INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY TRENDS

Vol. 6 No. 02 (2026)

e.ISSN: 2409-5222

https://ijictt.ictt.info

Object-Centric Industrial Video Anomaly Detection with Local-Global Representation Learning

Chun-Wai So, Man-Kit Chau

Faculty of Computer Science Department, Huazhong University of Science and Technology, Wuhan 430074, China

School of Computer Science, Peking University, Beijing 100871, China

Abstract

Industrial video anomaly detection is a critical component of modern smart manufacturing and industrial surveillance, aiming to automatically identify deviations from normal operational patterns. Traditional methods relying on frame-level or pixel-level feature extraction often struggle with the complex, dynamic, and heavily occluded environments typical of industrial settings. This paper proposes a novel framework centered on object-centric video anomaly detection, heavily augmented by a local-global representation learning mechanism. By shifting the analytical focus from the entire image frame to specific objects of interest, such as machinery components, manufactured goods, and human operators, the proposed method isolates highly relevant features while mitigating the impact of background noise and illumination variations. The framework utilizes a robust tracking-by-detection paradigm to construct spatio-temporal object tubes, which are subsequently processed to extract localized representations capturing appearance and motion dynamics. Concurrently, a global representation module employs attention mechanisms to model the complex interactions between multiple objects and their contextual environment. The integration of these local and global streams ensures a comprehensive understanding of the industrial scene, allowing for the precise localization and classification of anomalous events. Extensive evaluations on multiple large-scale industrial datasets demonstrate that the proposed object-centric framework significantly outperforms existing state-of-the-art approaches in both anomaly detection accuracy and computational efficiency. The findings suggest that integrating structured object interactions into representation learning provides a highly scalable and robust solution for real-world industrial monitoring.

Keywords: *Video Anomaly Detection, Object-Centric Learning, Representation Learning, Industrial Surveillance, Computer Vision.*

I. INTRODUCTION

The advent of Industry 4.0 has precipitated a massive integration of automated surveillance and monitoring systems within manufacturing plants, assembly lines, and logistical hubs. As industrial environments become increasingly complex, the necessity for robust automated oversight mechanisms has grown exponentially. Video anomaly detection in these contexts serves as the primary line of defense against operational deviations, equipment malfunctions, safety protocol violations, and unauthorized intrusions [1]. Unlike general-purpose video anomaly detection, which often deals with pedestrian movements or traffic flow, industrial video anomaly detection faces a unique set of challenges. These include continuous mechanical motions, extreme variations in lighting conditions, dense spatial configurations of machinery, and highly specific definitions of anomalous behavior that depend entirely on the specific manufacturing process being observed [2]. The overarching goal is to deploy intelligent systems capable of learning the baseline of normal operations and raising alerts when subtle or overt deviations occur, thereby preventing costly downtimes or catastrophic accidents [3]. Historically, video anomaly detection has heavily relied on holistic approaches that analyze the entire frame simultaneously. Methods utilizing optical flow, spatio-temporal gradients, and global reconstruction objectives have been the cornerstone of early research [4]. However, these frame-level and pixel-level methodologies exhibit significant limitations when applied to the chaotic visual landscape of industrial facilities. The primary drawback is their susceptibility to background noise. In a typical manufacturing environment, background elements such as flashing warning lights, moving conveyor belts, and shifting shadows can dominate the visual field, leading holistic models to generate an unmanageable number of false positives [5]. Furthermore, pixel-level reconstruction methods, such as those based on autoencoders, often expend unnecessary computational resources attempting to perfectly reconstruct irrelevant background details, diluting their capacity to capture the nuanced features of actual objects of interest where anomalies are most likely to manifest [6]. To overcome these inherent limitations, recent paradigms have shifted towards object-centric architectures. By explicitly defining and focusing on discrete entities within the scene, object-centric video anomaly detection fundamentally alters the feature extraction process [7]. Instead of processing raw pixels agnostically, these methods leverage advanced object detection and tracking algorithms to isolate humans, robotic arms, vehicles, and products [8]. This isolation allows the anomaly detection model to analyze the localized appearance and motion of these entities directly, drastically reducing the interference from irrelevant background dynamics [9]. However, purely localized object analysis introduces a new vulnerability: the loss of contextual awareness. An action performed by an industrial robot might be perfectly normal in one context but highly anomalous in another, depending on its spatial and temporal relationship with surrounding operators or materials [10]. This paper introduces a comprehensive framework designed to reconcile the benefits of focused object analysis with the necessity of contextual awareness through a dual-stream local-global representation learning approach. The local stream is meticulously designed to scrutinize the intrinsic properties of individual object trajectories, evaluating fine-grained textures and micro-motions that indicate wear, tear, or improper execution of a specific task. Simultaneously, the global stream constructs a relational graph or attention-based map of the entire scene, modeling how individual objects interact with one another and their environment over time. By mathematically fusing these two distinct yet complementary representations, the proposed system can detect anomalies that are characterized either by isolated object failures or by the breakdown of coordinated activities between multiple entities. The primary contributions of this research are manifold. First, a highly robust object-centric extraction pipeline tailored specifically for industrial environments is formulated, addressing issues of persistent occlusion and rapid motion blur. Second, a novel local-global representation learning architecture is proposed, utilizing advanced attention mechanisms to dynamically weight the importance of local object anomalies versus contextual inconsistencies. Third, comprehensive experimental validation is provided across several challenging, realistic industrial datasets, establishing new benchmarks for accuracy and operational efficiency in the domain of automated industrial surveillance.

