

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

This is supplementary material for VAND 3.0: Visual Anomaly and Novelty Detection, submission ID 23.

## 1. Motivation

Since the definition of anomaly can vary depending on the context, it is critical to handle a broad spectrum of anomalies in surveillance videos, anomalous events (e.g., running or throwing trash), abnormal interactions between normal objects (e.g., fighting or stealing), and appearance of anomalous objects (e.g., bicycle or vehicle on pedestrian walkways). To this end, we propose FusedVision, a branched framework capable of leveraging object-centric and normalcy learning modules to successfully detect a wide range of anomalies. Through rigorous experiments on ShanghaiTech, Avenue, and Ped2 datasets, we have demonstrated the remarkable efficacy of our proposed framework.

## 2. Anomaly Score Calculation

As discussed in (manuscript:Section 3.4) Anomaly Score Calculation, we use the widely popular in literature Peak Signal to Noise Ratio (PSNR) as the baseline anomaly scoring approach (manuscript:Section 3.4.1). Furthermore, we introduce three new anomaly scoring methods ( $\beta_t^{AVG}$ ,  $\beta_t^{MAX}$ , and  $\beta_t^{comb}$ ) which effectively utilize the two branched structure of our framework and yield significant performance gains over the baseline PSNR based scoring method (manuscript:Table 2 of Manuscript).

### 2.1. Mask values as anomaly score

As discussed in some of the existing anomaly detection literature, PSNR as anomaly scoring has some limitations [3, 1]. For example, using PSNR as anomaly score requires it to be min-max normalized over a whole sequence of video. Consequently, this means that PSNR is not easily applicable on real-time systems.

To overcome this limitation of PSNR based anomaly scoring method, we also explore a simple mask value-based anomaly scoring approach which takes the maximum pixel value of the fused mask generated by our framework ( $\mathcal{M}_t^{AVG}$  - manuscript: Eq. 3). The intuition behind this approach is that as the fused mask is created by merging the

NLM and the DM masks, higher values of it may correspond to the presence of anomalous event or objects in the input frame. Therefore, the pixel value can be considered as the anomaly score: Formally, for the mask  $\mathcal{M}_t^{AVG}$ , we compute this proposed anomaly score  $\eta_t$  as:

$$\eta_t = \max(\mathcal{M}_t^{AVG}). \quad (1)$$

As the values of this mask range between 0 and 1, the anomaly score for each input frame can be calculated independently without considering any normalization over the input video. This may make the system more applicable towards real-world anomaly detection.

The performance comparison of  $\eta_t$  as anomaly score with its counterpart PSNR based anomaly scoring methods proposed in our manuscript is provided in Table 1 of this Supplementary document. To limit the extent of the experiments, we utilize only the large-scale ShanghaiTech dataset for this study. As seen, anomaly score  $\eta_t$  demonstrates an AUC performance of 73.25% using Park *et al.* [5] as baseline NLM method. This performance is comparable to the PSNR based methods proposed in our manuscript with an additional benefit of not requiring any normalization.

## 3. Detection Module (DM) Settings

In this section, to facilitate the results reproducibility of our approach, we provide comprehensive details of the settings for our detection module (DM). As discussed in the existing object-centric anomaly detection methods, the anomaly is dependent on the context in which specific objects interact with each other or the environment [4, 2, 7]. To this end, for all datasets, we configure our DM module to detect the anomalous classes: 2, 3, 4, 8, 25, and 37, of the COCO dataset, based on the context provided by the dataset description. This configuration translates into our DM module identifying objects belonging to the bicycle, car, motorcycle, truck, backpack, and skateboard classes as anomalous. Interestingly, due to the presence of NLM in its branched architecture, our proposed FusedVision framework offers flexibility of detecting other anomalies such as event based anomalies or appearances of objects unknown to the DM. This property is in contrast with the existing

PSNR Baseline	Methods	Park et al. [5]	
		AUC (%)	$\Delta$
	$\alpha_t^{NLM}$ (manuscript: Eq. 6)	68.30	0.00
	$A_t^{comb}$ (manuscript: Eq. 10)	73.75	+ 5.45
Ours	$\beta_t^{AVG}$ (manuscript: Eq. 9)	72.03	+ 3.73
	$\beta_t^{MAX}$ (manuscript: Eq. 9)	71.64	+ 3.34
	$\eta_t$ (Supplementary: Eq. 1)	73.25	+ 4.95

Table 1. AUC % of our proposed mask based anomaly score calculation approaches compared with baseline PSNR method as well as our proposed PSNR based methods on ShanghaiTech Dataset. As seen, all of our approaches perform noticeably better than the PSNR baseline method. Moreover, the mask based method performs comparably with its counterpart PSNR based methods while having the flexibility of not requiring any normalization.

