FusedVision: A Knowledge-Infusing Approach for Practical Anomaly Detection in Real-world Surveillance Videos

Khaled Dawoud[1], Muhammad Zaigham Zaheer[1], Mustaqeem Khan[2],
Karthik Nandakumar[4,1], Abdulmotaleb El Saddik[3], Muhammad Haris Khan[1]
[1]Mohammed Bin Zayed University of Artificial Intelligence
[2]United Arab Emirates University [3]University of Ottawa [4]Michigan State University

Abstract

Object-centric approaches have gained attention as effective one-class classification methods for detecting anomalies in videos. These approaches rely on using an object detector to isolate all objects in the frames and subsequently leveraging either the objects themselves or their interactions to train a learning system. In this study, we put forth a novel perspective towards anomaly detection by proposing a branched network architecture that employs both an object detector and a normalcy learning model, working together in tandem to more effectively identify anomalies within the data. Through extensive experimentation, we analyze the optimal fusion mechanism as well as anomaly scoring proposed in our branched approach. Our approach is more practical towards real-world applications of anomaly detection where infusion of the knowledge about anticipated anomalies may result in better performance while maintaining a baseline performance nonetheless. To evaluate the general applicability of our approach, we integrate it with multiple existing recent anomaly detection methods and assess its efficacy on three widely used anomaly detection datasets: ShanghaiTech, Avenue, and Ped2. Our proposed approach noticeably outperforms existing methods, demonstrating its effectiveness in detecting anomalies across a range of contexts. The implementation of our method is available at https://github.com/kdawoud91/FusedVision.

1. Introduction

Anomaly detection in surveillance videos has several important applications, e.g., public safety, facility protection, and traffic monitoring[7, 20]. Due to the large amounts of incoming data, it is not scalable to assign human operators for data analysis making it critical to have autonomous anomaly detection systems. The mainstream research on anomaly detection addresses the problem as one-class classification (OCC) in which a familiarity model is trained on normal samples [17, 18, 22, 25, 32, 34, 41, 43, 50]. Then at test time, the model is expected to segregate anomalies from the normal data. With the introduction of autoencoders (AE), OCC research advanced significantly with researchers proposing several reconstruction based normalcy learning models [12, 23, 31, 41, 37, 48]. However, as reported widely [46, 47, 1], an AE can often generalize too well to start reconstructing anomalous inputs as well. To overcome such limitations of AEs, a more recent and well distinguished direction in OCC based anomaly detection is to address the task as object centric [10, 9, 40, 5, 6, 17]. In these approaches, a *pre-trained* object detector is often used to extract all objects and pass these to a learning system. As most of the anomalies in real-world can be tied to object appearances or their interactions, such pipelines make the anomaly detection task easier for the learning network, consequently making these methods superior against the conventional anomaly detection approaches [31, 1, 34].

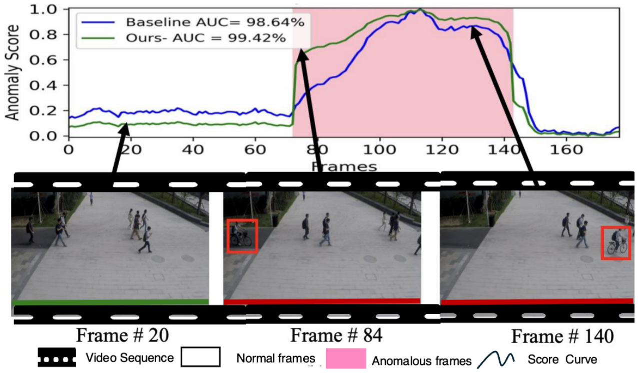

Figure 1. FusedVision responds rapidly to anomalies, achieving a significant anomaly score around frame #75 while maintaining consistently high anomaly detection performance up to frame #140. Once the anomaly disappears around frame #140, the anomaly score gradually decreases, returning to normal levels. Compared to the baseline, our approach demonstrates lower latency and higher anomaly scores.

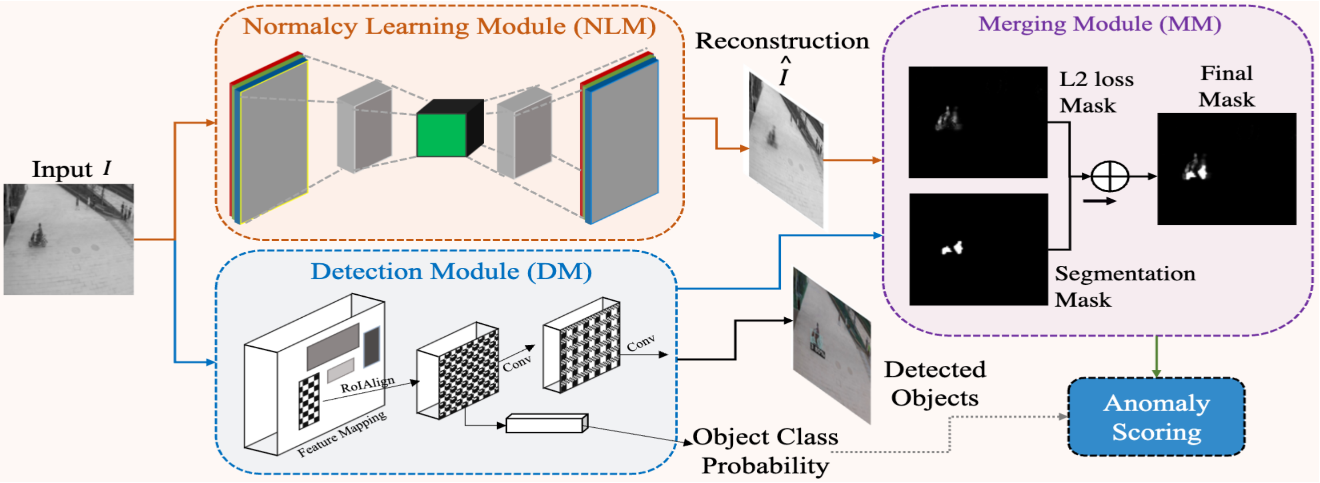

Figure 2. We present FusedVision, our proposed framework that integrates the Normalcy Learning Module (NLM) and the Detection Module (DM) through a Merging Module (MM) in a parallel branched configuration. NLM and DM are trained independently, and MM combines their outputs. The input image shows a person riding a motorcycle in a pedestrian area. Best viewed in color.