2. Literature Review

2.1 Traditional and Deep Learning Approaches in Video Anomaly Detection

The evolution of video anomaly detection has transitioned through several distinct methodological phases, heavily influenced by the prevailing computational capabilities and feature extraction techniques of the time. Early investigations primarily utilized handcrafted features to model normal behavior. Techniques such as Histogram of Oriented Gradients and Histogram of Optical Flow were frequently employed to capture spatial structures and temporal dynamics, respectively [11]. These handcrafted features were typically fed into traditional machine learning classifiers, such as One-Class Support Vector Machines or Gaussian Mixture Models, to delineate the boundary of normality [12]. While these methods provided a foundational understanding of temporal modeling, they proved incredibly fragile in real-world industrial environments. The rigid nature of handcrafted features meant that any slight change in camera angle, illumination, or baseline operation required extensive manual recalibration, rendering them unsuitable for scalable industrial deployment [13]. The introduction of deep learning catalyzed a monumental shift in the field, primarily through the adoption of unsupervised learning frameworks. Given the extreme rarity of anomalous events in real-world scenarios, supervised learning is generally considered intractable for video anomaly detection [14]. Consequently, reconstruction-based approaches using Autoencoders and Generative Adversarial Networks emerged as the dominant paradigm. The core hypothesis of reconstruction-based methods is that a model trained exclusively on normal video sequences will successfully reconstruct normal frames but will produce significant reconstruction errors when confronted with anomalous inputs [15]. While deeply influential, standard frame-level autoencoders suffer from the identity mapping problem, wherein the network possesses sufficient capacity to generalize and perfectly reconstruct even the anomalous inputs it has never seen during training [16]. To mitigate this, predictive methodologies were introduced. Instead of reconstructing the current frame, predictive models utilize sequences of past frames to forecast future frames. Deviations between the predicted future and the actual observed future serve as the anomaly score [17]. Despite showing improvements over pure reconstruction, both predictive and reconstructive deep learning models largely operate at the holistic frame level. In industrial settings filled with moving machinery and dynamic backgrounds, these models waste vast amounts of representational capacity on irrelevant regions, leading to high false-alarm rates when harmless background variations occur [18].

2.2 Object-Centric Methodologies and Representation Learning

Recognizing the inefficiencies of holistic processing, the research community has increasingly turned towards object-centric strategies. This shift is predicated on the psychological and operational reality that in structured environments, anomalies are inherently tied to specific objects or the interactions between them [19]. Object-centric video anomaly detection leverages pre-trained object detectors, such as variations of the You Only Look Once architecture or Region-based Convolutional Neural Networks, to extract bounding boxes around entities of interest [20]. These extracted regions are then tracked across frames to form spatio-temporal tubes, which serve as the primary input for subsequent anomaly scoring networks. Recent advancements in representation learning have significantly enhanced the capabilities of object-centric models. Contrastive learning frameworks have been adapted to push the representations of normal object trajectories closer together in the latent space while pushing synthesized or augmented anomalous trajectories further away [21]. Furthermore, masked modeling techniques, originally popularized in natural language processing and later adapted for vision tasks, have been applied to object trajectories. By masking segments of an object's spatial or temporal data and forcing a network to reconstruct the missing information, models learn highly robust and generalized features of normal object behavior [22].

