Track Any Anomalous Object: A Granular Video Anomaly Detection Pipeline

Yuzhi Huang[1*], Chenxin Li[2*†], Haitao Zhang[1], Zixu Lin[1], Yunlong Lin[1],
Hengyu Liu[2], Wuyang Li[2], Xinyu Liu[2], Jiechao Gao[3], Yue Huang[1†], Xinghao Ding[1] and Yixuan Yuan[2]

[1] Key Laboratory of Multimedia Trusted Perception and Efficien Computing,
Ministry of Education of China, Xiamen University
School of Informatics, Xiamen University
[2] The Chinese University of Hong Kong   [3] The University of Virginia

Abstract

Video anomaly detection (VAD) is crucial in scenarios such as surveillance and autonomous driving, where timely detection of unexpected activities is essential. Albeit existing methods have primarily focused on detecting anomalous objects in videos—either by identifying anomalous frames or objects—they often neglect finer-grained analysis, such as anomalous pixels, which limits their ability to capture a broader range of anomalies. To address this challenge, we propose an innovative VAD framework called Track Any Anomalous Object (TAO), which introduces a Granular Video Anomaly Detection Framework that, for the first time, integrates the detection of multiple fine-grained anomalous objects into a unified framework. Unlike methods that assign anomaly scores to every pixel at each moment, our approach transforms the problem into pixel-level tracking of anomalous objects. By linking anomaly scores to subsequent tasks such as image segmentation and video tracking, our method eliminates the need for threshold selection and achieves more precise anomaly localization, even in long and challenging video sequences. Experiments on extensive datasets demonstrate that TAO achieves state-of-the-art performance, setting a new progress for VAD by providing a practical, granular, and holistic solution. For more information, visit the project page at: https://tao-25.github.io/

1. Introduction

Video Anomaly Detection (VAD) involves identifying unusual or unexpected activities in surveillance videos and has significant applications in areas such as security monitoring (e.g., detecting violent behavior) and autonomous driving (e.g., recognizing traffic accidents). Current VAD research has developed along two main directions, which
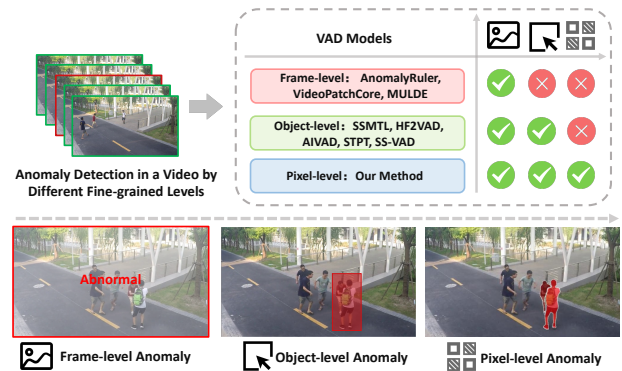
---

* Equal contribution   † Corresponding author

Figure 1. **Analyzing the Limitations of Existing VAD Models**. Video anomaly detection (VAD) models are predominantly frame-centric or object-centric. Frame-centric methods detect anomalies in frames without localizing them, while object-centric methods identify anomalous objects but lack pixel-level accuracy. Pixel-centric models address these gaps by providing pixel-level localization, delivering fine-grained segmentation and precise delineation of anomalies, particularly for overlapping objects where traditional methods struggle.

are reflected in both methodologies and evaluation benchmarks. Frame-centric methods [1, 2, 11, 32, 34, 39, 44, 57] and their corresponding frame-level benchmarks [18, 36, 49, 57] focus on analyzing entire frames to detect global anomalies (e.g., fires or smoke), typically evaluated using frame-level AUC metrics. In contrast, object-centric methods [3, 8–10, 19, 38, 41, 43, 48] and their object-level benchmarks [10, 19, 40] leverage pre-trained feature extractors to detect object-specific anomalies (e.g., human falls or vehicle accidents), evaluated using metrics such as TBDC [40] and RBDC [40], which comprehensively measure temporal consistency and spatial localization precision.

Despite these advances, a critical gap remains in real-world applications of VAD where fine-grained localization of anomalous objects is essential. As illustrated in Fig. 1,

current approaches suffer from fundamental shortage as frame-centric methods can only identify the presence of anomalies without localizing specific regions, while object-centric methods, though more precise, lack pixel-level accuracy. These limitations become particularly pronounced in complex scenarios with overlapping anomalies, where accurate delineation of object boundaries and shapes is crucial [4, 12, 14, 20, 28, 46, 50]. This observation motivates us to consider a more challenging yet practical scenario: Can we achieve both object-level structural integrity and pixel-level precision in a granular anomaly detection?

To address this challenge, we propose enhancing existing VAD benchmarks with pixel-level anomaly evaluation. While a straightforward approach would be to employ an anomaly segmentation task for pixel-wise classification [6, 13, 16, 22, 27, 62], this method falls short when dealing with multiple, potentially overlapping anomalous objects. Instead, inspired by instance segmentation techniques [15, 17, 24, 45, 52], we develop a novel evaluation framework that simultaneously considers both object-level and pixel-level anomaly detection accuracy. This framework ensures that detected anomalies accurately correspond to their true shapes, precise positions, and spatial distributions, while effectively minimizing false positives caused by local feature similarities or noise.

However, implementing precise pixel-level tracking in videos presents significant challenges, particularly in maintaining semantic and temporal consistency across frames. Traditional supervised approaches require extensive labeled data, which is scarce in existing datasets with pixel-level annotations. To overcome these challenges, we leverage SAM2 [42], a large-scale pre-trained segmentation model capable of processing both static images and video streams in real-time without requiring additional fine-tuning on specific anomaly datasets. Its ability to precisely delineate object contours and structures makes it particularly effective in handling complex scenarios involving overlapping anomalies or occlusions.