Despite being highly popular, the way object detectors are used in such approaches poses some fundamental drawbacks. As each frame of the input goes through object detector before being passed to the learning system, any failure at this step results in the subsequent network to fail in detecting anomalies. Although it may be argued that the object detectors are becoming robust, failures may still occur due to several reasons, e.g., low lighting, distant objects, or low resolution. Additionally, they may fail in detecting anomalies that are not related to objects. Moreover, object detectors typically cannot detect objects outside of the classes they have been trained on. Therefore, the conventional object centric methods, where learning system is placed subsequently after the object detector, may at times fail at anomaly detection[10, 40].

To alleviate these issues, we present a new method that enables modifications to the basic structure of the traditional one-class classification techniques. Instead of assuming that an object detector should assist the learning system, we modify the design by proposing a framework that, at an abstract level, utilizes both object detector and normalcy learning system side by side in a branched network. We hypothesize that such a design may help in overcoming the weaknesses of both components, with each assisting the other towards the ultimate goal of detecting anomalies. Fig.1 shows that our approach detects anomalies faster and achieves a better anomaly score compared to its baseline. In addition, to optimize the performance of such a branched network, we explore several design choices for branch fusion and anomaly score calculation. The results of these studies are thoroughly discussed in this work. We conduct extensive experiments on three anomaly detection datasets, i.e., ShanghaiTech [29], Avenue[28] and Ped2 [30]. Results demonstrate significant performance gains over the existing state-of-art (SOTA) methods. Moreover, analysis suggests that our approach holds a fail-safe design for real-world applications where infused knowledge about anomalies results in better anomaly detection performance while maintaining baseline performance nevertheless in case no information about anomalies is provided, making it a practical solution for real-world anomaly detection.

In summary our contribution is multifold:

- We propose a novel approach towards one-class classification based anomaly detection that utilizes both a normalcy learning model and an object detector in a parallel branched design, effectively addressing the limitations inherent to each component.
- We explore and discuss different fusion mechanisms as well as anomaly score calculation methods to facilitate future anomaly detection research.
- Our framework offers anomaly explanation as well as superior localization capability, which can facilitate categorization or prioritization of the responses to anomalous events.

2. Related Work

It is common among researchers to formulate anomaly detection as one-class classification problem where a network is typically trained on normal data and anomalies are present only at inference time [46, 1, 17, 18, 25, 32, 34, 35, 16]. Due to the nature of the problem, reconstruction based models have become the de facto standard in anomaly detection. Such models typically consist of regenerative autoendoers [1, 2, 47, 21, 44, 49, 15] which are trained on normal samples. At test time, it is expected that such models yield low reconstruction errors on normal samples and higher reconstruction errors on anomalous samples. One of the noticeable benefits of autoencoder based approaches is their general applicability on several types of anomalies and different modalities. For example, Zaheer *et al*. [46, 47] and Zhang *et al*. [49] utilized modified autoencoder architectures with similar underlying designs to detect anomalies in videos, images, and tabular datasets. However, a significant drawback of such approaches is that autoencoders are easy to overtrain. To counteract this, recent studies have introduced enhanced autoencoder frameworks. For example, Ristea *et al*. [37] proposed a self-distilled masked autoencoder that prioritizes high-motion areas using motion gradient weighting, preventing background overfitting and improving anomaly detection efficiency. Notably, object centric methods have been introduced recently to overcome the drawbacks of standalone autoencoders [8, 29, 31, 32, 35, 38]. In object centric methods, an object detector framework, e.g. YOLO [36] or Faster RCNN [11], is used to first detect all objects appearing in the input frames. A learning model is then trained using the cropped regions and their mutual relations which makes the learning task simpler and often more efficient [10, 9, 17, 14]. As the anomalies usually occur around moving objects/humans, object centric methods become more targeted, thus demonstrating superior performance on anomaly detection in surveillance videos [10, 9, 40, 17, 45]. Although generally helpful, object centric methods also pose several drawbacks on top of the learning models. For example, in [9] the authors have proposed to create tracks of objects over multiple frames, which are then used to train a convolution model. In case the object detector fails to detect an object, the subsequent network will not be able to recover it. Our approach to object-centric anomaly detection is novel because it combines the strengths of both autoencoders and object detectors in a branched network to calculate the anomaly scores. Unlike conventional approaches [40, 10, 17], our framework proposes a unique perspective, taking advantage of the synergistic integration of object detectors and autoencoders to achieve optimal results.