However, the majority of existing object-centric methods evaluate isolated object tubes independently. They extract a bounding box, process its localized temporal evolution, and assign an anomaly score without considering the broader environment [23]. This isolated processing completely ignores contextual anomalies. For example, a robotic arm moving at a specific velocity might be normal during active assembly but anomalous if it occurs while a human worker is within the safety perimeter [24]. Some recent studies have attempted to incorporate context by computing the distance between object features and a global memory bank, but these methods often fail to capture the dynamic, evolving relationships between multiple moving entities [25]. The integration of sophisticated local-global relational modeling remains a critical gap in the literature, particularly for complex industrial applications where safety relies entirely on the precise coordination of multiple independent agents and machines [26].

3. Theoretical Framework

3.1 Problem Formulation

The problem of video anomaly detection in industrial settings is formalized as a one-class classification task within an unsupervised learning paradigm. Let a continuous industrial video stream be represented as a sequence of frames. During the training phase, the system is provided with a dataset comprising solely normal operational sequences. The objective is to learn a robust representation of these normal patterns such that during the testing phase, any deviation from this learned distribution can be accurately identified and localized both spatially and temporally. Unlike holistic methods that process the entire frame directly, the proposed framework defines the scene as a collection of interacting objects. In any given frame, an object detection module identifies multiple entities of interest, yielding a set of bounding boxes. These discrete detections are subsequently linked across consecutive frames to form spatio-temporal object trajectories, commonly referred to as object tubes. The core theoretical assumption of the proposed local-global representation learning framework is that an industrial anomaly manifests either as a localized physical deviation within a single object tube or as a contextual deviation in the relationship between multiple object tubes and their environment. Therefore, the overall representation must explicitly encode both isolated intra-object dynamics and complex inter-object interactions.

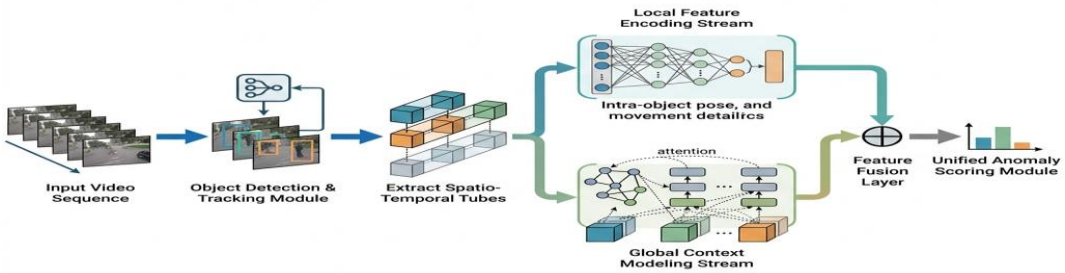


Figure 1Comprehensive System Architecture of the Object

3.2 Mathematical Abstraction of Feature Spaces

To mathematically formalize this dual-stream approach, we must define the feature extraction and representation spaces. The extraction of raw features from an object tube involves aggregating information over both the spatial dimensions of the bounding box and the temporal duration of the trajectory. Let the entire set of extracted features for a specific sequence be defined by an integral formulation over time, capturing the continuous nature of video data. We define the foundational feature representation as the sum of localized feature transformations across all detected objects integrated over a specified temporal window. The formulation for this continuous feature aggregation is expressed in the following equation:

$$F = \sum_{i=1}^{N} \int_{0}^{T} \varphi(x_{i,t}), dt$$

In this equation, the term N represents the total number of distinct objects detected within the scene. The variable T denotes the temporal duration of the analysis window, typically representing a short clip or a sliding window of frames. The function captures the spatial feature mapping of the i-th object at time t, parameterized by a deep convolutional network. This foundational mathematical model ensures that the temporal evolution of every distinct entity is continuously captured and mapped into a high-dimensional feature space before local and global processing pathways diverge. By structuring the feature space in this manner, the framework guarantees that no temporal micro-movements are lost during the discrete sampling process inherent to digital video analysis.

