Advancing Anomaly Detection: An Adaptation Model and a New Dataset

Liyun Zhu, **Arjun Raj** and **Lei Wang**
Australian National University
{u7778917, u7526852, lei.w}@anu.edu.au

Abstract

Industry surveillance is widely applicable in sectors like retail, manufacturing, education, and smart cities, each presenting unique anomalies requiring specialized detection. However, adapting anomaly detection models to novel viewpoints within the same scenario poses challenges. Extending these models to entirely new scenarios necessitates retraining or fine-tuning, a process that can be time-consuming. To address these challenges, we propose the **S**cenario-**A**daptive **A**nomaly **D**etection (**SA^{2}D**) method, leveraging the few-shot learning framework for faster adaptation of pre-trained models to new concepts. Despite this approach, a significant challenge emerges from the absence of a comprehensive dataset with diverse scenarios and camera views. In response, we introduce the **M**ulti-**S**cenario **A**nomaly **D**etection (**MSAD**) dataset, encompassing 14 distinct scenarios captured from various camera views. This real-world dataset is the first high-resolution anomaly detection dataset, offering a solid foundation for training superior models. MSAD includes diverse normal motion patterns, incorporating challenging variations like different lighting and weather conditions. Through experimentation, we validate the efficacy of SA^{2}D, particularly when trained on the MSAD dataset. Our results show that SA^{2}D not only excels under novel viewpoints within the same scenario but also demonstrates competitive performance when faced with entirely new scenarios. This highlights our method's potential in addressing challenges in detecting anomalies across diverse and evolving surveillance scenarios.

1 Introduction

Anomaly detection in surveillance videos presents a formidable challenge, requiring the prediction of anomalies that may manifest ambiguously in diverse scenarios [31]. Consider, for instance, the distinction between normal activities like walking on a sidewalk and abnormal activities like walking on a highway – a context-dependent determination.

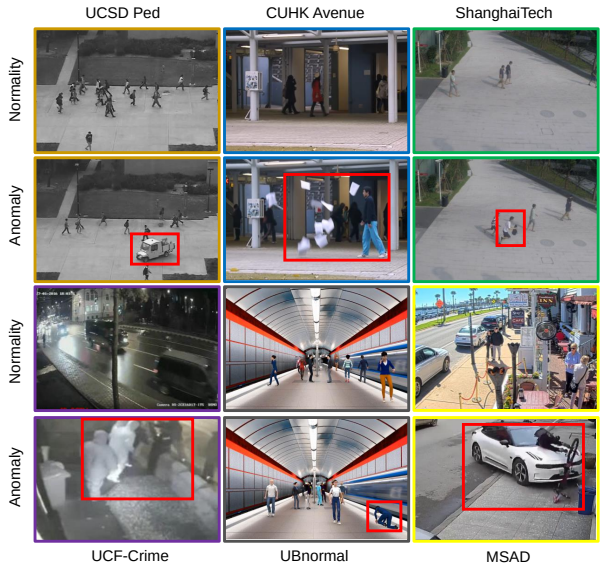

Figure 1: A comparison of our Multi-Scenario Anomaly Detection (MSAD) dataset *vs*. other existing datasets such as UCSD Ped, CUHK Avenue, ShanghaiTech, UCF-Crime and UBnormal. Our dataset (i) has higher resolution (ii) covers a wide range of scenarios and (iii) matches with the real-world scenarios.

Given the difficulty in obtaining anomaly-labeled videos, prevailing methods often treat anomaly detection as a one-class classification problem, relying on training solely with normal samples and testing on both normal and abnormal samples, treating anomalies as outliers.