Building upon these foundations, we propose *TAO*, an integrated framework that combines object-centric anomaly detection algorithms with SAM2 to create a simple yet effective fine-grained video anomaly detection system. For videos containing anomalous frames, our object-centric detection model generates bounding boxes for detected anomalies. After robust filtering, these boxes serve as prompting boxes for SAM2, which then generates segmentation masks for the anomalous objects. Evaluated on our proposed benchmark, our method achieves state-of-the-art performance on two widely-used datasets, demonstrating its effectiveness in bridging the gap between object-level detection and pixel-precise segmentation. The contributions of our paper are as follows:

- We present a new testing standard that unifies pixel-level and object-level assessment in video anomaly detection, addressing the limitations of existing metrics in complex scenarios with multiple or overlapping anomalies.
- We introduce *TAO*, a streamlined framework that integrates object-centric anomaly detection with the segmentation capabilities from vision foundation model, enabling precise pixel-level tracking of anomalous objects without requiring additional fine-tuning.
- Through extensive experiments on UCSD Ped2 and ShanghaiTech Campus datasets, we demonstrate that our approach achieves state-of-the-art performance under both traditional metrics and our proposed benchmark.

2. Related Work

Video Anomaly Detection. Video anomaly detection (VAD) presents significant challenges due to the rarity of anomaly data and the wide variety of abnormal events, which hinder the generalization capabilities of existing models across diverse scenarios. Traditional VAD methods can be broadly classified into frame-centric and object-centric approaches. Frame-centric methods [1, 11, 32, 34, 39, 44, 55–57] analyze entire frames or sequences, often leveraging reconstruction or prediction errors to identify anomalies. These methods are particularly effective for detecting global events, such as fires or smoke. However, they struggle with localized anomalies and are prone to interference from normal regions within a frame, limiting their precision in complex scenarios [11, 32, 34, 39, 53]. Object-centric methods [7–10, 19, 41, 43, 54, 58] focus on specific objects within frames by employing pre-trained object detection models to extract bounding boxes and assess their abnormality. This targeted approach is better suited for detecting anomalies related to individuals or objects, such as human falls or vehicle accidents. By concentrating on specific regions and reducing redundant information, object-centric methods achieve improved accuracy and robustness, especially in complex, multi-object environments. Building on these foundations, our approach is the first to integrate large-scale pre-trained models for pixel-level fine-grained detection in VAD. This advancement enables precise anomaly localization, bridging the gap between frame-level analysis and object-level detection, and significantly enhancing overall detection performance.

Vision Foundation Models. In recent years, large-scale pre-trained language models (LLMs) and vision-language models (VLMs) have shown remarkable potential in advancing video anomaly detection. Models such as BLIP-2 [26], LLaVA [29], LAVAD [59], VadCLIP [51], Video-ChatGPT [35], and Video-LLaMA [61] have significantly improved the understanding of complex visual tasks by integrating visual and linguistic information. Concurrently, prompting techniques have gained traction in large-scale vision models, enabling substantial progress in tasks such
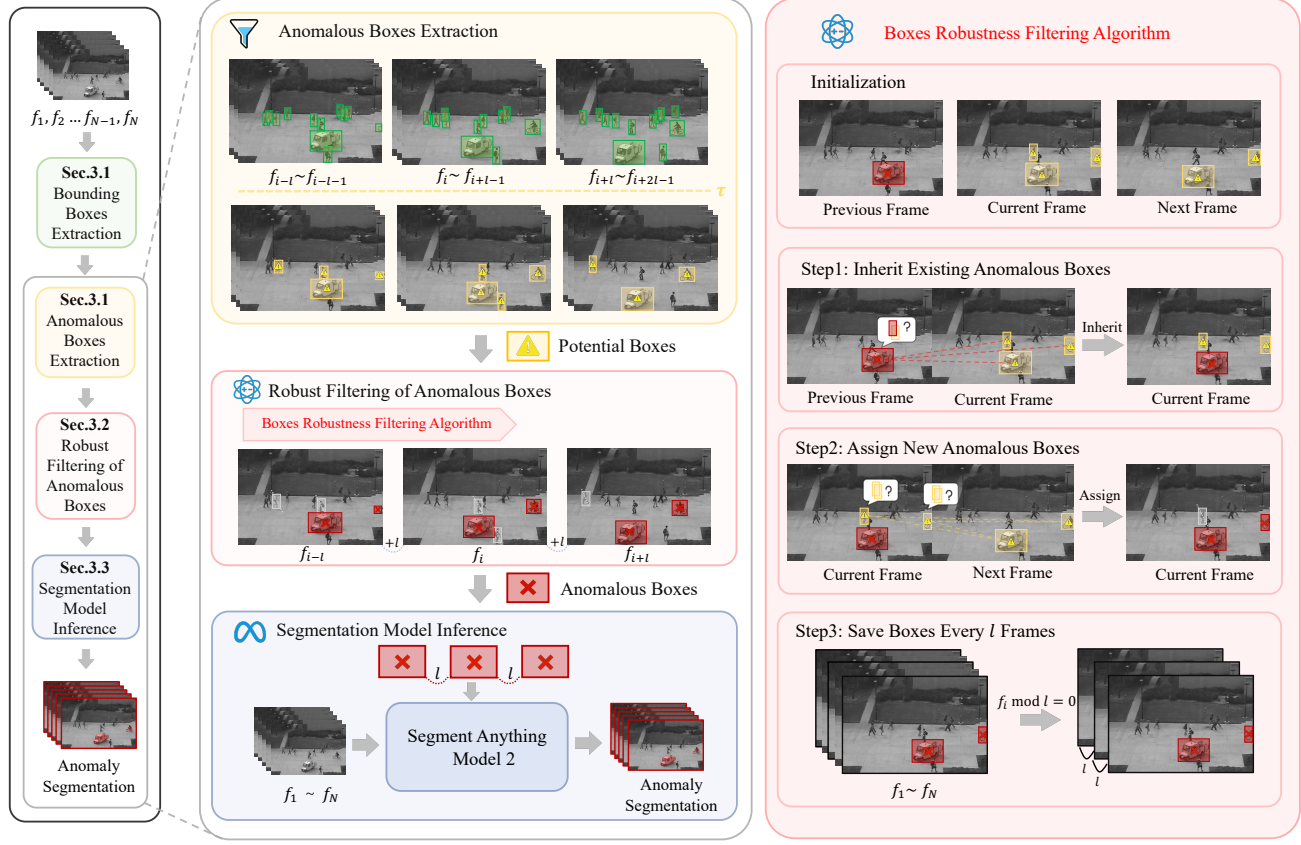
Figure 2. **Pipeline of our proposed *TAO*.** We first generate bounding boxes to identify objects in each frame. Next, we score these boxes using object-centric video anomaly detection algorithms to extract potential anomalous boxes. To ensure robustness, we apply filtering to eliminate redundant boxes. Finally, the filtered boxes and original frames are input into a prompt-based segmentation model to produce pixel-level anomaly segmentation masks.

