Track Any Anomalous Object: A Granular Video Anomaly Detection Pipeline

Supplementary Material

6. Supplementary

We propose a straightforward and effective pipeline for fine-grained video anomaly detection. The core approach consists of two key steps: **1)** generating anomaly prompts based on object detection results from video sequences and refining these prompts using a robust filtering algorithm; **2)** applying a prompt-based segmentation model to produce accurate pixel-level anomaly masks. This pipeline enables efficient identification and segmentation of anomalous objects, significantly improving detection precision and temporal consistency. In the supplementary materials, we provide additional details on the following aspects:

- We offer a more comprehensive explanation of the implementation process for anomalous boxes extraction and SAM2 segmentation inference in Sec. 6.1.
- We provide a detailed explanation of the object-level evaluation metrics utilized in our experiments, emphasizing their importance in assessing spatial and temporal anomaly detection performance in Sec. 6.2.
- We analyze the limitations of our proposed model and existing baselines, highlighting potential avenues for improvement in Sec. 6.3.
- we present additional experimental visualization results, highlighting the instance segmentation performance of our model, including examples from the ShanghaiTech Campus dataset in Sec. 6.4.

6.1. Comprehensive Implementation Details

Details of Anomalous Boxes Extraction. Our object-level VAD algorithm [27] utilizes features such as speed, pose, and depth to detect anomalies. During the training phase, a probabilistic density model, such as k-nearest neighbors or Mahalanobis distance, is constructed based on normal behavioral attributes. In the testing phase, the probability density of each object's features is calculated, where lower density values indicate greater deviation from normal behavior, resulting in higher anomaly scores. For the test data, only speed and depth features are used to compute anomaly scores. These scores are then standardized using the corresponding speed and depth anomaly scores from the training data, enhancing the prominence of anomalous objects and making them easier to detect. The standardized speed and depth anomaly scores are subsequently summed to produce a final overall anomaly score. A threshold is applied to this score to effectively filter and identify anomalous objects and their corresponding bounding boxes.

To address overlapping anomaly boxes that may correspond to the same anomalous object, we calculate the Intersection over Union (IoU) for each pair of filtered boxes B_i and B_j within the same frame. The IoU is computed as the ratio of the intersection area to the union area between two boxes, defined as:

$$\text{IoU}(B_i, B_j) = \frac{\text{Area}(B_i \cap B_j)}{\text{Area}(B_i \cup B_j)}.$$

If the IoU exceeds a threshold of $\tau = 0.3$, the two boxes are deemed to represent the same anomalous object. In such cases, a new box B_{new} is created by merging the two, with its top-left corner coordinates given by $\min(x_i^{\text{min}}, x_j^{\text{min}})$ and $\min(y_i^{\text{min}}, y_j^{\text{min}})$, and its bottom-right corner coordinates determined as $\max(x_i^{\text{max}}, x_j^{\text{max}})$ and $\max(y_i^{\text{max}}, y_j^{\text{max}})$. This merging process consolidates overlapping boxes into a single bounding box that accurately represents the anomalous object. By iteratively applying this procedure across all frames, the algorithm ensures a non-redundant and consistent representation of anomalies, improving the overall accuracy and reliability of localization.

Details of Segmentation Model Inference. We utilize SAM2 as the prompt-based segmentation model to perform instance segmentation for distinct anomalous objects in video clips. This process involves generating prompts for each object using a robust bounding box filtering algorithm, applied at fixed frame intervals. Each prompt is stored as a tuple $\mathcal{T}_{\text{box}} = (f_i, b_j, \mathcal{L}_j)$, where f_i represents the frame, b_j is the bounding box, and $\mathcal{L}_j$ is the corresponding object label. These tuples ensure consistent and accurate tracking of anomalous objects across frames, maintaining both spatial and temporal coherence. By consolidating bounding boxes and their associated labels into structured prompts, the model effectively localizes and segments anomalies within dynamic video contexts. Instance segmentation in SAM2 is performed by providing the corresponding object labels and bounding box prompts as inputs. The resulting segmentation outputs are processed to serve different evaluation purposes. For pixel-level metrics, the instance segmentation results are transformed into binarized segmentation masks that highlight anomalous regions at the pixel level. For object-level metrics, each instance segmentation result is converted into a bounding box that encapsulates the segmented object. This dual-processing approach allows for comprehensive evaluation of both fine-grained anomaly localization and high-level object tracking, ensuring the robustness and accuracy of the proposed method.

