in understanding the cause and effect of anomalies. However, unlike UCF-Crime [72], which exclusively collects surveillance videos from websites, CUVA sources its videos from news outlets and handheld cameras. This results in significant viewpoint shifts that may affect performance in surveillance.

The role of synthetic datasets. As obtaining videos that contain anomalies is challenging in the real world, and while some datasets simulate anomalies, researchers have begun exploring the use of synthetic datasets to generate normal and abnormal videos. For example, UBnormal [2], generated using Cinema4D software, simulates diverse events with 268 training, 64 validation, and 211 testing samples. Importantly, UBnormal is often used to augment real-world datasets rather than for direct model training, thereby enriching the diversity of VAD tasks. While synthetic datasets can produce multiple scenarios with diverse motions, there are challenges: (i) simulating non-human-related anomalies is difficult, and (ii) the simulated motions may not accurately reflect real-world dynamics [41]. Hence, more real-world datasets are still required to train robust VAD models, even though synthetic datasets can alleviate the issue of insufficient training sources.

Expanding anomaly types. Most existing benchmark datasets focus on human-related anomaly detection. UCSD Ped, CUHK Avenue, and ShanghaiTech center on pedestrian activities, with anomaly types including biking, cart movement, loitering, running, throwing objects, chasing, brawling, sudden motions, *etc*. UCF-Crime and XD-Violence focus on crime and violence, with anomalies such as assault, burglary, robbery, fighting, and shooting. Recent datasets like CUVA and UBnormal introduce multiple scenarios and a broader range of anomaly types, extending beyond human behavior-related anomalies. These datasets include non-human-related anomalies such as smoking, fireworks, forest fires, *etc*. Most existing datasets overlook the detection of non-human-related anomalies, because humans are the main source of anomalies. Therefore, there is still a demand to develop comprehensive benchmarks and methods to advance VAD.

Datasets are becoming closer to real-life scenarios. Early datasets such as ShanghaiTech and UCSD Ped have some drawbacks, including low resolution, grayscale imagery, and questionable anomaly types. For instance, activities like cycling, running, or skating on the street may be considered normal in these datasets. While datasets like Street Scene [56] and IITB Corridor [63] have expanded diversity with more frames and anomaly events, they are still not representative enough for learning real-world normality. These datasets generally neglect scene-dependent anomalies (events that are normal in one scene but abnormal in another) [6]. To address these limitations, the new NWPU Campus dataset introduced in [6] contains 43 scenes (camera views). Additionally, the recent CUVA dataset introduced in [14] serves as a large-scale, comprehensive benchmark for understanding the causation of video anomalies. With high-quality annotations, this dataset advances prompt-based VAD approaches, bringing datasets closer to real-life scenarios. However, since most videos in this

dataset come from news sources, they suffer from dynamic subtitles, frequent scene changes, *etc*. Fig. 1 shows a visual comparison of some VAD datasets. As shown in this comparison, the datasets are increasingly resembling real-life scenarios in terms of resolution, scenario complexity, anomaly types, and external variations.

H Limitation and future work

Although our dataset covers a wide range of scenarios and anomaly types, there are still more scenarios and anomaly types we need to consider for real-world applications. Some anomaly types, such as uncontrolled animals and sudden crowd gatherings, are still missing from our dataset. Our future work will focus on collecting more scenarios and anomaly types.

Videos for industrial scenarios, like adherence to safety protocols such as wearing hard hats or operating equipment safely, are difficult to obtain due to privacy issues. We will explore alternative solutions for these anomalies. Moreover, we will continue refining our dataset for existing anomalies, such as no right turns at specific crossings, for traffic-related anomalies.

Additionally, we will provide more modalities to further expand our dataset, such as audio, video captions, and human skeletons. Given recent advances, there is a greater focus on reusable large-scale pretrained models and multi-modal fusion for robust VAD.

I Potential societal impact

Our new dataset will enhance robust VAD, providing support and guidance to improve the safety and security of our communities.

However, since our dataset contains human identity information, it is crucial to explore methods that balance VAD performance while removing personally identifiable information, such as faces, poses, or gaits of individuals, even though we have obtained permissions from YouTube, Itemfix, and Bilibili to use their videos for research purposes. Our long-term plan is to explore new data modalities, such as the use of audio and skeleton sequences that do not contain human identity information, to advance VAD.