as image segmentation. By employing semantic prompts (e.g., free-form text) and spatial prompts (e.g., points or bounding boxes), these models achieve highly accurate segmentation guided by input cues. Segmentation Foundation Models (SFMs), such as the Segment Anything Model (SAM) [21] and SEEM [64], have demonstrated exceptional performance, particularly in zero-shot generalization and multi-modal prompting, allowing for cross-task segmentation. Building on these advancements, the Segment Anything Model 2 (SAM2) [42] further improves segmentation, particularly for video-based applications. SAM2 integrates prompting techniques to not only segment static objects with precision but also capture dynamic, complex objects in video contexts. This capability enables SAM2 to deliver finer segmentation and enhanced localization accuracy in video anomaly detection, addressing the challenges posed by real-world variability and complexity. This work leverages SAM2's advanced video segmentation capabilities to tackle the challenge of fine-grained video anomaly detection, achieving precise and robust anomaly identification in

complex scenarios.

3. Method

Preliminary. In the proposed video anomaly detection (VAD) benchmark, a video clip is represented as a sequence $\{f_1, f_2, \ldots, f_{N_f}\} \subseteq \mathcal{C}$, where $\mathcal{C}$ is the set of all possible video clips, and N_f denotes the total number of frames. Each frame f_i is expressed as $f_i = [p_{i,1}, p_{i,2}, \ldots, p_{i,N_i}]$, where N_i is the total number of pixels in the frame, and $p_{i,j} \in \mathcal{P}$ represents the j-th pixel, with $\mathcal{P}$ as the set of possible pixel values. Each pixel $p_{i,j}$ is classified as either "normal" or "anomalous".

Overview. To achieve precise anomaly localization in videos, we design a streamlined pipeline for pixel-level segmentation in video anomaly detection. Our pipeline comprises four stages: bounding boxes extraction, anomalous boxes extraction, robust filtering, and segmentation inference. The process begins by generating bounding boxes to identify objects of interest. These boxes are then scored us-
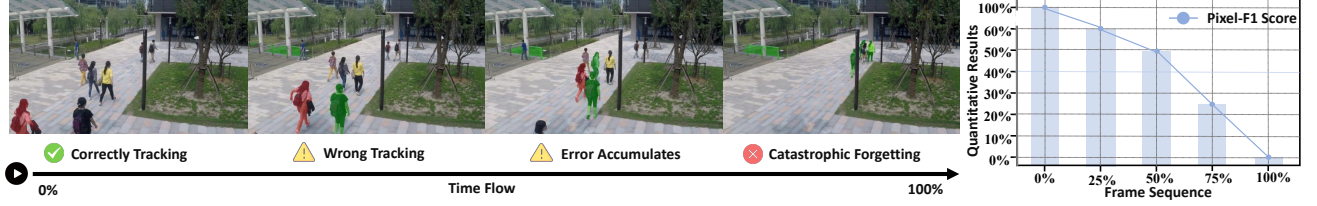
Figure 3. **Impact of Redundant Segmentation on Performance**. Potential anomalous boxes are used as prompts for SAM2 segmentation (red masks: true anomalies, green masks: redundant results). Initially, SAM2 tracks true anomalies accurately, but tracking errors accumulate, leading to catastrophic forgetting and a collapse in Pixel-F1 performance.

Algorithm 1 Boxes Robustness Filtering

1: **Input:** A set of anomalous bounding boxes $\mathcal{B}_{\text{anomaly}} = \{b_i \mid s_i > \tau\}$ for each frame $f_i \in \{f_1, f_2, \ldots, f_n\}$, where s_i is the anomaly score and τ is the threshold; parameters k (tracking window size), h (overlap threshold), l (save interval), and m (frame match threshold)

2: **Output:** Filtered bounding boxes $\mathcal{B}_{\text{filtered}}$ with their corresponding object labels $\mathcal{L}_j$

3: Initialize the list $\mathcal{B}_i$ to store the filtered bounding boxes from frame f_i

4: **for** each frame $f_i \in \{f_1, \ldots, f_n\}$ **do**

5:     **Step 1: Inherit Existing Anomalous Boxes**

6:     **for** each bounding box $b_j \in \mathcal{B}_{\text{anomaly}}$ and b_p from frame f_{i-k} to f_{i-1} **do**

7:         **if** $\sum_{p=i-k}^{i-1} \mathbb{I}(\text{IoU}(b_j, b_p) > h) \geq m$ **then**

8:             Assign $\mathcal{L}_j = \mathcal{L}_p$

9:             Remove b_j from $\mathcal{B}_i$

10:         **end if**

11:     **end for**

12:     **Step 2: Assign New Anomalous Boxes**

13:     **for** each remaining $b_j \in \mathcal{B}_{\text{anomaly}}$ and b_p from frame f_{i+1} to f_{i+k} **do**

14:         **if** $\sum_{p=i+1}^{i+k} \mathbb{I}(\text{IoU}(b_j, b_p) > h) \geq m$ **then**

15:             Assign a new object label $\mathcal{L}_j = \mathcal{F}(b_j)$

16:             Save the tuple $\mathcal{T}_{\text{box}} = (f_i, b_j, \mathcal{L}_j)$

17:         **end if**

18:     **end for**

19:     **Step 3: Save Boxes Every l Frames**

20:     **if** $i \bmod l == 0$ **then**

21:         Save all bounding box tuples $\mathcal{T}_{\text{box}} = (f_i, b_j, \mathcal{L}_j)$ in the current frame

22:     **end if**

23: **end for**

---

ing object-centric anomaly detection algorithms to extract potential anomalies. A tracking-based filtering step follows to eliminate redundant boxes and retain only robust ones. Finally, the filtered boxes and original frames are fed into

a prompt-based segmentation model to generate pixel-level anomaly masks. The pipeline is illustrated in Fig. 2.