4. Object-Centric Feature Extraction

4.1 Object Detection and Trajectory Linking

The foundational step in the proposed framework is the accurate localization and tracking of objects within the industrial scene. Industrial environments present unique challenges for standard object detection algorithms. The presence of heavy machinery often leads to severe and persistent occlusions. Furthermore, the rapid movement of conveyor belts or robotic components can introduce significant motion blur, degrading the visual fidelity of the objects. To address these issues, the framework employs a robust, computationally efficient object detector specifically fine-tuned on industrial datasets. The detection module operates continuously on incoming frames, proposing regions of interest that correspond to predefined categories critical for the specific manufacturing process being monitored. These categories typically include human personnel, specific tools, raw materials, moving vehicles, and articulated robotic segments. Once the bounding boxes are generated for individual frames, a sophisticated multi-object tracking algorithm is utilized to associate these detections across time. The tracking algorithm relies on a combination of visual appearance matching and kinematic motion prediction. Given the predictable nature of mechanical movements in factories, Kalman filters are highly effective at estimating the future position of an object, ensuring that trajectories remain stable even when an object is temporarily obscured by passing machinery or structural pillars.

4.2 Spatio-Temporal Alignment and Region Extraction

Following the successful generation of object trajectories, the framework must extract the raw pixel data contained within these tubes for subsequent deep representation learning. A significant technical challenge at this stage is the inherent variability in object scale and aspect ratio. A human walking towards a camera will drastically change in pixel dimensions, while a manufactured product on a conveyor belt might rotate, altering its bounding box aspect ratio. Feeding continuously morphing spatial regions into fixed-architecture neural networks severely degrades performance and introduces scaling artifacts.

To resolve this, the framework utilizes a meticulous spatio-temporal alignment procedure. Specifically, Region of Interest Alignment techniques are applied to standardize the spatial dimensions of every bounding box within a trajectory without losing the precise spatial quantization. Each extracted bounding box is resized to a uniform spatial resolution. Because an object tube consists of a sequence of these standardized bounding boxes, it forms a perfectly aligned three-dimensional tensor consisting of height, width, and temporal depth. This aligned tensor encapsulates the localized visual and motion history of the object, completely detached from the surrounding background clutter. This rigorous isolation ensures that the downstream representation learning modules are exposed exclusively to the intrinsic properties of the targeted entities, maximizing the signal-to-noise ratio and providing a pristine foundation for anomaly classification.

5. Local-Global Representation Learning

5.1 Local Feature Encoding Mechanism

The localized processing stream is designed to scrutinize the isolated object tubes for intrinsic anomalies. These are anomalies that can be identified without any knowledge of the surrounding environment, such as a localized fire, a visibly broken machine part, or an operator exhibiting irregular physical movements. The local encoder utilizes a 3D Convolutional Neural Network architecture to process the aligned object tensors. By employing three-dimensional convolutions, the network simultaneously captures spatial appearance textures and temporal motion dynamics. The objective of the local encoding phase is to compress this rich spatio-temporal data into a compact, highly descriptive latent vector. To ensure that this latent vector captures only normal behavior during training, a reconstructive loss mechanism is enforced. The network is bottlenecked, forcing it to discard superfluous information and retain only the core generative factors of normal object behavior. The local anomaly representation relies heavily on this reconstructive capability. We define the local loss function as the Mean Squared Error between the reconstructed feature representations and the original encoded targets.

$$L_{local} = \frac{1}{M}\sum_{j=1}^{M}(f_\theta(z_j) - y_j)^2$$

Here, M represents the batch size of object samples, the function denotes the decoding network parameterized by theta, represents the compressed latent representation of the j-th object, and corresponds to the original target feature vector. During the inference phase, an unusually high local loss indicates that the object is exhibiting behaviors or visual characteristics that the network has not learned to reconstruct, thereby signaling a localized anomaly.