3. Method

FusedVision consists of three distinct components namely Normalcy Learning Module (NLM), Detection Module (DM), and Merging Module (MM). Each of these are visualized in Fig. 2 and discussed below:

3.1. Normalcy Learning Module (NLM)

In order to learn from the normal training videos, mainstream approaches in one-class classification (OCC) use some variants of autoencoders (AE) which learn to reconstruct the normal input data. The intuition behind using AEs is that, as the network only sees normal data during training, anomalous samples will be reconstructed with a higher error at inference time. Given $\mathbf{V} = \left[I_{1,} \, \ldots, I_T\right]$ is a video containing T frames, the output of an AE for an input image $I_{t\in[1,T]}$ can typically be defined as:

$$\hat{I}_t = \mathscr{D}_\theta(\mathscr{E}_\gamma(I_t)), \tag{1}$$

where $\hat{I}_t$, $\mathscr{E}_\gamma$, and $\mathscr{D}_\theta$ are reconstructed image, encoder and decoder, respectively. Typically, AEs are trained to minimize the reconstruction loss between input frame and its reconstruction as follows:

$$L_{RE} = \frac{1}{C \times H \times W} \left\|\hat{I}_t - I_t\right\|_F^2, \tag{2}$$

where C, H, and W are the channels, height, and width of the video frames, respectively. $\|\cdot\|_F$ is the Frobenius norm.

NLM is trained using only normal training data which effectively makes it a one-class classification model. This makes our design generic and gives us the flexibility to use just about any off-the-shelf one-class classification architecture to detect unseen anomalies. To this end, in this paper, we have experimented with several different state-of-the-art normalcy learning architectures previously introduced in the anomaly detection literature [25, 32, 1, 42, 9, 37],. The final output of NLM is a mask $\mathscr{M}_t^{NLM}$ with values ranging from 0 to 1 calculated by using the L2 loss between the input image and its reconstruction at each pixel. It is pertinent to mention that the reconstruction loss function described in Eq. 2 is generic. Particularly, we follow the loss functions of the original normalcy learning architectures we utilize in our approach as NLM. For instance, in addition to the reconstruction loss defined in Eq. 2, Park *et al*. [32] included feature compactness and feature separateness losses for the memory module inserted between the encoder and the decoder of their approach. However, the fundamental normalcy learning concept remains unchanged.

3.2. Detection Module (DM)

The Detection Module (DM) in our approach consists of an object detection network that is responsible for detecting and localizing objects within the input frame. To achieve this, we integrate a pre-trained object detector, following the essense of existing object-centric methods [9, 17]. Our DM is designed to output a single channel binary mask $\mathscr{M}_t^{DM}$ of the same size as the input image I_t ($H \times W$), highlighting the regions where anomalous objects are detected. To detect these objects, the DM utilizes the class IDs of predefined anomalous objects in a given dataset, while disregarding other classifications.

3.3. Merging Module (MM)

To achieve successful anomaly detection and localization, our proposed FusedVision needs to utilize a fusion strategy between the output of the Detection Module (DM) and the output of the Normalcy Learning Module (NLM). The goal is to fuse the two masks, $\mathscr{M}_t^{NLM}$ and $\mathscr{M}_t^{DM}$, in such a way that effectively allows us to reap the maximum benefits of the two modules. To this end, the two suitable design choices are discussed below:

Pixel-wise average based fusion. In order to fuse $\mathscr{M}_t^{NLM}$ and $\mathscr{M}_t^{DM}$, we may perform a pixel-wise average as:

$$\mathscr{M}_t^{AVG} = \frac{\mathscr{M}_t^{NLM} + \mathscr{M}_t^{DM}}{2}, \tag{3}$$

The intuition behind this pixel-wise average is to assign equal importance to each component of our FusedVision approach.

Pixel-wise max value based fusion. Another possibility is to fuse $\mathscr{M}_t^{NLM}$ and $\mathscr{M}_t^{DM}$ by taking a pixel-wise max between the two masks as:

$$\mathscr{M}_t^{MAX} = max(M_t^{NLM}, M_t^{DM}). \tag{4}$$

The intuition here is to emphasize on the anomalous regions highlighted by both or either of the two branched modules in our FusedVision approach. A comparison analysis of these two fusion methods are provided in Section 4.

3.4. Anomaly Score Calculation

To compute the final anomaly scores, we first discuss the peak signal-to-noise ratio (PSNR) based approach commonly used in existing literature for anomaly calculation in reconstruction-based models [1, 32]. Subsequently, we propose several alternative superior methods to compute anomaly scores that optimize the utilization of both the normalcy learning and detection modules.

3.4.1 PSNR as Anomaly Score

In accordance with recent literature on anomaly detection [1, 4, 32], we employ Peak Signal to Noise Ratio (PSNR) as a metric to quantify the similarity between an input image and its corresponding reconstruction. The PSNR score is computed as follows:

$$\hat{P}_t = 10\log_{10} \frac{max(\hat{I}_t)^2}{\frac{1}{C \times H \times W} \left\|\hat{I}_t - I_t\right\|_F^2}. \tag{5}$$

The anomaly score is then computed using the following min-max normalization over whole video frames as:

$$\alpha_t^{NLM} = 1 - \frac{\hat{P}_t - \min_t\left(\hat{P}_t\right)}{\max_t\left(\hat{P}_t\right) - \min_t\left(\hat{P}_t\right)}, \tag{6}$$

where α_t^{NLM} denotes the normalized PSNR based anomaly score ranging between 0 and 1 of an image at time t in a given video. This conventional anomaly scoring approach solely considers the reconstructed image and the original input image, thus cannot make use of the detection module of our system. Therefore, if this approach is used as is, it may not be able to leverage the benefits offered by our proposed architecture. To address this issue, we propose two modified PSNR-based anomaly scoring approaches that incorporate both the normalcy learning module (NLM) and the detection module (DM) of our system. The details of these approaches are discussed in the subsequent sections.