3.1. Anomalous Boxes Extraction

Bounding Boxes Extraction. To achieve precise pixel-level segmentation in video anomaly detection, pre-trained object detection algorithms are used to generate bounding boxes for objects within a frame, serving as potential prompts for guiding the segmentation model. For a given input frame f, the object detection model $\mathcal{D}$ outputs m bounding boxes $b_1, b_2, \ldots, b_m$, each paired with a class label $y_1, y_2, \ldots, y_m$:

$$\{(b_1, y_1), (b_2, y_2), \ldots, (b_m, y_m)\} = \mathcal{D}(f) \qquad (1)$$

After the extraction of object bounding boxes in each frame, object-centric VAD algorithms compute an anomaly score s_i for each object within its bounding box b_i. This score is derived from features such as pose, speed, and depth, and can be formally represented as:

$$s_i = \mathcal{A}(b_i) \qquad (2)$$

where $\mathcal{A}$ represents the scoring function used in object-centric video anomaly detection algorithms. To isolate anomalous objects, we filter the bounding boxes based on an anomaly score threshold. Specifically, we retain only the bounding boxes where the anomaly score exceeds a predefined threshold τ, as shown in the following expression:

$$\mathcal{B}_{\text{anomaly}} = \{b_i \mid s_i > \tau\} \qquad (3)$$

where τ denotes the anomaly score threshold, s_i is the anomaly score for the i-th object, and b_i is the corresponding bounding box.

3.2. Robust Filtering of Anomalous Boxes

The anomalous boxes detection from Sec. 3.1 contains numerous false positives due to overlapping characteristics between normal and abnormal boxes. This overlap complicates threshold setting and leads to misclassifications. These redundant boxes can significantly degrade segmentation performance by generating inaccurate results and

overwhelming the model's tracking capabilities, potentially causing the loss of anomalous targets in subsequent frames.

As illustrated in Fig. 3, using potential anomalous boxes as SAM2 prompts at fixed intervals reveals the impact of tracking errors. While red masks indicate true anomalies and green masks show redundant results, the model initially tracks anomalous objects effectively. However, tracking errors accumulate over time, leading to an increasing number of incorrectly tracked objects. This deterioration ultimately results in catastrophic forgetting of true anomalies, reflected in steadily declining Pixel-F1 scores.

To enhance robustness and address redundancy, we present the *Boxes Robustness Filtering* algorithm, which exploits the temporal consistency of true anomalous boxes. Unlike redundant boxes that appear sporadically, true anomalous boxes maintain consistent tracking of the same target. The algorithm, detailed in Alg. 1, operates in three key steps. First, the *inheritance step* computes the similarity between each bounding box b_j in the current frame f_i and bounding boxes b_p from the previous k frames. If the Intersection over Union (IoU) exceeds a threshold h for at least m frames, b_j inherits the label $\mathcal{L}_p$ from earlier frames, ensuring consistent tracking of previously identified anomalies. Second, the *assignment step* addresses bounding boxes that fail to inherit a label by comparing them with bounding boxes in the subsequent k frames. If the IoU exceeds h for at least m frames, a new label $\mathcal{L}_j = \mathcal{F}(b_j)$ is assigned, effectively labeling newly emerging anomalies. Finally, the *saving step* records bounding boxes and their labels every l-th frame, systematically capturing all identified anomalies. By iteratively applying these steps, the *Boxes Robustness Filtering* algorithm achieves robust tracking and localization of anomalous objects. Each bounding box and its label are stored as a tuple $\mathcal{T}_{\text{box}} = (f_i, b_j, \mathcal{L}_j)$, ensuring spatial and temporal consistency across video frames. This approach reduces redundancy and strengthens anomaly detection accuracy, providing a reliable basis for subsequent segmentation and analysis.

3.3. Inference of the Segmentation Model

Using the tuple $\mathcal{T}_{\text{box}} = (f_i, b_j, \mathcal{L}_j)$ obtained from Sec. 3.2, we extract the center point of each bounding box b_j. Specifically, given the coordinates of the box $b_j = [x_1, y_1, x_2, y_2]$, the center point is calculated: $\mathbf{c}_j = \left(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}\right)$. Effectively, these prompts are not saved for every frame but are instead collected at regular intervals, specifically every l-th frame, as outlined in Sec. 3.2. This ensures efficient storage and processing while still capturing the necessary information for accurate segmentation [23, 25]. For each selected frame, both the center point $\mathbf{c}_j$ and the bounding box b_j are stored as prompts. Rather than processing each frame individually, we aggregate prompts from all saved frames $\{f_1, f_{1+l}, f_{1+2l}, \ldots, f_{N_f}\}$, where l denotes the saving in-terval. The aggregated prompts are:

$$\mathcal{I} = \{(\mathbf{c}_j, b_j, f_i) \mid i = 1, 1+l, 1+2l, \ldots, N_f\} \quad (4)$$

where $\mathcal{I}$ represents the set of prompts that includes the center points, bounding box coordinates b_j, and original frames f_i for the selected frames.

By inputting this prompt information $\mathcal{I}$ saved at regular intervals into SAM2, which serves as the prompt-based segmentation model $\mathcal{M}$, we leverage its feature propagation capabilities to efficiently generate pixel-level segmentation results for the entire video. SAM2 propagates the segmentation results from the prompted frames to the remaining frames, producing pixel-level segmentation across the entire video:

$$\mathcal{S} = \mathcal{M}(\mathcal{I}) = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{N_f}\} \quad (5)$$

where $\mathcal{S}_i = [p_1, p_2, \ldots, p_M]$ denotes the segmentation result for frame f_i, where each pixel p_j is either "normal" or "anomalous". This approach enables the generation of the complete segmentation set $\mathcal{S}$, efficiently utilizing contextual information from the prompts.

4. Experiments

4.1. A New Comprehensive VAD Benchmark

Benchmark Design. We present a novel benchmark framework that integrates both pixel-level and object-level evaluation metrics to advance VAD model assessment. This dual-level approach addresses the limitations of using single evaluation metrics. *Pixel-level metrics* excel at identifying subtle irregularities through independent pixel evaluation, but struggle to capture broader spatial patterns. *Object-level metrics* complement this by assessing global characteristics like shape and position, enabling coherent anomaly tracking and recognition. By combining these complementary perspectives, our framework provides a comprehensive evaluation of both fine-grained details and holistic anomaly patterns, particularly valuable for real-world applications.