6.2. Object-Level Evaluation Metrics

RBDC. The Region-Based Detection Criterion (RBDC) evaluates the spatial accuracy of anomaly detection by quantifying the proportion of correctly matched predicted regions relative to the ground truth regions. For a predicted bounding box B_{pred} and a ground truth box B_{gt}, the Intersection over Union (IoU) is computed as:

$$\text{IoU} = \frac{\text{Area}(B_{\text{pred}} \cap B_{\text{gt}})}{\text{Area}(B_{\text{pred}} \cup B_{\text{gt}})}.$$

A match is deemed correct if IoU $> \alpha$, where α is a predefined threshold (e.g., $\alpha = 0.1$). The RBDC score is defined as:

$$\text{RBDC} = \frac{\text{Number of Correctly Matched Regions}}{\text{Total Number of Ground Truth Regions}}.$$

RBDC is essential for assessing spatial precision, ensuring that detected anomalies accurately align with ground truth regions, particularly in scenarios with complex or overlapping anomalies.

TBDC. The Track-Based Detection Criterion (TBDC) evaluates the temporal consistency of anomaly detection by measuring the proportion of correctly tracked anomaly trajectories relative to the total number of ground truth tracks. A trajectory is considered correctly tracked if the IoU between the predicted bounding box and the ground truth box exceeds the threshold α in each frame of the track. The TBDC score is calculated as:

$$\text{TBDC} = \frac{\text{Number of Correctly Tracked Anomaly Tracks}}{\text{Total Number of Ground Truth Tracks}}.$$

TBDC is critical for evaluating temporal robustness, capturing the model's ability to maintain consistent anomaly detection across consecutive frames, which is particularly important in dynamic video scenarios.

Together, RBDC and TBDC provide a comprehensive framework for evaluating object-level anomaly detection, addressing both spatial precision and temporal coherence. These metrics are particularly well-suited for real-world applications such as surveillance and autonomous systems, where accurate spatial localization and robust temporal tracking are paramount.

6.3. Limitations and Future Directions

Limitations. The performance of our model is closely tied to the robustness of the prompts provided to the prompt-based segmentation model, such as SAM2. These prompts heavily rely on the effectiveness of the object-level anomaly detection algorithm in assigning accurate anomaly scores to objects within anomalous frames. A precise anomaly detection algorithm that effectively distinguishes between normal and anomalous objects generates higher-quality prompts, resulting in improved segmentation accuracy. However, false-positive prompts pose significant challenges. First, they can introduce cumulative tracking errors in SAM2, leading to catastrophic forgetting of actual anomalous objects. As these errors propagate across frames, the model may progressively lose its ability to detect critical anomalies, severely undermining its reliability. Second, false-positive prompts increase the computational burden during SAM2 inference. By prompting the model to process non-anomalous objects, they degrade inference efficiency, resulting in slower processing times and unnecessary computational overhead. Addressing these issues is essential to ensure both the accuracy and efficiency of the proposed framework. This underscores the need to enhance the anomaly detection algorithm's precision and robustness, thereby minimizing the impact of false-positive prompts and maximizing the overall performance of the system.

Future Directions. Existing video anomaly detection datasets primarily focus on frame-level and object-level anomalies, with pixel-level annotations being extremely limited. Among the few datasets that provide pixel-level annotations, these are often coarse, offering only rough outlines of anomalous objects rather than precise contours. Additionally, current pixel-level annotations are predominantly binary masks, which pose significant challenges in scenarios where anomalous objects overlap, as binary masks fail to differentiate between overlapping objects, making accurate evaluation difficult. To address these limitations, we propose adopting instance-level pixel annotations for anomalous objects. Instance-level annotations would uniquely identify each anomalous object at the pixel level, even in complex scenarios involving overlapping objects. This approach would enhance the precision of pixel-level anomaly detection and enable more granular evaluations in video anomaly detection tasks. Furthermore, the adoption of instance-level pixel annotations would support the development of more robust algorithms capable of handling real-world scenarios, where anomalies often appear in complex and overlapping forms. By bridging the gap in current datasets, instance-level annotations could serve as a critical foundation for advancing video anomaly detection and promoting consistent benchmarking across future methods.