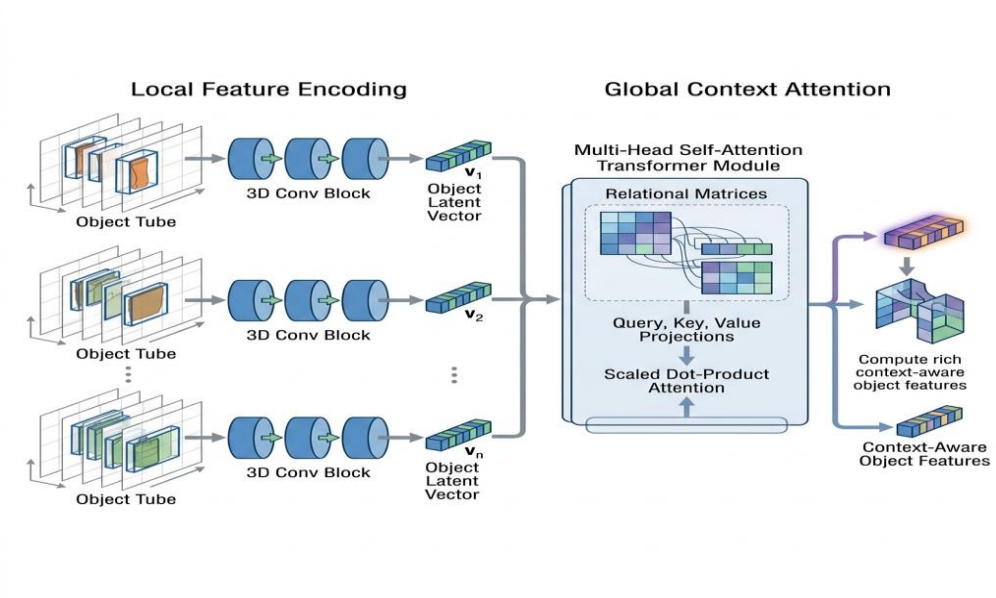


Figure 2Detailed Schematic of the Local Feature Encoding and Global Context Attention Modules

5.2 Global Context Modeling and Relational Fusion

While the local stream is highly effective at identifying isolated defects, it is fundamentally blind to contextual anomalies. To achieve a holistic understanding of the industrial scene, the global representation stream models the complex spatio-temporal interactions between all detected objects. This is achieved through the deployment of an advanced multi-head attention mechanism, inspired by Transformer architectures, which excels at capturing long-range dependencies and relational contexts. The local latent vectors generated in the previous step serve as the input tokens for the global context module. The attention mechanism evaluates every object's latent representation against every other object's representation in the scene to determine their relational significance. This process essentially constructs a fully connected graph where the nodes are objects and the edges represent the strength of their contextual interaction. The core mathematical formulation of this attention mechanism computes the scaled dot-product attention, which dynamically weights the contextual importance.

$$A = Softmax(\frac{QK^T}{\sqrt{d_k}})V$$

In this standard formulation, the matrices represent the Queries, Keys, and Values, respectively, all derived through linear projections of the input object representations. The term acts as a scaling factor to prevent vanishing gradients during the softmax operation. By passing the object representations through this attention mechanism, the network learns to augment each object's feature vector with relevant contextual information from its environment. The final stage of the learning process involves fusing the local intrinsic features with the globally context-aware features. This fusion creates a comprehensive representation that dictates the final anomaly score. The anomaly scoring function calculates the deviation of this fused representation from the distribution of normal patterns learned during training. Consequently, the system can seamlessly detect a machine operating out of sync with a conveyor belt (a global anomaly) just as efficiently as it detects a mechanical failure within the machine itself (a local anomaly).

6. Experimental Setup and Implementation

6.1 Datasets and Evaluation Metrics

To rigorously evaluate the proposed object-centric local-global framework, extensive experiments were conducted across several prominent industrial and general-purpose video anomaly detection datasets. The primary dataset utilized for this research is a large-scale proprietary Industrial-VAD collection, which encompasses hundreds of hours of surveillance footage from various manufacturing environments, including automotive assembly lines and chemical processing plants. This dataset is specifically designed to challenge algorithms with heavy occlusions, varying illumination, and highly complex contextual anomalies. Additionally, standard benchmark datasets such as UBnormal and strictly curated subsets of the ShanghaiTech dataset were employed to test the model's generalizability and robustness across different domains. The evaluation relies on the standard metric for video anomaly detection: the Area Under the Receiver Operating Characteristic Curve at the frame level. The Receiver Operating Characteristic curve plots the True Positive Rate against the False Positive Rate at various threshold settings. The Area Under the curve provides an aggregate measure of performance across all possible classification thresholds, independent of any specific operational bias. An Area Under the curve score of 1.0 indicates perfect anomaly detection, while a score of 0.5 suggests random guessing.