3.4.2 Reinforced Reconstruction based PSNR as Anomaly Score

In order to maximize the benefits of both the normalcy learning module (NLM) and detection module (DM), we propose a novel approach for computing anomaly scores, called reinforced reconstruction-based PSNR. The method involves using pixel-wise average based fusion mask $\mathscr{M}_t^{AVG}$ generated in Eq. 3 to mask the anomalous regions of the reconstructed image $\hat{I}$. Specifically, we create a reinforced reconstruction $\tilde{I}$ of the input image I_t as:

$$\tilde{I}_t^{AVG} = (1 - \mathscr{M}_t^{AVG}) \odot \hat{I}_t \tag{7}$$

where $\odot$ denotes element-wise multiplication. Subsequently, Eq. 5 to calculate the PSNR score takes the form:

$$\tilde{P}_t^{AVG} = 10\log_{10} \frac{max(\tilde{I}_t^{AVG})}{\frac{1}{C \times H \times W} \left\|\tilde{I}_t^{AVG} - I_t\right\|_F^2} \tag{8}$$

Finally, the anomaly score using the reinforced PSNR $\tilde{P}_t^{AVG}$ is given as:

$$\beta_t^{AVG} = 1 - \frac{\tilde{P}_t^{AVG} - \min_t\left(\tilde{P}_t^{AVG}\right)}{\max_t\left(\tilde{P}_t^{AVG}\right) - \min_t\left(\tilde{P}_t^{AVG}\right)} \tag{9}$$

In addition, we explore the reinforced PSNR evaluation using the pixel-wise max value based fusion mask $\mathscr{M}_t^{MAX}$ obtained using Eq. 4. The resulting anomaly score is denoted as β_t^{MAX} in the subsequent parts of our manuscript. The overall reinforced reconstruction based PSNR calculation method is depicted in Fig. 3.

3.4.3 PSNR-Based Anomaly Detection with Object Confidence Score

In line with object-centric anomaly detection techniques [9, 17], we can incorporate the anomalous object confidence score generated by the detector in anomaly scoring. The detection module (DM) outputs a confidence score between 0 and 1, indicating the likelihood of an anomalous object's presence in the input frame. To unleash the true potential of our branched FusedVision architecture, we propose combining the PSNR based anomaly score calculated in Eq. 6

with the anomalous object confidence score (ζ) predicted by DM:

$$A_t^{comb} = \frac{\alpha_t^{NLM} + \zeta_t}{2} \tag{10}$$

Here, A_t^{comb} represents the combined anomaly score of an image at time t in a video. If multiple anomalous objects are detected in a frame, we consider the maximum confidence score among the objects present within that frame as ζ_t. This ensures that the most significant anomalies are considered for the final score calculation.

4. Experiments

In this section, we discuss the experimental details to evaluate FusedVision, compare the performance with previous state-of-the-art methods, and perform an ablation study to determine the significance of each component.

4.1. Experimental Setup

4.1.1 Datasets

We evaluate our method on three publicly available video anomaly detection datasets, namely ShanghaiTech [29], Avenue[28], and Ped2 [30], which are designed by following one-class classification (OCC) protocol in which only normal frames are used for training and both normal and anomalous frames are used during testing.

ShanghaiTech [29]. This dataset is the most comprehensive surveillance video anomaly detection dataset that includes 330 training videos and 107 testing videos. It contains a wide range of anomalous events such as fighting, running, and vehicles appearing in pedestrian areas.

Avenue [28]. This dataset includes 16 training videos and 21 testing videos featuring diverse anomaly events, such as bicycles in pedestrian areas, and throwing items.

Ped2 [30]. This dataset contains 16 training and 12 testing videos. Normal events include pedestrian walking, while abnormal events include trucks, cyclists and skateboarders.

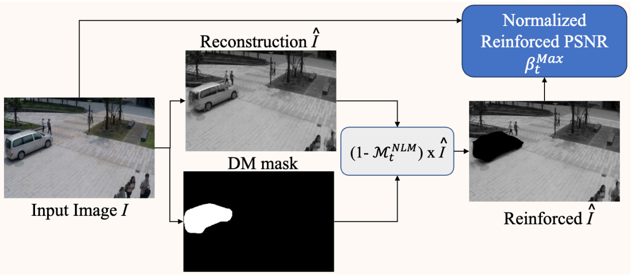
Reconstruction $\hat{I}$
Normalized Reinforced PSNR β_t^{Max}
Input Image I
DM mask
(1- $\mathcal{M}_t^{NLM}$) x $\hat{I}$
Reinforced $\hat{I}$

Figure 3. Anomaly score calculation using max reinforced reconstruction based PSNR.

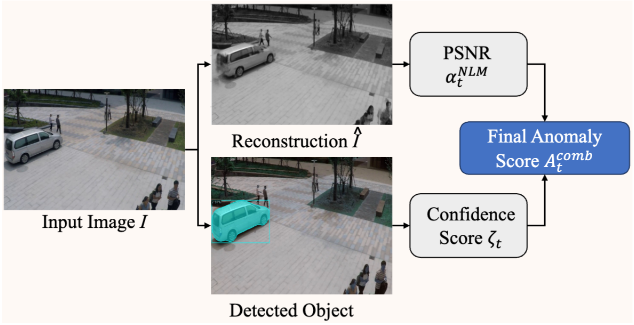

Figure 4. Shows anommaly score calculation using normalized PSNR with anomalous object confidence score

4.1.2 Evaluation Metrics

For frame-level anomaly detection, we use the widely popular [32, 37, 25, 1], area under the receiver operating characteristic curve (AUC) metric to evaluate our approach. The receiver operating characteristic (ROC) curve is calculated by plotting the true positive rate (TPR) against the false positive rate (FPR). For pixel-level anomaly detection, we use the RBDC and TBDC metrics, recently introduced in [33]. RBDC only marks a detected region as true positive if the intersection over union (IOU) with the ground truth region is above a predefined threshold a, while TBDC marks an anomaly track as true positive if the number of detections in a track is greater than a threshold b. We use default values of $a = 0.1$ and $b = 0.1$, following [33, 10].