High-quality videos are crucial for effective training, and commonly used datasets include CUHK Avenue [12], UCSD Ped [24], ShanghaiTech [14], UCF-Crime [23], and UBnormal [1]. However, these datasets exhibit limitations. Firstly, many lack high resolution, while contemporary surveillance videos are characterized by rich color information and higher quality, essential for capturing detailed object-level information. Secondly, existing datasets often fail to represent real-world scenarios accurately. Examples like CUHK Avenue and UCSD Ped are recorded in controlled environments with few pedestrians and vehicles, limiting their applicability to complex real-world scenarios with crowded spaces and heavy traffic. Moreover, challenging conditions like diverse light-

Dataset	Year	#Training	#Testing	#Scenario	Source	Resolution	Variations	Human-Related	Non-Human Related
UCSD Ped1	2010	34	36	1	Surveillance	238×158	✗	✓	✗
UCSD Ped2	2010	16	12	1	Surveillance	238×158	✗	✓	✗
CUHK Avenue	2013	16	21	1	Self-recorded	640×360	✗	✓	✗
ShanghaiTech	2017	330	107	1	Surveillance	856×480	✗	✓	✗
UCF-Crime	2018	**1610**	**290**	-	YouTube/LiveLeak	320×240	✓	✓	✗
UBnormal	2022	332	211	8	3D modeling	1080×720	✓	✓	✓
MSAD (ours)	2024	358	75	**14**	YouTube/Surveillance	$\mathbf{1920\times1080}$	✓	✓	✓

Table 1: A comparison of our newly introduced dataset *vs*. existing anomaly detection datasets in terms of the number of training videos, test videos, total number of scenarios, source, resolution, and whether it contains (i) challenging dynamic environments (Variations) and (ii) human-related and/or non-human related anomalies. Compared to existing datasets, ours is more comprehensive, offering high-resolution videos across a diverse range of scenarios, including challenging dynamics such as variations in weather and lighting.

ing, weather changes, and potential camera motions are often overlooked. Thirdly, anomalies in some datasets may differ from real-world occurrences; for instance, activities like cycling, running, or skating on the street may be deemed normal. Furthermore, human-centric anomalies are emphasized, neglecting non-human-related anomalies such as falling objects, fires, or water leaks, which can also result in significant disasters. Most critically, existing datasets lack diversity in scenarios, focusing primarily on different camera views within the same scenario rather than encompassing multiple scenarios. A single scenario may include various camera views, emphasizing the need for models trained with samples from diverse scenarios and fine-tuned for specific ones to enhance generalization capabilities.

To address these challenges, we introduce the **M**ulti-**S**cenario **A**nomaly **D**etection (**MSAD**) dataset. This real-world dataset comprises surveillance videos from 14 distinct scenarios, including roads, malls, parks, sidewalks, and more. It incorporates various objects like pedestrians, cars, trunks, and trains, along with dynamic environmental factors such as different lighting and weather conditions. Notably, MSAD stands out as the first high-resolution anomaly detection dataset offering a comprehensive representation of real-world scenarios. In tandem with the dataset, we propose a novel anomaly detection model called **S**cenario-**A**daptive **A**nomaly **D**etection (**SA^{2}D**). Recognizing the challenges of adapting pre-trained models to novel viewpoints within the same scenario and extending them to entirely new scenarios, SA^{2}D leverages a few-shot learning framework [3]. This approach facilitates rapid adaptation of a universal model to new concepts, allowing for easy reuse and fine-tuning for new camera viewpoints and scenarios. To validate the efficacy of SA^{2}D, we conducted experiments using the MSAD dataset. Our results demonstrate that SA^{2}D achieves competitive performance not only under novel viewpoints within the same scenario but also when faced with entirely new scenarios. This emphasizes the model's capability to generalize effectively across diverse scenarios, showcasing its potential as a robust and versatile anomaly detection solution. We summarize the main points below.

i. We present the Multi-Scenario Anomaly Detection (MSAD) dataset, a high-resolution, real-world anomaly detection dataset encompassing diverse scenarios and anomalies, both human and non-human-related.

ii. Alongside the dataset, we propose a novel Scenario-Adaptive Anomaly Detection (SA^{2}D) model, leveraging a few-shot learning framework for efficient adaptation to new concepts and scenarios.

iii. Our experiments demonstrate that SA^{2}D achieves competitive performance, highlighting its effectiveness not only within the same scenario but also in entirely new scenarios, showcasing its potential for robust and versatile anomaly detection.