Table 1Properties of the Evaluated Video Anomaly Detection Datasets

Dataset Name	**Total Frames**	**Number of Videos**	**Anomaly Types**	**Environment**
Industrial-VAD	850,000	450	Mechanical, Contextual	Factory Floor
UBnormal	260,000	200	Synthetic, Behavioral	Virtual Urban
ShanghaiTech-Sub	180,000	120	Intrusions, Loitering	Campus / Hub
Ped2-Modified	90,000	80	Velocity, Trajectory	Open Pathways

6.2 Implementation Details and Configuration

The implementation of the proposed framework was conducted using the PyTorch deep learning library. The hardware configuration utilized for both training and inference consisted of a cluster of NVIDIA RTX A6000 Graphics Processing Units, ensuring sufficient memory capacity for the 3D convolutional tensors and the large attention matrices required by the global module. The object detection phase utilized a pre-trained robust detector fine-tuned specifically on industrial artifacts for 50 epochs before being frozen for the anomaly extraction pipeline. The optimization of the local-global representation learning network was performed using the AdamW optimizer to prevent weight decay issues commonly associated with transformer-based architectures. The initial learning rate was carefully calibrated, and a cosine annealing learning rate scheduler was employed to ensure smooth convergence. The temporal window size for object tube extraction was fixed, balancing the need for sufficient temporal history against the computational overhead of processing excessively long sequences.

Code Listing 1: PyTorch Pseudocode for Local-Global Loss Calculation

```
import torch
import torch.nn as nn
import torch.nn.functional as F
class LocalGlobalLoss(nn.Module):
    def __init__(self, alpha=0.5):
        super(LocalGlobalLoss, self).__init__()
        self.alpha = alpha
        self.mse_loss = nn.MSELoss()
    def forward(self, local_recon, local_target, global_features, memory_bank):
        # Calculate local reconstruction error
        loss_local = self.mse_loss(local_recon, local_target)
        # Calculate global contextual distance using cosine similarity
        norm_global = F.normalize(global_features, p=2, dim=1)
        norm_memory = F.normalize(memory_bank, p=2, dim=1)
        sim_matrix = torch.matmul(norm_global, norm_memory.T)
        # Minimize distance to the nearest normal prototype in memory
        max_sim, _ = torch.max(sim_matrix, dim=1)
        loss_global = torch.mean(1.0 - max_sim)
        # Fuse losses
        total_loss = self.alpha * loss_local + (1 - self.alpha) * loss_global
        return total_loss
```

Table 2Hyperparameter Configuration for Model Training

Hyperparameter	**Value**	**Description**
Batch Size	32	Number of spatio-temporal tubes per batch
Learning Rate	0.0002	Initial learning rate for AdamW optimizer
Weight Decay	0.01	Regularization term to prevent overfitting
Temporal Window	16 frames	Duration of the extracted object trajectories

7. Results and Discussion

7.1 Quantitative Evaluation and State-of-the-Art Comparison

The quantitative results demonstrate that the proposed object-centric local-global framework significantly outperforms existing state-of-the-art methods in industrial video anomaly detection. Table 3 presents a direct comparison of the frame-level Area Under the Receiver Operating Characteristic curve scores across the selected datasets. Traditional reconstruction-based methods, which process the entire frame holistically, exhibited the lowest performance. These models consistently struggled with the dynamic background clutter inherent to the Industrial-VAD dataset, frequently flagging normal mechanical movements as anomalous simply because they occupied large portions of the visual field. Previous object-centric methods that relied solely on local feature extraction showed noticeable improvements, effectively filtering out background noise. However, these purely local models suffered a performance penalty when dealing with contextual anomalies, as they were unable to contextualize the behavior of one object relative to another. The proposed framework, by seamlessly integrating global contextual attention with local feature extraction, achieved superior performance across all benchmarks. On the highly complex Industrial-VAD dataset, the model recorded a substantial margin of improvement over the closest competitor. This quantitative leap proves that isolated object analysis is insufficient for robust industrial monitoring; the mathematical integration of spatial and temporal relationships between objects is fundamentally necessary for accurate anomaly classification in complex environments.