Benchmark Datasets. *UCSD Ped2* [47] contains 16 training and 12 test videos (240×360 pixels) captured by a fixed overhead camera. The training set consists of normal pedestrian activities, while the test set includes anomalous events such as bikers, skateboarders, and vehicles. *ShanghaiTech Campus* [30], one of the largest VAD datasets, comprises 330 training and 107 test videos (480×856 pixels). While the training set features normal scenarios, the test set contains various anomalies including robbery, fighting, and unauthorized cycling in pedestrian zones.

Benchmark Metrics. For *pixel-level* evaluation, we employ four metrics: Pixel-AUROC measures discrimination between normal and anomalous pixels across thresholds, Pixel-AP assesses detection precision by balancing false

Figure 4. Qualitative comparison with anomaly detection models adapted for video anomaly detection on the UCSD Ped2 dataset.
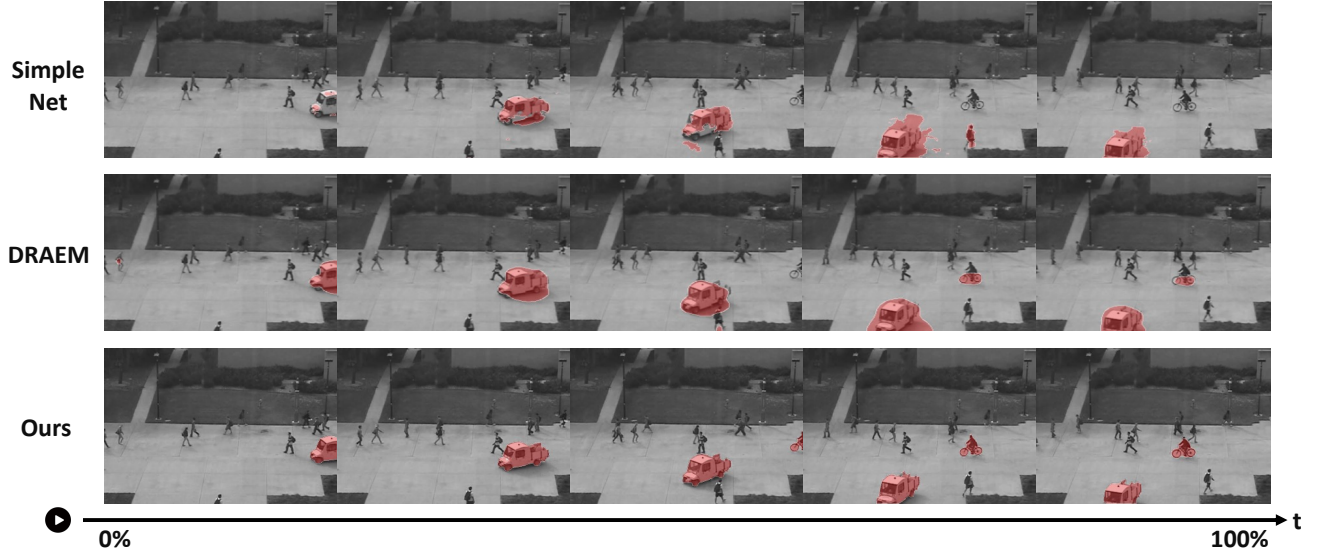

Table 1. Quantitative comparison on video anomaly detection with pixel-level and object-level metrics on the UCSD Ped2 dataset.

Method	Pixel-level				Object-level	
	Pixel-AUROC ↑	Pixel-AP ↑	Pixel-AUPRO ↑	Pixel-F1 ↑	RBDC ↑	TBDC ↑
AdaCLIP (Zero-shot) [5]	51.02	1.32	33.98	2.61	5.8	10.6
AdaCLIP (Fully fine-tuned) [5]	53.06	4.97	50.66	11.19	12.3	15.5
AnomalyCLIP (Zero-shot) [63]	51.63	21.20	36.34	5.92	7.5	11.2
AnomalyCLIP (Fully fine-tuned) [63]	54.25	23.73	38.59	7.48	13.1	21.0
DDAD (Fully trained) [37]	55.87	5.61	15.12	2.67	18.01	13.29
SimpleNet (Fully trained) [33]	52.49	20.51	44.05	10.71	51.18	27.75
DRAEM (Fully trained) [60]	69.58	30.63	35.78	10.89	44.26	70.64
TAO (Partially fine-tuned)	75.11	50.78	72.97	64.12	83.6	93.2

Table 2. Quantitative comparison on video anomaly detection with RBDC and TBDC metrics on the UCSD Ped2 dataset.

Method	RBDC ↑	TBDC ↑
OCAD [19]	52.7	72.8
SS-VAD [40]	62.5	80.5
SiVL [41]	74	89.3
AED-SSMTL [10]	72.8	91.2
TAO (Ours)	83.6	93.2

Table 3. Quantitative comparison on video anomaly detection with RBDC and TBDC metrics on the ShanghaiTech Campus dataset.

Method	RBDC ↑	TBDC ↑
OCAD [19]	20.7	44.5
BAF-AT [9]	41.3	78.8
AED-SSMTL [10]	43.2	84.1
HF2VAD [31]	45.4	84.5
STPT [38]	51.6	84.6
TAO (Ours)	62.1	85.4

positives and negatives, Pixel-AUPRO quantifies segmentation accuracy through region overlap, and Pixel-F1 provides an integrated precision-recall measure. For *object-level* assessment, we utilize Region-Based Detection Criterion (RBDC) and Track-Based Detection Criterion (TBDC). RBDC measures spatial accuracy using Intersection over Union (IoU) with threshold α, while TBDC evaluates temporal consistency in tracking anomalous regions across

frames. Together, these metrics provide comprehensive evaluation of both spatial precision and temporal coherence.