$$RBDC = \frac{\text{num. of anomalous regions detected}}{\text{total num. of anomalous regions}} \tag{11}$$

$$TBDC = \frac{\text{num. of anomalous tracks detected}}{\text{total num. of anomalous tracks}} \tag{12}$$

4.1.3 Normalcy Learning Module (NLM) and Detection Module (DM) Settings

To demonstrate the general applicability of our proposed FusedVision approach, we have selected six recent state-of-the-art one-class classification-based anomaly detection models [25, 32, 1, 9, 42, 37] as candidates for our proposed NLM. Our selection was made arbitrarily and based on the public availability of the official code. For DM, we utilize Mask RCNN [13] Resnet101-FPN pre-trained on COCO dataset [24]. We adopt pre-defined anomalous objects classes from the COCO dataset. Particularly, for all datasets, the anomalous classes are set to be [2,3,4,8,25,37], based on the context provided by the dataset descriptions. Additional details are provided in the Supplementary. We exclude humans from the detection process of the object detector as they are typically associated with interaction-based

Method	ShanghaiTech	Avenue	Ped2
Astrid et al. [3]	73.7	72.0	98.40
Szymanowicz et al. [40]	70.4	75.3	84.4
Liu et al. [26]	74.2	89.9	99.3
Liu et al. [27]	85.0	93.6	-
Sun et al. [39]	78.6	91.5	99.4
Liu et al. [25]	72.80	85.09	95.40
Liu et al. [25] + FusedVision†	77.16 (+ 4.36)	85.09 (0.00)	96.30 (+ 0.9)
Park et al. [32]	68.30	80.69	94.30
Park et al. [32] + FusedVision†	73.75 (+ 5.45)	81.68 (+ 0.99)	95.94 (+ 1.64)
Astrid et al. [1]	75.97	84.67	96.50
Astrid et al. [1] + FusedVision†	78.83 (+2.86)	85.04 (+0.13)	97.42 (+ 0.92)
Georgescu et al. [9]	70.32	92.80	95.29
Georgescu et al. [9] + FusedVision†	73.78 (+3.46)	92.80 (0.00)	95.01 (+0.28)
Wang et al. [42]	84.25	92.20	98.88
Wang et al. [42] + FusedVision†	85.16 (+0.91)	91.93 (-0.27)	98.88 (0.00)
Ristea et al. [37]	79.1	91.30	95.40
Ristea et al. [37] + FusedVision†	83.58 (+4.48)	91.30 (0.00)	98.86 (3.46)

Table 1. Performance comparisons of our framework with existing state-of-the-art approaches on ShanghaiTech, Avenue, and Ped2 datasets. Generally, our FusedVision approach shows significant performance gains over the six existing anomaly detection methods [25, 1, 32, 9, 42, 37] used as baselines. In cases where performance gains are not observed, the score is either maintained or experiences only a negligible drop, attributed to the branched structure of our approach. Note that, we report the anomaly score using Eq. 10. Blue color marks performance gains while Red color marks performance drops.

anomalies (motions and actions). NLM is responsible for detecting human actions and interactions related anomalies.

4.2. Analyzing Anomaly Scoring Methods

As discussed in Section 3, we examine various approaches, including reinforced reconstruction based PSNR anomaly scores (β_t^{AVG} & β_t^{MAX}) in Eq. 9 and PSNR based anomaly detection with object confidence score (A_t^{comb}) in Eq. 10 within our framework. Tab.2 presents our evaluation results of these approaches, compared to the conventional PSNR-based anomaly scoring (α_t^{NLM}) in Eq. 6, on two different normalcy learning baselines [32, 1]. As shown, all of our proposed approaches successfully outperform the baseline PSNR anomaly scoring method. This shows the importance of using both DM and NLM together towards the final anomaly scoring compared to using NLM standalone. Among our proposed methods, A_t^{comb}, which is computed by averaging the confidence score output by DM and the PSNR value by NLM, outperforms its counterpart methods by a significant gain of 5.45% and +4.48% in terms of AUC compared to their respective baselines [32, 37]. Based on these analyses, A_t^{comb} is our default anomaly scoring configuration throughout the rest of the manuscript.

	Methods	Park et al. [32]		Astrid et al.[1]	
		AUC (%)	**Δ**	**AUC (%)**	**Δ**
PSNR Baseline	α_t^{NLM} (Eq. 6)	68.30	0.00	75.97	0.00
Ours	A_t^{comb} (Eq. 10	**73.75**	+ 5.45	**78.83**	+ 2.86
	β_t^{AVG} (Eq. 9)	72.03	+ 3.73	77.04	+ 1.07
	β_t^{MAX} (Eq. 9)	71.64	+ 3.34	77.45	+ 1.48

Table 2. AUC % of our proposed anomaly score calculation approaches on ShanghaiTech Dataset. As seen, all of our approaches perform noticeably better than the corresponding baseline categories. Top method in each category is highlighted as bold.

4.3. Comparisons with State-Of-The-Art (SOTA)

The results, summarized in Table 1, show that on all datasets, our approach noticeably enhances the performance of the corresponding baseline methods in most cases. Notably, our approach achieved a remarkable improvement of 4.48% in terms of AUC on ShanghaiTech dataset when compared to using Ristea *et al.* [37] method as a baseline.