6.4. More Visualization about Experiments

We provide additional visualization results from comparative experiments, showcasing detailed visual comparisons between our method and four baselines—SimpleNet [17], DRAEM [36], DDAD [21], and AnomalyCLIP [38]—on three selected video clips from the UCSD Ped2 dataset (see Fig. 7–9). Binary segmentation masks are employed for consistency and clarity, and the results clearly demonstrate that our model significantly outperforms the baseline methods. Additionally, we present instance segmentation results of our model on the ShanghaiTech Campus dataset, further illustrating its effectiveness and robustness in segmenting anomalous objects at the instance level (see Fig. 10).

Figure 7. Visual comparisons on the UCSD Ped2 dataset. Red masks represent the binary segmentation results.

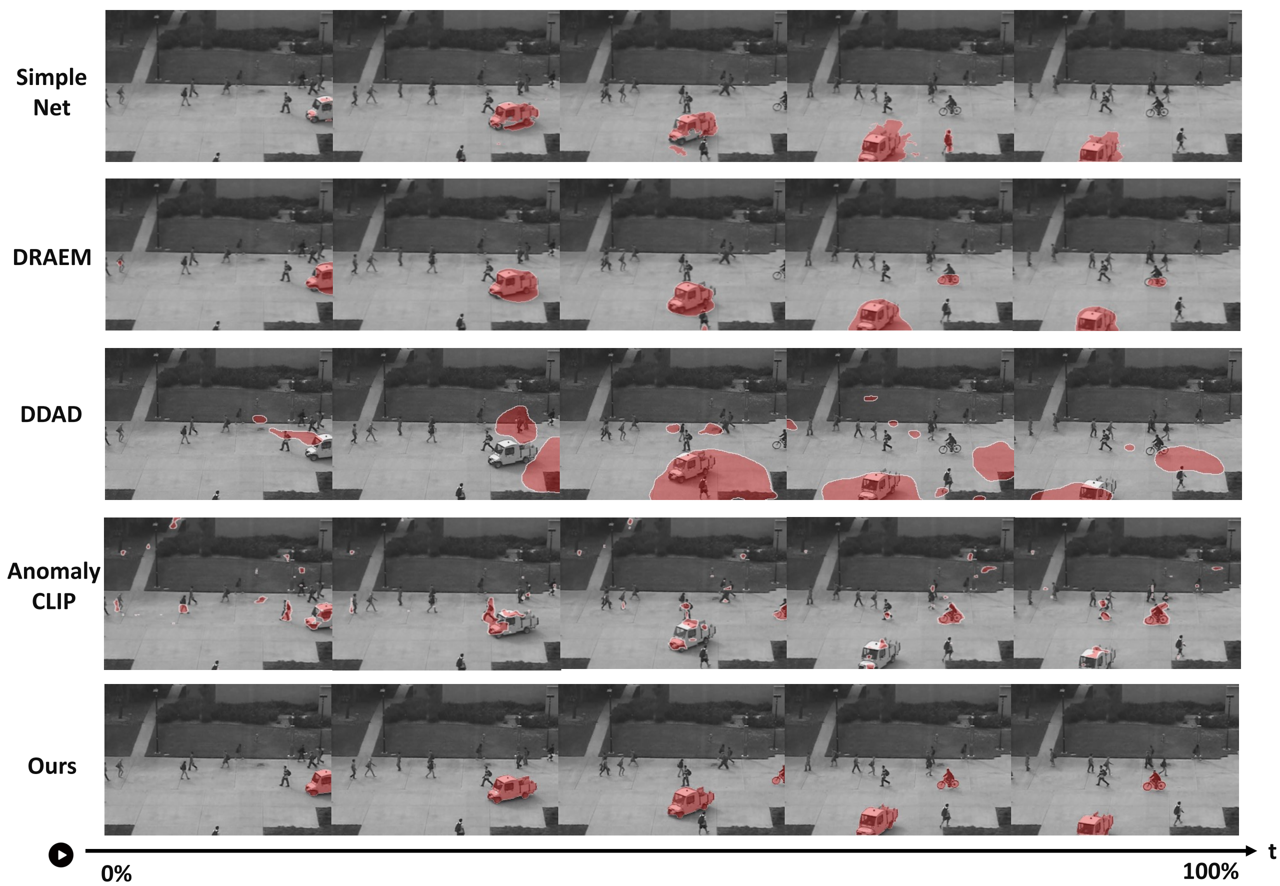

Figure 8. Visual comparisons on the UCSD Ped2 dataset. Red masks represent the binary segmentation results.

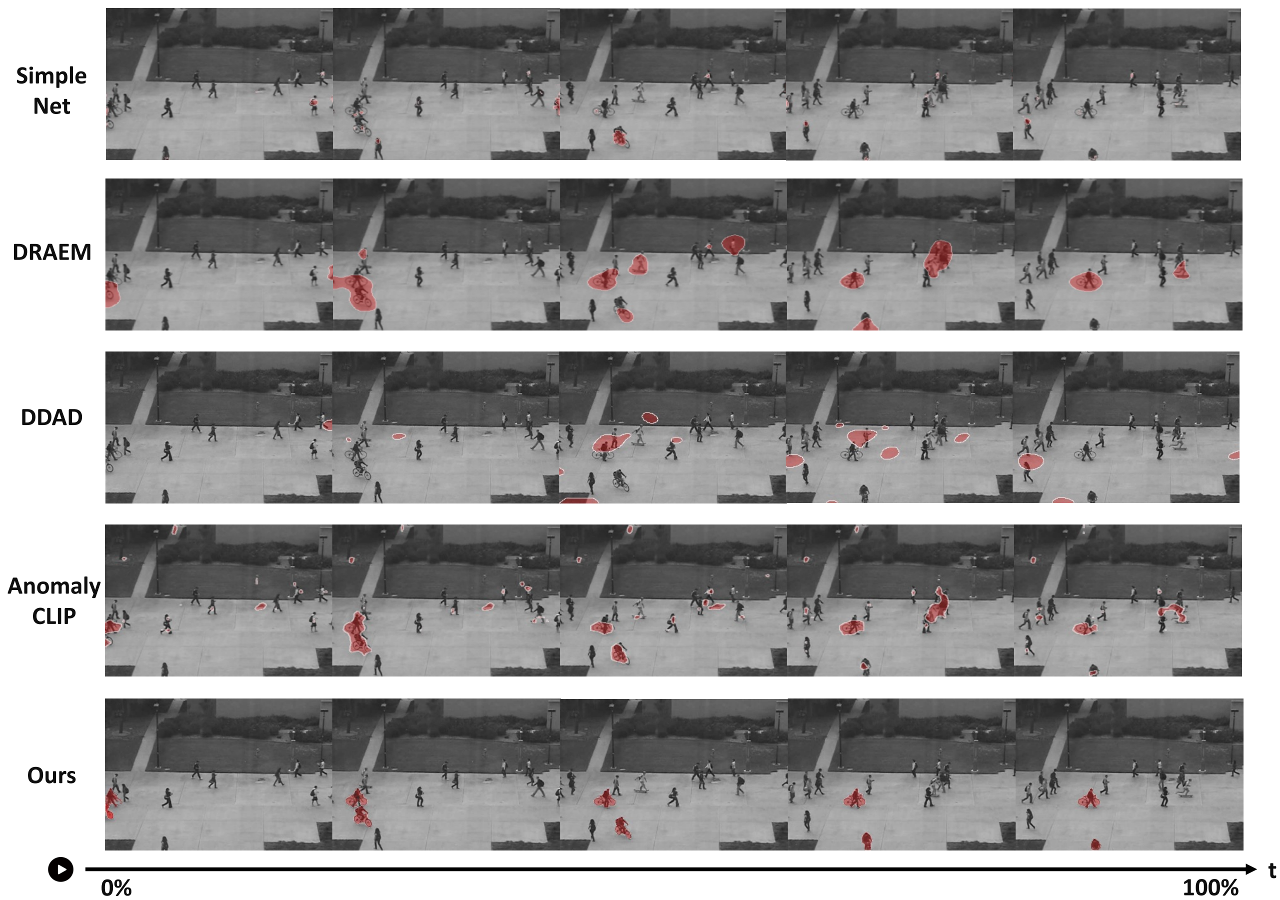

Figure 9. Visual comparisons on the UCSD Ped2 dataset. Red masks represent the binary segmentation results.