4.2. Implementation Details

Our framework builds upon SAM2 [42], utilizing an object-level anomaly detection algorithm [43] to extract anomalous bounding boxes as segmentation prompts. We em-

ploy SAM2 [42] as the prompt-based segmentation model, which integrates these prompts with original frames for precise anomaly segmentation. Our implementation adopts a partial fine-tuning strategy, where only the object-level VAD algorithm is fine-tuned on target datasets, while SAM2 retains its pre-trained weights for inference. In the bounding box threshold filtering stage, we assign anomaly scores based on pose and depth features, with thresholds set to $\tau = 1.5$ for UCSD Ped2 and $\tau = 1.6$ for ShanghaiTech Campus. The subsequent robustness filtering stage employs a tracking window of $k = 5$, frame match threshold of $m = 3$, and box overlap threshold of $h = 0.2$, with save intervals of $l = 5$ and $l = 15$ for UCSD Ped2 and ShanghaiTech Campus respectively. For segmentation, we utilize SAM2 with hiera base-plus weights on an NVIDIA RTX 4090 GPU. To ensure robust evaluation, overlapping boxes of the same anomaly are merged, and segmentation probability maps are uniformly binarized. For object-level metrics, we derive bounding boxes from segmentation extremal coordinates. Given anomaly pixels $P = (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, we compute the box coordinates as: $x_{\min} = \min_{(x_i, y_i) \in P} x_i$, $x_{\max} = \max_{(x_i, y_i) \in P} x_i, y_{\min} = \min_{(x_i, y_i) \in P} y_i$, $y_{\max} = \max_{(x_i, y_i) \in P} y_i$, where the resulting bounding box $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ is used for computing RBDC and TBDC metrics, providing robust spatial evaluation at the object level.

4.3. Comparison to Anomaly Detection Models

Given the lack of pixel-level segmentation capabilities in current video anomaly detection models, we benchmark several state-of-the-art image anomaly detection algorithms, including AdaCLIP [5], AnomalyCLIP [63], DRAEM [60], DDAD [37], and SimpleNet [33], to evaluate both pixel-level and object-level metrics on the UCSD Ped2 dataset. AdaCLIP and AnomalyCLIP were selected for their pre-trained vision-language architectures, which closely align with SAM2's design and demonstrate strong zero-shot capabilities. To ensure a fair comparison, we assess both AdaCLIP and AnomalyCLIP under zero-shot and fully fine-tuned settings. In contrast, DRAEM, DDAD, and SimpleNet, as fully trained models, provide robust baselines for pixel-level and object-level anomaly detection, representing diverse methodological approaches. For anomaly detection, video streams are processed frame-by-frame through each algorithm, generating heatmaps that identify potential anomalies. Pixel-level segmentation is achieved by applying a threshold to these heatmaps, allowing precise localization of anomalous regions.

As shown in Tab.1, the experimental results clearly demonstrate the advantages of our model in both pixel-level and object-level anomaly detection. While most benchmarked models struggle to balance these two aspects, of-

ten excelling in one while underperforming in the other, our model achieves robust performance across both dimensions. This balance reflects its strong and versatile design, enabling precise pixel-level segmentation and maintaining spatial-temporal consistency at the object level. Moreover, our partial fine-tuning approach effectively leverages pre-trained features, minimizing the need for extensive dataset-specific training. In contrast, fully trained models such as DRAEM, DDAD, and SimpleNet, as well as vision-language models like AdaCLIP and AnomalyCLIP, fail to achieve comparable results. Fig. 4 further illustrates the strengths of our method. The segmentation results produced by our model are more complete and detailed, accurately delineating object boundaries while preserving the integrity of anomalous objects.

4.4. Comparsion to Conventional Video Anomaly Detection Models

To assess the performance of our proposed method, we conducted comprehensive experiments on two widely used datasets, UCSD Ped2 and ShanghaiTech Campus. Our approach was benchmarked against several state-of-the-art video anomaly detection models, including SS-VAD [40], SiVL [41], AED-SSMTL [10], HF2VAD [31], BAF-AT [9], and STPT [38]. Due to the absence of fine-grained segmentation capabilities in traditional methods, the evaluation was focused on object-centric anomaly detection and limited to object-level metrics, TBDC [40] and RBDC [40]. These metrics provide a robust and meaningful basis for comparing the effectiveness of our model with existing approaches.

Our method achieves state-of-the-art performance on both the UCSD Ped2 and ShanghaiTech Campus datasets. On UCSD Ped2, it outperforms existing methods in TBDC and RBDC metrics, demonstrating precise tracking and localization of anomalous objects across video frames. On the more challenging ShanghaiTech Campus dataset, characterized by complex backgrounds and subtle anomalies, our method consistently surpasses baselines by delivering balanced performance across all evaluation metrics. Unlike existing approaches, which often excel in isolated metrics but fail to generalize, our model effectively integrates object-centric detection and fine-grained segmentation, enabling robust anomaly localization and tracking under diverse real-world conditions, as shown in Tabs. 2 and 3.

To conclude, our method establishes a new standard in video anomaly detection by combining object-centric detection with fine-grained segmentation. This integration enables precise localization and robust tracking of anomalous objects, addressing the limitations of traditional methods that struggle with balanced performance across metrics. Our approach demonstrates versatility and effectiveness across diverse datasets, highlighting its potential for real-world applications.

Table 4. Ablation study on proposed key components, evaluated with pixel-level and object-level metrics on the UCSD Ped2 dataset.

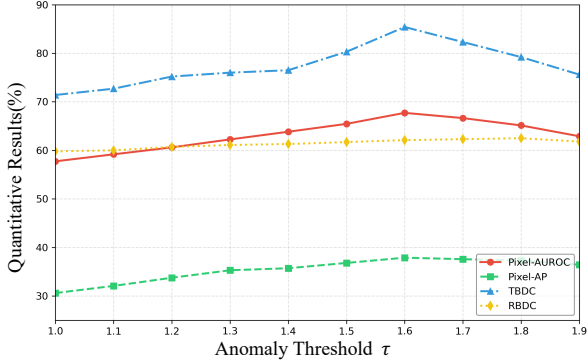| Video Track Mode	Box Robust Filtering Algorithm	Pixel-level				Object-level	
		Pixel-AUROC ↑	Pixel-AP ↑	Pixel-AUPRO ↑	Pixel-F1 ↑	RBDC ↑	TBDC ↑
×	×	41.72	25.30	40.55	32.99	41.0	71.6
×	✓	39.06	22.80	38.09	30.56	39.3	71.7
✓	×	68.90	38.47	67.23	52.17	70.8	76.9
✓	✓	75.11	50.78	72.97	64.12	83.6	93.2

Figure 5. **Sensitivity Analysis of Anomaly Threshold (τ).** We evaluate performance under varying threshold on ShanghaiTech Campus dataset.
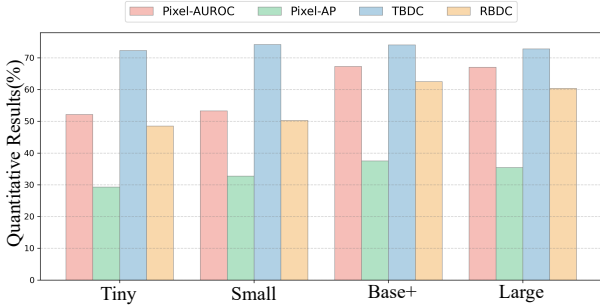

Figure 6. **Robust Testing of SAM2 Backbones.** We compare the performance across different SAM2 model variants.