Furthermore, using Wang *et al.* [42] as baseline on the same dataset, FusedVision achieves a remarkable AUC performance of 85.16% on ShanghaiTech dataset. Interestingly, in the cases where FusedVision does not improve over the baselines, for example, Wang*et al.* [42] on Avenue dataset, the performance drop is almost negligible, i.e., 0.27. This is owed to the branched structure of our approach in which if the detection module fails to improve the anomaly detection, the performance almost stays similar to the baseline model. This fail-safe property is highly useful in terms of practical/real-world anomaly detection systems where if our FusedVision approach does not show any improved performance, it does not drop the performance of the overall system either. In addition, the performance gains of 12.79% and 15.42% by our approach in terms of RBDC and TBDC are observed in Table 3 when using Park *et al.* [32] as baseline. These noticeable gains are owed to the fusion of the masks output by NLM and DM in FusedVision which significantly enhances the quality of anomaly localization.

Method	**RBDC**	**TBDC**
Park et al.	19.81	53.35
Park et al + Ours	32.60 (+ 12.79)	68.77 (+ 15.42)

Table 3. RBDC and TBDC scores for [32] and our FusedVision framework. The improvement is indicated in blue.

In Fig. 5 we present the ROC curve comparisons of the baseline models [32, 25, 1, 9] with the performance after integrating these into our approach for anomaly detection on ShanghaiTech dataset. As seen, our FusedVision approach attains consistent and significant improvements over the baseline methods. Another interesting observation can be seen when comparing the ROC curves of the standalone Detection Module (DM) with our approaches. Although the object detector is capable of detecting anomalous objects on its own, the performance of the standalone detection module is considerably low. However, when the DM is combined with the NLM in FusedVision, the performance gains are significant highlighting its importance.

4.4. Ablation Study

We conducted an ablation study by systematically removing different branches and evaluating their individual contributions to the overall performance (Tab.4). For this study, we limit the experiments to using two baseline models, Park *et al*. [32] and Astrid *et al*. [1], on ShanghaiTech dataset. As seen, the performance of our proposed approach is significantly better than that of each branched module in a standalone configuration. DM on its own reaches only 63.40% AUC. Considering Park *et al*. [32] and Astrid *et al*. [1] baselines, the NLM demonstrates an AUC performance of 68.30% and 75.97% respectively. With our FusedVision approach, the AUC performance of 73.75% and 78.83% is observed respectively. Compared to the performance of standalone DM, its a significant improvement of 10.35% and 15.43% respectively. These experiments suggest that relying solely on object detectors or normalcy learning models may not be sufficient for detecting anomalies in complex environments. This also highlights the advantage of a more comprehensive approach like ours.

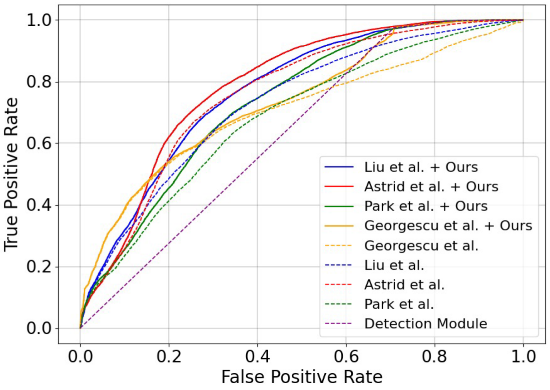

Figure 5. ROC curve comparisons of our approach with each of the baselines used in our experiments.

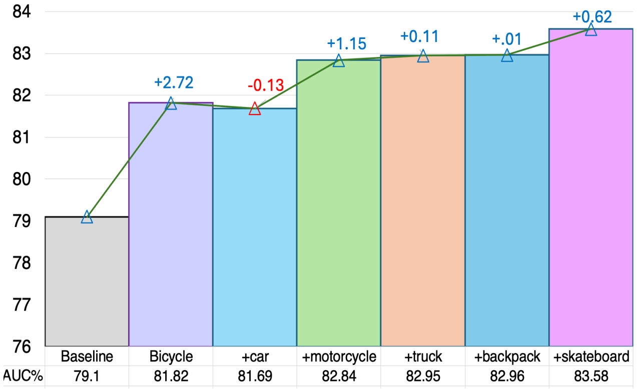

Figure 6. AUC gains over Ristea *et al*. [37] baseline when object knowledge is infused to the detection module. While FusedVision generally maintains baseline performance in case object information is witheld, it demonstrates notable gains with each object added to the list of known anomalies. This highlights the practical application of FusedVision for real-world anomaly detection.

4.4.1 Knowledge Infusion

Fig 6 demonstrates the gradual improvements in performance as anomaly knowledge is infused via defining each object element in the Detection Module (DM). The experiment uses the model proposed by Ristea *et al*. as (NLM) baseline trained on ShanghaiTech dataset that achieves an AUC of 79.1%. When incorporating bicycle alone as an anomaly, the accuracy improves to 81.82%. Further infusing the (DM) with additional object knowledge systematically enhances performance, reaching 83.58%, resulting in a total accuracy gain of 4.48%. These results demonstrate that incorporating relevant knowledge into the DT module plays a crucial role in improving the overall network's accuracy. The results also highlight the fail-safe design of our approach, where no knowledge about the anomalous objects will result in FusedVision maintaining the baseline performance while improving over it significantly as the knowledge is added to the DM module, highlighting its practical application in real-world anomaly detection scenarios.