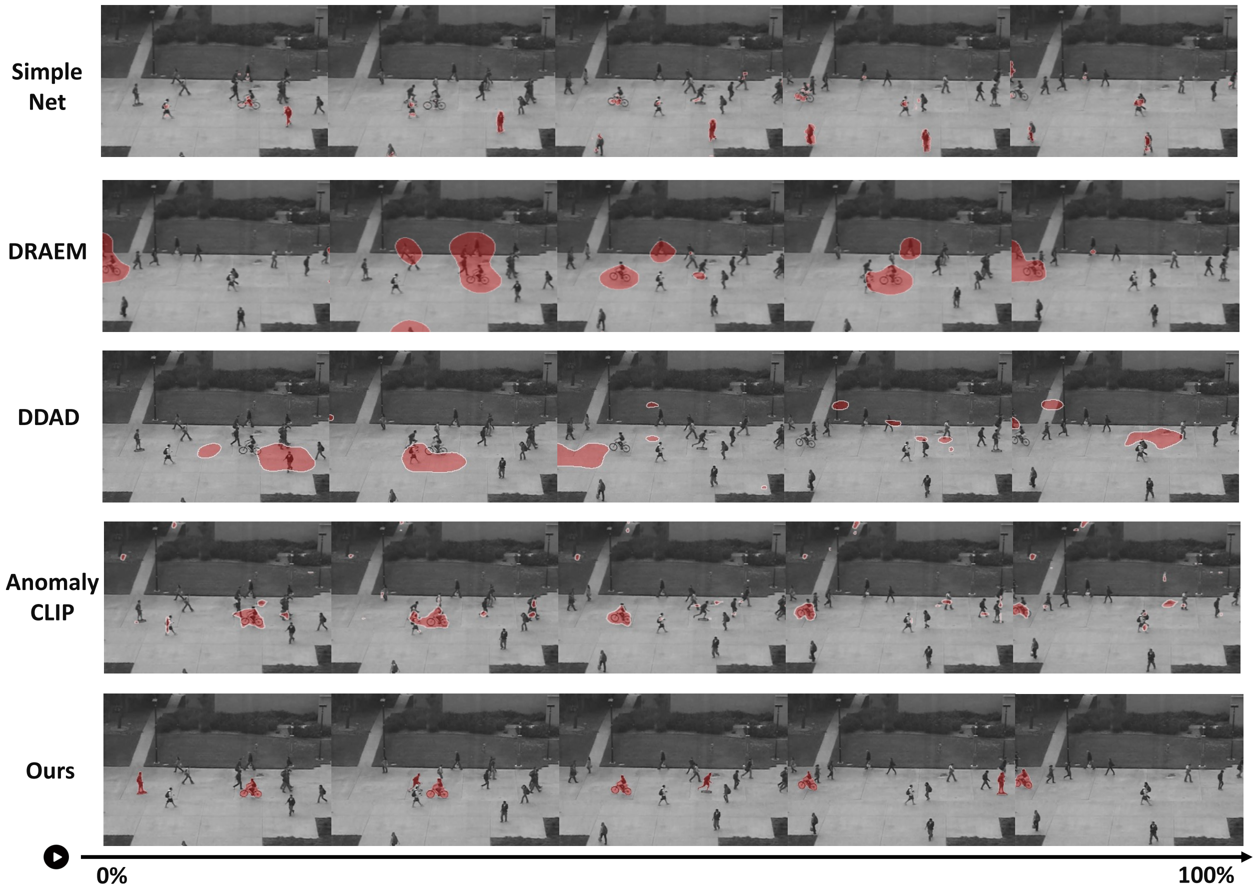

Figure 10. Instance segmentation visualization results on the ShanghaiTech Campus dataset. Masks in different colors represent the segmentation results for distinct instances.