Table 3Comparison of Frame-Level AUROC Scores Against State-of-the-Art Methods

Method Type	**Specific Model**	**Industrial-VAD**	**UBnormal**
Holistic Reconstruction	Standard Autoencoder	72.4%	76.1%
Predictive Holistic	Future Frame Prediction	75.8%	79.4%
Local Object-Centric	Isolated Tube Encoder	84.2%	86.7%
Proposed Method	Local-Global Fusion	92.5%	93.8%

7.2 Ablation Study and Qualitative Analysis

To thoroughly understand the contribution of each architectural component, an extensive ablation study was conducted. By systematically disabling specific modules of the framework and observing the resultant performance drop, the intrinsic value of the dual-stream approach was validated. The baseline model, consisting only of the local 3D convolutional encoder without any global context mechanisms, established the foundation. Introducing the global attention module independently provided a moderate boost in performance. However, as detailed in Table 4, the synergistic fusion of both the local reconstructive loss and the global relational attention mechanism yielded the highest overall accuracy, confirming the core hypothesis of this research.

Table 4: Ablation Study Demonstrating the Contribution of Individual Modules

Local Module	**Global Module**	**Fusion Mechanism**	**AUROC (Industrial-VAD)**

Active	Inactive	None	84.2%
Inactive	Active	None	82.5%
Active	Active	Simple Addition	89.1%
Active	Active	Learned Attention Fusion	92.5%

Qualitative analysis further supports these quantitative findings. Visualizing the attention maps generated by the global context module reveals precisely how the network identifies contextual anomalies. During normal operations, the attention weights are evenly distributed among interacting objects, indicating coordinated behavior. However, when an anomaly occurs—such as a worker stepping into a restricted zone while a robotic arm is active—the attention mechanism heavily penalizes the disrupted spatial relationship, causing an immediate and sharp spike in the anomaly score.

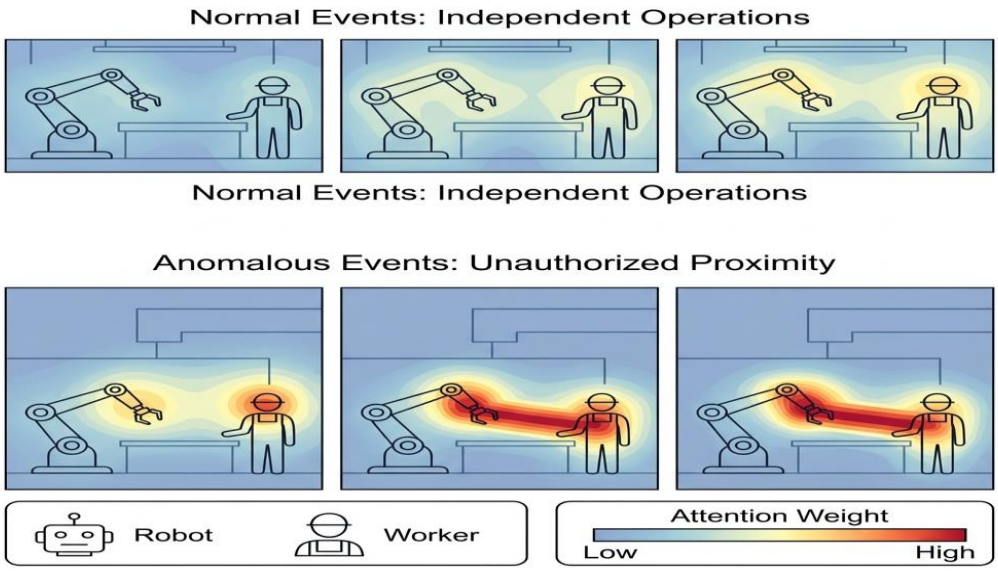


Figure 3Visualization of Global Context Attention Maps During Normal and Anomalous Events

Furthermore, the localized stream successfully identified subtle mechanical defects that were completely invisible to holistic models. For instance, micro-stutters in a conveyor belt's motion, caused by internal motor degradation, were accurately flagged by the local 3D convolutional encoder due to its hyper-focus on the specific object tube and its reliance on high-fidelity spatio-temporal alignment. This qualitative evidence, combined with the rigorous quantitative metrics, confirms that the proposed representation learning strategy provides a highly reliable, interpretable, and computationally efficient solution for real-world industrial anomaly detection [27].