2 Related Work

Revisiting anomaly detection methods. Many recent works approach anomaly detection as a one-class classification problem, relying on the features of normal events. These methods exclusively train on normal videos and subsequently test on both normal and abnormal videos. For example, reconstruction-based techniques, such as those employing Auto-Encoder [6] and Generative Adversarial Network (GAN) [41], aim to reconstruct normal events to minimize the reconstruction loss. However, the assumption that these methods lack an understanding of normal events is not always accurate, as they can sometimes generalize well to anomalies [19]. It's crucial to note that achieving a good reconstruction does not necessarily guarantee effective anomaly detection, as these methods can be sensitive to object speeds, often treating sudden motions as anomalies.

Prediction-based methods enhance model discrimination by incorporating complementary information. These models predict various tasks like future frame prediction [11, 13, 2], middle box prediction [4], and puzzle solving [35]. The predictive convolution attentive block [20], using dilated convolutions with a masked central area, has shown promise for anomaly detection. Distance-based methods focus on learning a distance function between normal and abnormal videos. In [18], video patches are employed to learn a distance function for anomaly localization. Probability-based approaches utilize features like motion patterns [9] or pose sequences [8] to model the distribution of normal events and estimate abnormal probabilities during inference. While effective, these methods, relying on high-level features for video representation, often have limited interpretability. In addition to self-

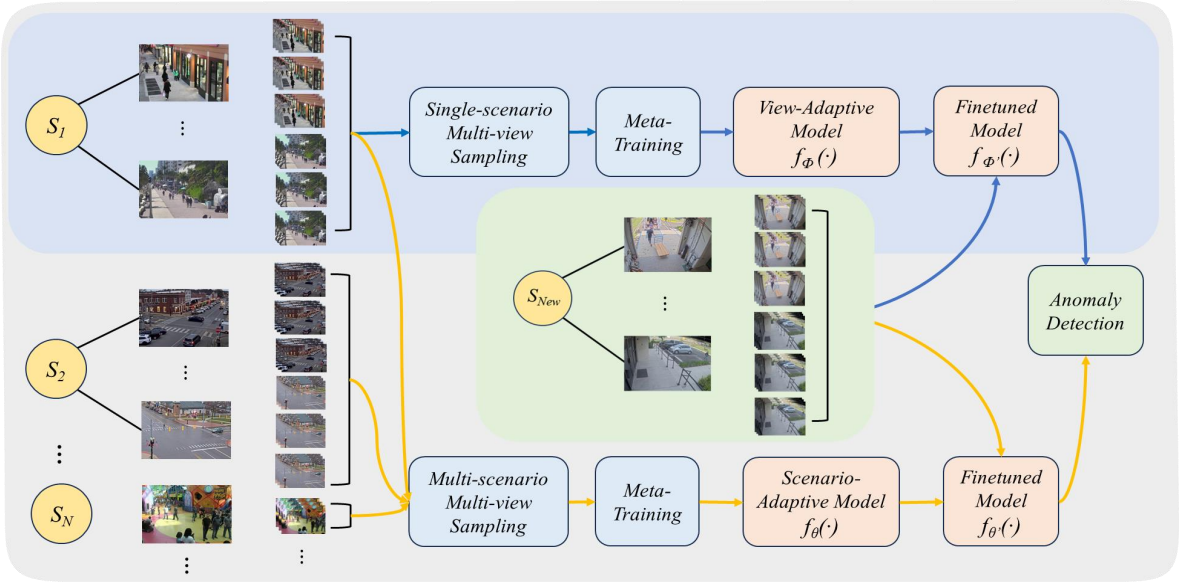

Figure 2: A comparison between the existing few-shot scene-adaptive (view-adaptive) anomaly detection model (depicted by the light blue block) and our proposed Scenario-Adaptive Anomaly Detection (SA^{2}D) model (illustrated by the light gray block). On the left-hand side of the figure, the first column, $\{S_1, S_2, \cdots, S_N\}$, 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.

supervised approaches, weakly supervised methods [23, 40] are employed. Trained on both normal and abnormal videos, these methods annotate abnormal videos at the video level, indicating the presence of abnormal behavior without specifying its exact location.

In contrast to existing methods, our SA^{2}D model takes a unique approach, utilizing a few-shot learning framework for faster adaptation to novel concepts and scenarios. This adaptability, coupled with its competitive performance on diverse scenarios, sets SA^{2}D apart from traditional models.