4.5. Qualitative Analysis

Fig. 7 presents two different input images and their corresponding output images obtained at various stages of our framework. First image is a frame containing a person riding a motorcycle within a padestrian area Fig. 7 (a). The output of the NLM, represented by Fig. 7 (b), indicates that some parts of the motorcycle are not properly reconstructed, leading to their highlighting in the reconstruction mask Fig. 7 (d). On the other hand, Fig. 7 (c) demonstrates the detection of this motorcycle by the DM and Fig. 7 (e) shows the

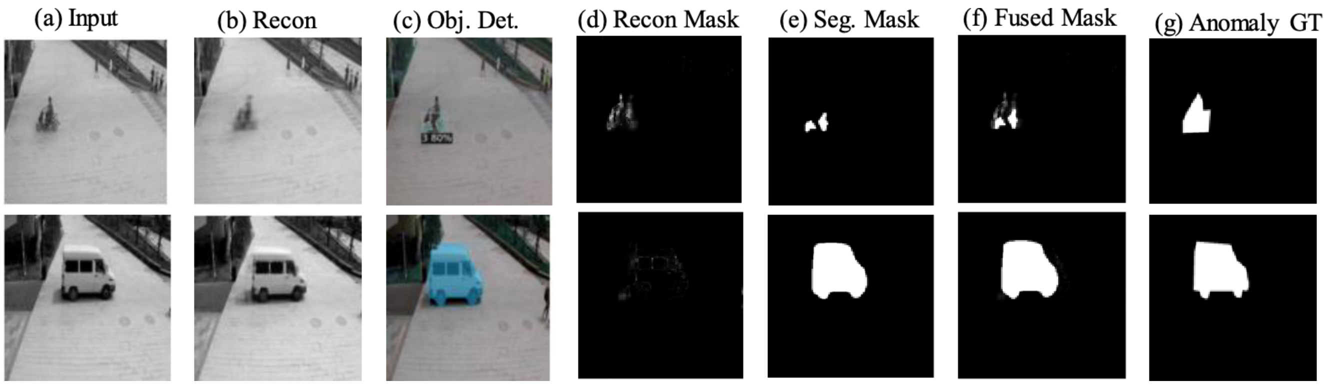

Figure 7. Visualization of inputs and outputs at several intermediate steps of our framework. (a) Input image (b) Reconstructed image thorough NLM. (c) Output of the DM (d) Reconstruction mask. (e) The anomalous object mask created by DM. (f) Final mask after fusion in MM. (g) Anomaly groundtruth. As seen, our final masks produced in (f) localize anomalies noticeably well.

DM	NLM	AUC (%) Park et al. [32]	AUC (%) Astrid et al. [1]
✓		63.40	63.40
	✓	68.30	75.97
✓	✓	73.75	78.83

Table 4. Ablation study of our approach. Detection Module (DM) shows subpar performance if used standalone. Normalcy Learning Module (NLM), although demonstrates better performance than DM, the best performance is achieved when both modules are utilized side-by-side as proposed in our approach.

corresponding segmentation mask. As seen, the masks in Figs. 7 (d) & (e) are individually not fully adequate. Compared to these, the fused mask shown in Fig. 7 (f), generated by our proposed FusedVision approach, represents the anomaly better and is closer to the anomaly ground truth Fig. 7 (g). Another example is presented in the second row of the same figure, where the input image is a vehicle within a pedestrian area as shown in Fig. 7 (a). The NLM reconstructs the object sufficiently enough Fig. 7 (b) to not demonstrate any significant reconstruction error Fig. 7 (d). In a conventional OCC method, this failure may simply result in a misdetection of the anomalous region. However, the DM in our FusedVision approach successfully detectes the vehicle Fig. 7 (c) and localizes it Fig. 7 (e). Consequently, the final fused mask output by our approach successfully localizes the anomalous region demonstrating the superiority of our proposed approach.

4.6. Time Complexity Analysis

We compare the overhead of FusedVision with several existing methods, demonstrating the trade-off between detection accuracy and processing speed. The results show that FusedVision achieves significantly faster processing while maintaining strong anomaly detection. Our experiments are conducted on a single NVIDIA GeForce RTX 4090 with 24GB of VRAM. FusedVision achieves a processing speed of 47 FPS. This is nearly twice as fast as recent object-centric approaches, such as Georgescu *et al*. [9] with 23 FPS and Georgescu *et al*. [10] with 18 FPS.

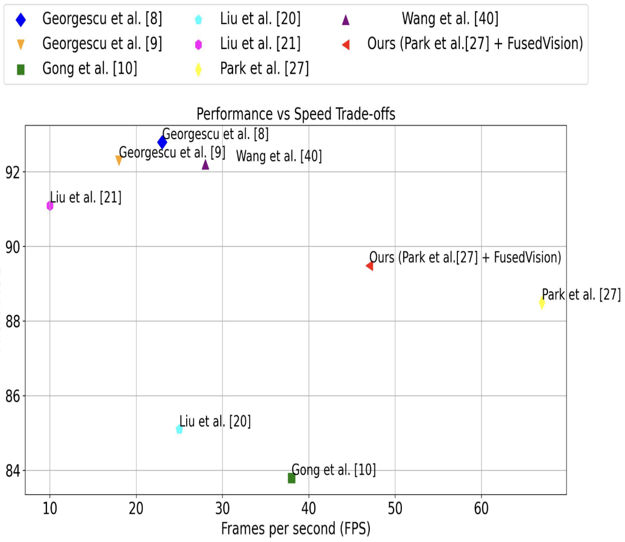

Figure 8. Performance vs. speed trade-offs for our FusedVision approach and several state-of-the-art methods [9, 10, 12, 26, 32, 42], with publicly available code. Our model (Park et al. [32] + FusedVision) achieves an optimal balance between detection accuracy and processing speed, reaching 47 FPS using YOLOv8 [19] as the detection module. Results reported on the Avenue dataset.