4.5. Further Empirical Study

Ablation Study. We conduct ablation experiments on the UCSD Ped2 dataset to evaluate two components: *video tracking mode* and *box robust filtering*. Video tracking mode, a SAM2 component, enables temporal analysis, while disabling it reduces the model to SAM, performing frame-by-frame segmentation. Disabling box robust filtering uses raw anomaly-thresholded boxes as prompts. Tab. 4 shows the effects of each component on pixel- and object-level metrics. The baseline model, with both components disabled, performs the worst. Enabling box robust filter-

ing alone yields minimal improvement, while activating video tracking mode significantly boosts pixel-level accuracy. The best performance is achieved with both components enabled, demonstrating their synergistic effect. These results confirm the necessity of both temporal tracking and robust filtering in our framework.

Sensitivity Analysis. We conduct a sensitivity analysis to assess how variations in the anomaly threshold τ affect our framework's performance. Fig. 5 shows the impact of τ on four metrics—Pixel-AUROC, Pixel-AP, TBDC, and RBDC—using the ShanghaiTech Campus dataset. As τ ranges from 1.0 to 1.9, the framework performs stably, with optimal results at $\tau = 1.6$. TBDC is most sensitive, peaking at 85% at $\tau = 1.6$, while RBDC stays around 62%. Pixel-level metrics, Pixel-AUROC and Pixel-AP, also show steady trends, reflecting the framework's balance of fine-grained anomaly detection and object consistency. This analysis highlights the robustness and adaptability of our framework across threshold variations, ensuring reliable detection in diverse scenarios.

Robustness Testing. We evaluate our framework's robustness through extensive testing across different SAM2 backbone architectures (Fig. 6). Our analysis covers models from the lightweight Tiny variant to the full-scale Large model, with Base+ consistently yielding the best results. This stability suggests that the framework's effectiveness is not dependent on backbone capacity. While larger models show slight improvements in pixel-level metrics, performance remains strong across all variants. These results validate the framework's reliability and robustness, demonstrating its suitability for deployment in a wide range of scenarios, irrespective of computational constraints.

5. Conclusion

This paper presents *TAO*, a framework integrating coarse object detection with fine-grained anomaly localization in videos. Leveraging SAM2 and our dual-level evaluation, it enables precise segmentation and robust tracking. Experiments on UCSD Ped2 and ShanghaiTech datasets demonstrate state-of-the-art performance in pixel- and object-level metrics, advancing video anomaly detection.

6. Acknowledgments

References

[1] Sunghyun Ahn, Youngwan Jo, Kijung Lee, and Sanghyun Park. Videopatchcore: An effective method to memorize normality for video anomaly detection. *arXiv preprint arXiv:2409.16225*, 2024. 1, 2

[2] Jinbin Bai, Zhen Dong, Aosong Feng, Xiao Zhang, Tian Ye, Kaicheng Zhou, and Mike Zheng Shou. Integrating view conditions for image synthesis. *arXiv preprint arXiv:2310.16002*, 2023. 1

[3] Jinbin Bai, Wei Chow, Ling Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Shuicheng Yan. Humanedit: A high-quality human-rewarded dataset for instruction-based image editing. *arXiv preprint arXiv:2412.04280*, 2024. 1

[4] Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Qing-Guo Chen, Xiangtai Li, Zhen Dong, Lei Zhu, and Shuicheng Yan. Meissonic: Revitalizing Masked Generative Transformers for Efficient High-Resolution Text-to-Image Synthesis. *arXiv preprint arXiv:2410.08261*, 2024. 2

[5] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pages 55–72. Springer, 2025. 6, 7

[6] Zhiyuan Ding, Qi Dong, Haote Xu, Chenxin Li, Xinghao Ding, and Yue Huang. Unsupervised anomaly segmentation for brain lesions using dual semantic-manifold reconstruction. In *International Conference on Neural Information Processing*, pages 133–144. Springer International Publishing Cham, 2022. 2

[7] Keval Doshi and Yasin Yilmaz. Any-shot sequential anomaly detection in surveillance videos. 2020. 2

[8] Keval Doshi and Yasin Yilmaz. Continual learning for anomaly detection in surveillance videos. 2020. 1

[9] Mariana-Iuliana Georgescu, RaduTudor 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,IEEE Transactions on Pattern Analysis and Machine Intelligence*. 6, 7

[10] 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. 1, 2, 6, 7

[11] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[12] Chunming He, Kai Li, Guoxia Xu, Yulun Zhang, Runze Hu, Zhenhua Guo, and Xiu Li. Degradation-resistant unfolding network for heterogeneous image fusion. In *ICCV*, pages 12611–12621, 2023. 2

[13] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *NeurIPS*, 36, 2024. 2

[14] Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Zhenhua Guo, Xiu Li, Martin Danelljan, and Fisher Yu. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. *ICLR*, 2024. 2

[15] Chunming He, Chengyu Fang, Yulun Zhang, Kai Li, Longxiang Tang, Chenyu You, Fengyang Xiao, Zhenhua Guo, and Xiu Li. Reti-diff: Illumination degradation image restoration with retinex-based latent diffusion model. *ICLR*, 2025. 2

[16] Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey. *TPAMI*, 2025. 2

[17] Chunming He, Rihan Zhang, Fengyang Xiao, Chenyu Fang, Longxiang Tang, Yulun Zhang, Linghe Kong, Deng-Ping Fan, Kai Li, and Sina Farsiu. Run: Reversible unfolding network for concealed object segmentation. *arXiv preprint arXiv:2501.18783*, 2025. 2

[18] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[19] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 6