Comparing anomaly detection datasets: A review. Various research groups have developed datasets to facilitate the creation of anomaly detection models. Figure 1 showcases samples from both existing datasets and our newly introduced dataset, while Table 1 offers statistical insights into these datasets. Existing datasets fall into two main categories: single-view datasets, such as UCSD Ped1, Ped2, and CUHK Avenue [24, 12], and multi-view datasets like ShanghaiTech and UBnormal. Unlike many datasets with fixed viewpoints and limited scenarios, our MSAD dataset boasts a broader range of scenarios and camera viewpoints.

Single-scenario datasets, like UCSD Ped [24] and CUHK Avenue [12], primarily originate from campus surveillance videos. UCSD Ped, captures scenes on campus streets but is outdated with monochrome videos. Similarly, CUHK Avenue, recorded at a university, includes 16 training videos and 21 testing videos, defining anomalies as running, throwing objects, and loitering. Both datasets feature pixel-level annotations. ShanghaiTech [14], a pioneering multi-view dataset with 13 camera views, has served as a benchmark for various methods. However, akin to other university datasets, it lacks object diversity and deviates from real-world scenarios. UCF-Crime [23] was the first real-world dataset with multiple views, comprising 1900 videos from YouTube and LiveLeak. Despite encompassing diverse anomaly types, including abuse, explosions, stealing, and fighting, its quality is suboptimal due to monochrome video footage, low resolution, moving cameras, and redundancy. UBnormal [1], 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.

These datasets employ different annotation methods, including video-level, frame-level, and pixel-level annotations, corresponding to distinct evaluation metrics. To balance complexity and cost, our MSAD dataset opts for frame-level annotations, ensuring accuracy in identifying abnormal clips.

3 Our Method: Scenario-Adaptive Anomaly Detection

Below we present our **S**cenario-**A**daptive **A**nomaly **D**etection (**SA^{2}D**) model. First, we introduce the notations.

Notations: $\mathcal{I}_K$ represents the index set $1, 2, \cdots, K$. Calligraphic mathcal fonts denote tensors (*e.g.*, $\boldsymbol{\mathcal{V}}$), 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

Figure 3: Some video frames selected from our training set. Our dataset includes a diverse range of scenarios, both indoor and outdoor, featuring various objects such as 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.

model [13]. This model leverages the meta-learning framework [3, 34, 26, 27] to train a model using video data collected from various camera views within the same scenario[1], 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 [13], 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. 2 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 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 $f_\theta : \mathbf{X} \rightarrow \boldsymbol{Y}$.

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 [21] for future frame prediction, incorporating a ConvLSTM block [22] for sequential modeling to retain long-term temporal information. Following [11] and [13, 36], we incorporate GAN [7] 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 [3, 13], 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 [13], we define $L(\cdot)$ as the summation of the least absolute deviation (L_1 loss) [16], multi-scale structural similarity measurement [30], and the gradient difference loss [15]. The updated model parameters θ' are

[1] 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.

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:

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

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.

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 [11], 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 the frame is well predicted. Following [15], 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.

4 Our Dataset: Multi-Scenario Anomaly Detection Dataset

We present the **M**ulti-**S**cenario **A**nomaly **D**etection (**MSAD**) dataset tailored for our few-shot multi-scenario multi-view learning framework.

Resolution. All the videos we've collected are high resolution, such as 1920×1080. 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.

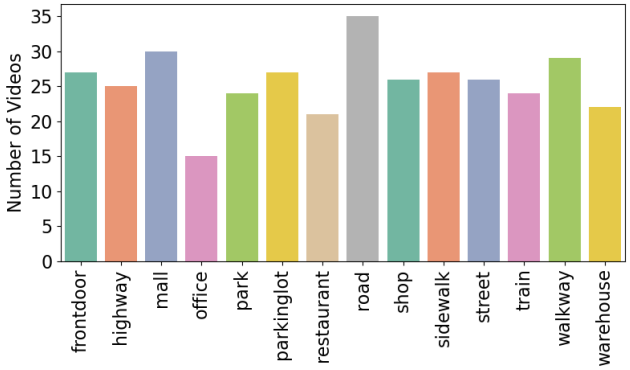

Figure 4: Distribution of videos per scenario in our MSAD training set.