5. Conclusion

We proposed a one-class classification approach with a branched design in which a normalcy learning module and a detection module work side by side to detect anomalies. Two fusion mechanisms are also explored to facilitate further research. Moreover, to unleash the full capability of our tendem design, three anomaly score calculation methods are introduced. Our approach achieves notable performance on three anomaly detection datasets.

References

[1] Marcella Astrid, Muhammad Zaigham Zaheer, Jae-Yeong Lee, and Seung-Ik Lee. Learning not to reconstruct anomalies. *arXiv preprint arXiv:2110.09742*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[2] Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Synthetic temporal anomaly guided end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 207–214, 2021. 2

[3] Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. Synthetic temporal anomaly guided end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 207–214, October 2021. 6

[4] Fei Dong, Yu Zhang, and Xiushan Nie. Dual discriminator generative adversarial network for video anomaly detection. *IEEE Access*, 8:88170–88176, 2020. 4

[5] Keval Doshi and Yasin Yilmaz. Any-shot sequential anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 934–935, 2020. 1

[6] Keval Doshi and Yasin Yilmaz. Continual learning for anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 254–255, 2020. 1

[7] Huu-Thanh Duong, Viet-Tuan Le, and Vinh Truong Hoang. Deep learning-based anomaly detection in video surveillance: A survey. *Sensors*, 23(11):5024, 2023. 1

[8] Ye Fei, Chaoqin Huang, Cao Jinkun, Maosen Li, Ya Zhang, and Cewu Lu. Attribute restoration framework for anomaly detection. *IEEE Transactions on Multimedia*, 2020. 3

[9] 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, 3, 4, 5, 6, 7, 8

[10] 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. 1, 2, 3, 5, 8

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 3

[12] 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. 1, 8

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 5

[14] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *Proceedings of the IEEE international conference on computer vision*, pages 3619–3627, 2017. 3

[15] Jaehui Hwang, Junghyuk Lee, and Jong-Seok Lee. Anomaly score: Evaluating generative models and individual generated images based on complexity and vulnerability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8754–8763, June 2024. 2

[16] Jiin Im, Yongho Son, and Je Hyeong Hong. Fun-ad: Fully unsupervised learning for anomaly detection with noisy training data. *arXiv preprint arXiv:2411.16110*, 2024. 2

[17] 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. 1, 2, 3, 4

[18] Radu Tudor Ionescu, Sorina Smeureanu, Marius Popescu, and Bogdan Alexe. Detecting abnormal events in video using narrowed normality clusters. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1951–1960. IEEE, 2019. 1, 2

[19] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 8

[20] Sardar Waqar Khan, Qasim Hafeez, Muhammad Irfan Khalid, Roobaea Alroobaea, Saddam Hussain, Jawaid Iqbal, Jasem Almotiri, and Syed Sajid Ullah. Anomaly detection in traffic surveillance videos using deep learning. *Sensors*, 22(17):6563, 2022. 1

[21] Mingyu Lee and Jongwon Choi. Text-guided variational image generation for industrial anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26519–26528, June 2024. 2

[22] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013. 1

[23] Zhenyu Li, Ning Li, Kaitao Jiang, Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Superpixel masking and inpainting for self-supervised anomaly detection. In *Bmvc*, 2020. 1

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[25] 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. 1, 2, 3, 5, 6, 7

[26] 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, 8

[27] Zuhao Liu, Xiao-Ming Wu, Dian Zheng, Kun-Yu Lin, and Wei-Shi Zheng. Generating anomalies for video anomaly detection with prompt-based feature mapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24500–24510, 2023. 6

[28] 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. 2, 5

[29] 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. 2, 3, 5

[30] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1975–1981. IEEE, 2010. 2, 5

[31] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1273–1283, 2019. 1, 3

[32] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14372–14381, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[33] 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. 5

[34] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, Enver Sangineto, and Nicu Sebe. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1689–1698. IEEE, 2018. 1, 2

[35] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE international conference on image processing (ICIP)*, pages 1577–1581. IEEE, 2017. 2, 3

[36] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[37] Nicolae-C Ristea, Florinel-Alin Croitoru, Radu Tudor Ionescu, Marius Popescu, Fahad Shahbaz Khan, Mubarak Shah, et al. Self-distilled masked auto-encoders are efficient video anomaly detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15984–15995, 2024. 1, 3, 5, 6, 7

[38] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021. 3

[39] Che Sun, Chenrui Shi, Yunde Jia, and Yuwei Wu. Learning event-relevant factors for video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2384–2392, 2023. 6

[40] Stanislaw Szymanowicz, James Charles, and Roberto Cipolla. X-man: Explaining multiple sources of anomalies in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3224–3232, 2021. 1, 2, 3, 6

[41] Yao Tang, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129:123–130, 2020. 1

[42] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 494–511. Springer, 2022. 3, 5, 6, 8

[43] Peng Wu, Jing Liu, and Fang Shen. A deep one-class neural network for anomalous event detection in complex scenes. *IEEE transactions on neural networks and learning systems*, 31(7):2609–2622, 2019. 1

[44] Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5527–5537, October 2023. 2

[45] 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*, pages 583–591, 2020. 3

[46] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14193, 2020. 1, 2, 3

[47] Muhammad Zaigham Zaheer, Jin-Ha Lee, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Stabilizing adversarially learned one-class novelty detection using pseudo anomalies. *IEEE Transactions on Image Processing*, 31:5963–5975, 2022. 1, 2, 3

[48] Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, and Jianxin Liao. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17385–17394, 2024. 1

[49] Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, and Jianxin Liao. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17385–17394, June 2024. 2, 3

[50] Xinyi Zhang, Naiqi Li, Jiawei Li, Tao Dai, Yong Jiang, and Shu-Tao Xia. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6782–6791, October 2023. 1