[20] Hamza Karim, Keval Doshi, and Yasin Yilmaz. Real-time weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6848–6856, 2024. 2

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[22] Chenxin Li, Yunlong Zhang, Jiongcheng Li, Yue Huang, and Xinghao Ding. Unsupervised anomaly segmentation using image-semantic cycle translation. *arXiv preprint arXiv:2103.09094*, 2021. 2

[23] Chenxin Li, Yunlong Zhang, Zhehan Liang, Wenao Ma, Yue Huang, and Xinghao Ding. Consistent posterior distributions under vessel-mixing: a regularization for cross-domain retinal artery/vein classification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 61–65. IEEE, 2021. 5

[24] Chenxin Li, Wenao Ma, Liyan Sun, Xinghao Ding, Yue Huang, Guisheng Wang, and Yizhou Yu. Hierarchical deep network with uncertainty-aware semi-supervised learning for vessel segmentation. *Neural Computing and Applications*, pages 1–14, 2022. 2

[25] Chenxin Li, Xinyu Liu, Wuyang Li, Cheng Wang, Hengyu Liu, Yifan Liu, Zhen Chen, and Yixuan Yuan. U-kan makes strong backbone for medical image segmentation and generation. *arXiv preprint arXiv:2406.02918*, 2024. 5

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[27] Huangxing Lin, Yunlong Lin, Jingyuan Xia, Linyu Fan, Feifei Li, Yingying Wang, and Xinghao Ding. Fusion2void: Unsupervised multi-focus image fusion based on image inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[28] Yunlong Lin, Tian Ye, Sixiang Chen, Zhenqi Fu, Yingying Wang, Wenhao Chai, Zhaohu Xing, Lei Zhu, and Xinghao Ding. Aglldiff: Guiding diffusion models towards unsupervised training-free real-world low-light image enhancement. *arXiv preprint arXiv:2407.14900*, 2024. 2

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[30] 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. 5

[31] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13588–13597, 2021. 6, 7

[32] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[33] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, 2023. 6, 7

[34] Shao-Yuan Lo, Poojan Oza, and Vishal M. Patel. Adversarially robust one-class novelty detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–12, 2022. 1, 2

[35] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 2

[36] Jakub Micorek, Horst Possegger, Dominik Narnhofer, Horst Bischof, and Mateusz Koziński. MULDE: Multiscale Log-Density Estimation via Denoising Score Matching for Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18868–18877, 2024. 1

[37] Arian Mousakhan, Thomas Brox, and Jawad Tayyub. Anomaly detection with conditioned denoising diffusion models. *arXiv preprint arXiv:2305.15956*, 2023. 6, 7

[38] Yassine Naji, Aleksandr Setkov, Angelique Loesch, Michèle Gouiffès, and Romaric Audigier. Spatio-temporal predictive tasks for abnormal event detection in videos. 2022. 1, 6, 7

[39] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[40] Bharathkumar Ramachandra and Michael 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. 1, 6, 7

[41] Bharathkumar Ramachandra, Michael J. Jones, and Ranga Raju Vatsavai. Learning a distance function with a siamese network to localize anomalies in videos. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 1, 2, 6, 7

[42] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 3, 6, 7

[43] Tal Reiss and Yedid Hoshen. Attribute-based representations for accurate and interpretable video anomaly detection. 2022. 1, 2, 6

[44] Chenrui Shi, Che Sun, Yuwei Wu, and Yunde Jia. Video anomaly detection via sequentially learning multiple pretext tasks. 1, 2

[45] Liyan Sun, Chenxin Li, Xinghao Ding, Yue Huang, Zhong Chen, Guisheng Wang, Yizhou Yu, and John Paisley. Few-shot medical image segmentation using a global correlation network with discriminative embedding. *Computers in biology and medicine*, 140:105067, 2022. 2

[46] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4975–4986, 2021. 2

[47] Shu Wang and Zhenjiang Miao. Anomaly detection in crowd scene. In *IEEE 10th International Conference on Signal Processing Proceedings*, pages 1220–1223. IEEE, 2010. 5

[48] Yingying Wang, Yunlong Lin, Ge Meng, Zhenqi Fu, Yuhang Dong, Linyu Fan, Hedeng Yu, Xinghao Ding, and Yue Huang. Learning high-frequency feature enhancement and alignment for pan-sharpening. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 358–367, 2023. 1

[49] Yingying Wang, Yunlong Lin, Xuanhua He, Hui Zheng, Keyu Yan, Linyu Fan, Yue Huang, and Xinghao Ding. Learning diffusion high-quality priors for pan-sharpening: A two-stage approach with time-aware adapter fine-tuning. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 1

[50] 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. 2

[51] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6074–6082, 2024. 2

[52] Fengyang Xiao, Sujie Hu, Yuqi Shen, Chengyu Fang, Jinfa Huang, Chunming He, Longxiang Tang, Ziyun Yang, and Xiu Li. A survey of camouflaged object detection and beyond. *CAAI AIR*, 2024. 2

[53] Zhaohu Xing, Lequan Yu, Liang Wan, Tong Han, and Lei Zhu. Nestedformer: Nested modality-aware transformer for brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 140–150. Springer, 2022. 2

[54] Zhaohu Xing, Liang Wan, Huazhu Fu, Guang Yang, and Lei Zhu. Diff-unet: A diffusion embedded network for volumetric segmentation. *arXiv preprint arXiv:2303.10326*, 2023. 2

[55] Zhaohu Xing, Sicheng Yang, Sixiang Chen, Tian Ye, Yijun Yang, Jing Qin, and Lei Zhu. Cross-conditioned diffusion model for medical image to image translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 201–211. Springer, 2024. 2

[56] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 578–588. Springer, 2024.

[57] Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: Reasoning for video anomaly detection with large language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1, 2

[58] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 2

[59] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536, 2024. 2

[60] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draema discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8330–8339, 2021. 6, 7

[61] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2

[62] Yunlong Zhang, Chenxin Li, Xin Lin, Liyan Sun, Yihong Zhuang, Yue Huang, Xinghao Ding, Xiaoqing Liu, and Yizhou Yu. Generator versus segmentor: Pseudo-healthy synthesis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*, pages 150–160. Springer International Publishing, 2021. 2

[63] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023. 6, 7

[64] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 3