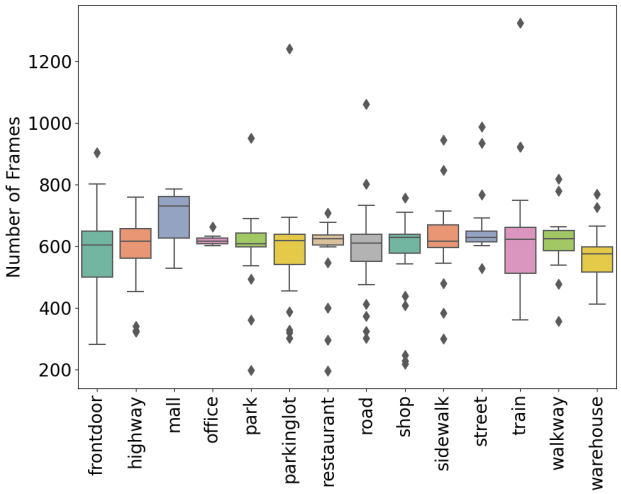

Figure 5: Boxplot illustrating frame number variations across scenarios in our MSAD training set.

Scenario. Traditional datasets commonly define a scene as the perspective captured by a camera, often referred to as a camera viewpoint. For instance, the ShanghaiTech dataset [14] 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. Figure 3 showcases 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-related *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 [38, 31, 33, 32, 25, 17, 28, 29, 39, 37, 10]. 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 such as water leaks, fires, and other non-human-related events.

Dataset statistics. As illustrated in Table 1, our MSAD dataset comprises 358 training videos and 75 testing videos, sourced from YouTube and real-world surveillance videos. We meticulously removed watermarks and texts to mitigate potential performance issues caused by visual distractions. In the context of anomaly detection, extracting dynamic frames from our training videos is crucial for ensuring representation. Thus, during video processing, we focus on selecting dynamic video clips while maintaining each video's duration around 15-20 seconds. 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.

The distribution of video numbers across different scenarios is depicted in Figure 4, while Figure 5 provides frame counts per scenario. In Figure 4, there are approximately 25 videos per scenario, and the total number of videos per scenario is well-balanced. Figure 5 shows an average of around 600 frames per video scenario. However, significant variation, particularly with outliers in the parking lot and road categories, is observed. This variation may be attributed to the slightly varying lengths of videos. The boxplot representing variation within the 'office' category exhibits an unusually small spread, likely due to a limited number of captured videos originating from a small set of locations that generally share the same camera settings.

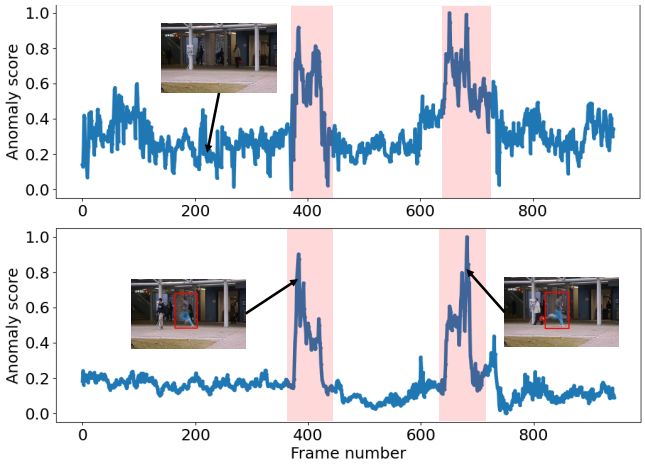

Figure 6: Frame-level anomaly scores (normalised between 0 and 1) are depicted for a test video from CUHK Avenue. The light red color block highlights the time period when the anomaly occurs. The top row illustrates the curve generated by the view-adaptive model, while the bottom row displays our scenario-adaptive 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.

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.

Evaluation metrics. Consistent with prior studies [11, 13], we employ the area under the curve (AUC) as our evaluation metric. However, it is important to note that two approaches exist for computing frame-level AUC, as discussed in [5]. The first approach involves concatenating frames from all the samples and calculating the overall AUC score, referred to as Micro AUC. The second approach computes individual scores for each video's frames and then averages them, known as Macro AUC. We consider both Micro and Macro AUC scores.