object centric methods in which if the detector fails to detect an object, the learning system predicts normalcy for the input. The results presented in manuscript:Table 4 validate this property. As seen, when only DM is used in a standalone configuration to detecting anomalies, it demonstrates a mere AUC performance of 63.40%. However, when Park *et al.* [5] is added as NLM in our FusedVision framework, the performance increases to 73.75%. Moreover, in another similar experiment, adding Astrid *et al.* [1] as NLM in our FusedVision framework brings the performance to 78.83%. Moreover, from manuscript:Table 1, adding Ristea *et al.* [6] as NLM bring the performance of our proposed approach to 83.58%. This consistent performance gains by adding different NLM baseline demonstrate the significant flexibility of our approach in performing beyond anomalous object detection capability of the DM used in our approach.

## 4. Qualitative Results

In this section, we present additional visualizations comparing our FusedVision approach with its baseline counterpart. Our analysis focuses on two critical aspects of video anomaly detection: the Area Under the Curve (AUC) score and the response time when an anomaly is introduced. As shown in Fig. 1, the baseline method proposed by Astrid *et al.* [1] exhibited an increase in detection scores around frame #80, despite these being normal frames, indicating suboptimal performance in distinguishing between normal and anomalous events. In contrast, our FusedVision approach demonstrated a more robust and accurate response, detecting the anomaly precisely around frame #100 when the bicycle first appeared. Furthermore, FusedVision maintained a consistently high AUC score over the anomalous frames, whereas the Astridet al. [1] method struggled to maintain reliable performance beyond frame #160. This improvement can be attributed to the integration of the detection module (DM), which enhances both the capture speed and accuracy of anomaly detection in FusedVision.

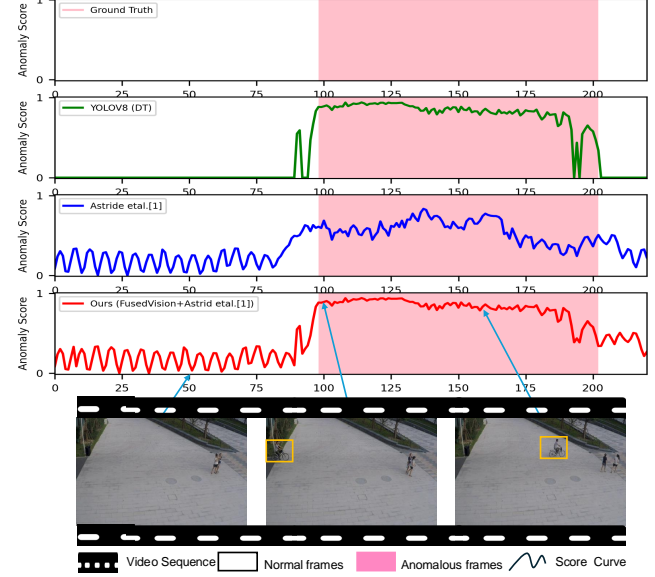


Figure 1. Anomaly score comparison between the generated Astrid *et al.* baseline [1] and our proposed approach (FusedVision). Our approach captures anomalies more quickly and provides a more robust anomaly score.

## 5. Demo Video

For additional qualitative results compilation of the proposed FusedVision framework on videos taken from ShanghaiTech dataset, please refer to the following link [https://drive.google.com/file/d/1r4bQr1Bc1ls4B02A-qKISxFXzEosvnuH/view?usp=drive\\_link](https://drive.google.com/file/d/1r4bQr1Bc1ls4B02A-qKISxFXzEosvnuH/view?usp=drive_link)

## 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
- [2] Ali Enver Bilecen and Hüseyin Özkan. Object-centric video anomaly detection with covariance features. In *2022 30th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4, 2022. 1
- [3] Wei Dai and Daniel Berleant. Discovering limitations of image quality assessments with noised deep learning image sets. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3735–3744. IEEE, 2022. 1
- [4] Yang Liu, Zhengliang Guo, Jing Liu, Chengfang Li, and Liang Song. Osin: Object-centric scene inference network for unsupervised video anomaly detection. *IEEE Signal Processing Letters*, 30:359–363, 2023. 1
- [5] 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
- [6] 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. [2](#)

- [7] Zhengye Yang and Richard J. Radke. Context-aware Video Anomaly Detection in Long-Term Datasets . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4002–4011, Los Alamitos, CA, USA, June 2024. IEEE Computer Society.

[1](#)