5 Experiment

5.1 Setup

We first show the superior performance of our scenario-adaptive model by comparing SA^{2}D with the few-shot scene adaptive model [13] on CUHK Avenue. To demonstrate the enhanced anomaly detection performance of our MSAD dataset, we train two adaptive models: one utilizing ShanghaiTech, and the other leveraging our MSAD. We conduct two sets of experiments for both models:

i. 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 [13], we partition the ShanghaiTech dataset into 7 scenes (views), specifically, scene

2, 4, 7, 9, 10, 11, and 12 for training as in [13]. Subsequently, the remaining 5 camera views (there are no test videos for scene 13, so we only consider 5 views during testing) are individually used for testing purposes.

ii. In the *cross-scenario evaluation*, the model is trained on one scenario and tested on a completely different one. For model training, we employ ShanghaiTech (or our MSAD), while testing is performed on the test sets from UCSD Ped and CUHK Avenue.

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 adhere to our predefined evaluation metrics. Below we show the experimental results, comparisons and discussions.

5.2 Evaluation

Scenario-adaptive outperforms view-adaptive. We commence by comparing our SA^{2}D with the view-adaptive model (few-shot scene adaptive [13]), utilizing test videos from CUHK Avenue for evaluation. In Fig. 6, we present the visualization of frame-level anomaly scores for both models. The illustration indicates that our SA^{2}D (Fig. 6 *bottom*) produces significantly smoother curves compared to the view-adaptive model (Fig. 6 *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.

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 single-scenario/cross-view evaluation falls short compared to our SA^{2}D. Notably, our SA^{2}D, trained on our MSAD dataset 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 observations reveal that our model, trained on MSAD, demonstrates superior performance compared to the view-adaptive model. 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 on MSAD 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.

Further insights. 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.

Training set	Testing set	AUC			
		Micro		Macro	
ShanghaiTech	ShanghaiTech ($v1$)	61.36		55.34	
	ShanghaiTech ($v3$)	26.51		26.58	
	ShanghaiTech ($v5$)	53.40		53.32	
	ShanghaiTech ($v6$)	78.36		78.27	
	ShanghaiTech ($v8$)	50.02		52.54	
	UCSD Ped2	57.38		58.36	
	CUHK Avenue	69.98		78.32	
	MSAD	63.92		64.92	
MSAD	ShanghaiTech ($v1$)	68.96	↑7.6	77.89	↑22.6
	ShanghaiTech ($v3$)	67.59	↑41.1	73.43	↑46.9
	ShanghaiTech ($v5$)	55.74	↑2.3	54.02	↑0.7
	ShanghaiTech ($v6$)	75.47	↓2.9	72.35	↓5.9
	ShanghaiTech ($v8$)	60.85	↑10.8	61.52	↑9.0
	UCSD Ped2	70.35	↑13.0	65.74	↑7.4
	CUHK Avenue	79.57	↑9.6	84.49	↑6.2
	MSAD	69.96	↑6.0	69.60	↑4.7

Table 2: Experimental results on single-scenario/cross-view and cross-scenario evaluations. The light gray color highlights the cross-scenario experiments. Although we train on our MSAD training set and test on the MSAD test set, these experiments are still considered cross-scenario. This is because the scenarios in our test set are disjoint from those in our training set. On ShanghaiTech, only 7 views are used during training and the rest views are individually used for testing. The notation $v*$ in parentheses denotes the use of different camera views. The enhancements in comparison to the view-adaptive model are indicated in red.

6 Conclusion

In this paper, we address the challenges in anomaly detection through the introduction of the **M**ulti-**S**cenario **A**nomaly **D**etection (**MSAD**) dataset and the innovative **S**cenario-**A**daptive **A**nomaly **D**etection (**SA^{2}D**) model. Our MSAD dataset is the first dataset of its kind, providing a solid foundation for training advanced models in anomaly detection. Along with the dataset, our SA^{2}D model leverages the few-shot learning framework to efficiently adapt to new concepts and scenarios. The experimental results showcase the model's robustness, excelling not only new camera views of the same scenario but also in novel scenarios.

Acknowledgements

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, with computational resources provided by NCI Australia, an NCRIS-enabled capability supported by the Australian Government.

References

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak 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.

[2] Congqi Cao, Yue Lu, and Yanning Zhang. Context recovery and knowledge retrieval: A novel two-stream framework for video anomaly detection. *arXiv preprint arXiv:2209.02899*, 2022.

[3] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[4] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak 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.

[5] Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak 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.

[6] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den 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.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

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

[9] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling 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.

[10] Piotr Koniusz, Lei Wang, and Anoop Cherian. Tensor representations for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[11] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua 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.

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

[13] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang 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.

[14] Weixin Luo, Wen Liu, and Shenghua 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.

[15] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

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

[17] Zhenyue Qin, Yang Liu, Pan Ji, Dongwoo Kim, Lei Wang, Bob McKay, Saeed Anwar, and Tom Gedeon. Fusing higher-order features in graph neural networks for skeleton-based action recognition. *IEEE TNNLS*, 2022.

[18] Bharathkumar Ramachandra, Michael Jones, and Ranga 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.

[19] Tal Reiss and Yedid Hoshen. Attribute-based representations for accurate and interpretable video anomaly detection. *arXiv preprint arXiv:2212.00789*, 2022.

[20] Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak 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.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas 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.

[22] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[23] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Pro-*

ceedings of the IEEE conference on computer vision and pattern recognition, pages 6479–6488, 2018.

[24] Mahadevan Vijay, Wei-Xin LI, Bhalodia Viral, and Vasconcelos Nuno. Anomaly detection in crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1975–1981, 2010.

[25] Lei Wang and Piotr Koniusz. Self-supervising action recognition by statistical moment and subspace descriptors. In *ACM-MM*, page 4324–4333, 2021.

[26] Lei Wang and Piotr Koniusz. Temporal-viewpoint transportation plan for skeletal few-shot action recognition. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 4176–4193, December 2022.

[27] Lei Wang and Piotr Koniusz. Uncertainty-dtw for time series and sequences. In *European Conference on Computer Vision*, pages 176–195. Springer, 2022.

[28] Lei Wang and Piotr Koniusz. 3Mformer: Multi-order multi-mode transformer for skeletal action recognition. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, June 2023.

[29] Lei Wang and Piotr Koniusz. Flow dynamics correction for action recognition. *ICASSP*, 2024.

[30] Zhou Wang, Eero P Simoncelli, and Alan 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.

[31] Lei Wang, Du Q Huynh, and Moussa Reda Mansour. Loss switching fusion with similarity search for video classification. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 974–978. IEEE, 2019.

[32] Lei Wang, Piotr Koniusz, and Du Q. Huynh. Hallucinating IDT descriptors and I3D optical flow features for action recognition with cnns. In *ICCV*, 2019.

[33] Lei Wang, Du Q. Huynh, and Piotr Koniusz. A comparative review of recent kinect-based action recognition algorithms. *IEEE TIP*, 29:15–28, 2020.

[34] Lei Wang, Jun Liu, and Piotr Koniusz. 3D skeleton-based few-shot action recognition with JEANIE is not so naïve. *arXiv preprint arXiv: 2112.12668*, 2021.

[35] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *European Conference on Computer Vision*, pages 494–511. Springer, 2022.

[36] Ze Wang, Yipin Zhou, Rui Wang, Tsung-Yu Lin, Ashish Shah, and Ser Nam Lim. Few-shot fast-adaptive anomaly detection. *Advances in Neural Information Processing Systems*, 35:4957–4970, 2022.

[37] Lei Wang, Ke Sun, and Piotr Koniusz. High-order tensor pooling with attention for action recognition. *ICASSP*, 2024.

[38] Lei Wang. Analysis and evaluation of Kinect-based action recognition algorithms. Master's thesis, School of the Computer Science and Software Engineering, The University of Western Australia, 2017.

[39] Lei Wang. *Robust Human Action Modelling*. PhD thesis, The Australian National University, Nov 2023.

[40] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *European Conference on Computer Vision*, pages 729–745. Springer, 2022.

[41] Jongmin Yu, Younkwan Lee, Kin Choong Yow, Moongu Jeon, and Witold Pedrycz. Abnormal event detection and localization via adversarial event prediction. *IEEE transactions on neural networks and learning systems*, 33(8):3572–3586, 2021.