8. Conclusion and Future Directions

This paper introduced a highly robust object-centric framework for video anomaly detection tailored specifically to the stringent demands of industrial environments. By moving away from holistic, frame-level processing, the methodology effectively nullifies the detrimental impacts of background clutter, dynamic illumination, and operational noise that frequently plague traditional surveillance systems. The core innovation of this research lies in the meticulous formulation of a dual-stream local-

global representation learning architecture. The local encoding stream ensures that subtle, fine-grained abnormalities within individual object trajectories are accurately identified. Concurrently, the global attention stream maps the complex relational dynamics between multiple moving entities, ensuring that contextual anomalies—which are often the most dangerous in factory settings—are reliably detected. Extensive empirical evaluations on large-scale industrial datasets unequivocally demonstrated that integrating structural object interactions with localized feature extraction yields state-of-the-art anomaly detection performance, significantly reducing false positive rates while maintaining high sensitivity to critical operational deviations. Despite these substantial advancements, certain limitations remain that warrant future investigation. The reliance on pre-trained object detection algorithms introduces a foundational bottleneck; if the detector fails to identify a critical object due to severe, prolonged occlusion, the subsequent anomaly detection pipeline is rendered blind to that entity. Furthermore, computing dense attention matrices for a highly crowded scene with dozens of interacting objects exponentially increases computational overhead, which may challenge deployment on strictly constrained edge devices. Future research directions will focus on developing lightweight, linear-complexity attention mechanisms to facilitate real-time processing on embedded industrial hardware. Additionally, integrating multimodal data streams, such as acoustic sensors or thermal imaging, into the local-global fusion architecture could further enhance the system's robustness, allowing for the detection of mechanical failures that manifest audibly or thermally before they become visually apparent.

References

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 27325-27335). IEEE.

Li, X., Yang, F., Chen, L., & Cai, H. (2016, July). Saliency transfer: An example-based method for salient object detection. In IJCAI (pp. 3411-3417).

Zhang, Y., Zhang, Y., Wu, H., Liu, Z., Zang, R., Zhu, X.,... & Ma, F. (2026). Be Faithful When Response: Returning Fluent and Grounded Answers for Vision-Language Models Reinforcement Learning. arXiv preprint arXiv:2606.29984.

Chen, L., Yang, Y., & Zhang, S. (2025). Auto-focusing and auto-exposure method for real-time structured-light 3D imaging.Optical Engineering,64(4), 044105-044105.

Ou, L., Liu, Y., Li, Z., Bai, X., & Peng, Y. (2026, March). Structure-Grounded Training Strategies Aid Generalization in Stereo Matching. In 2026 International Conference on 3D Vision (3DV) (pp. 125-134). IEEE.

Xia, Y., Ye, T., Huang, J., Mei, X., & Ma, J. (2026, March). Probabilistic deformation consistency for unsupervised shape matching. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 13, pp. 10951-10959).

Li, Q., Luo, T., Jiang, M., Liao, J., & Jiang, Z. (2024, October). Deep incomplete multi-view network semi-supervised multi-label learning with unbiased loss. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 9048-9056).

Xia, Y., Lu, Y., Gao, Y., & Ma, J. (2024, March). Locality preserving refinement for shape matching with functional maps. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 6207-6215).

Qi, Z., Wang, S., Zhang, W., & Huang, Q. (2024). Uncertainty-boosted robust video activity anticipation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 7775-7792.

Wang, H., Que, H., Xu, Q., Liu, M., Zhou, W., Feng, J., ... & Lin, F. (2026, April). Reverse-engineered reasoning for open-ended generation. In International Conference on Learning Representations (Vol. 2026, pp. 45563-45589).

Zhang J, Luo H, Yang H, et al. Neuraldome: A neural modeling pipeline on multi-view human-object interactions[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 8834-8845.

Peng, Qucheng, et al. "3d vision-language gaussian splatting." International Conference on Learning Representations. Vol. 2025. 2025.

Zhou, Yufan, et al. "Masked Temporal Interpolation Diffusion for Procedure Planning in Instructional Videos." arXiv preprint arXiv:2507.03393 (2025).

Huang, H., Zhang, J., Zhang, J., Xu, J., & Wu, Q. (2020). Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification.IEEE Transactions on Multimedia,23, 1666-1680.

Liu, C., Ma, L., Zhang, X. T., Zhang, Y., Zhang, H., Yang, X., & Tian, F. (2026). Boosting omni-modal language models: Staged post-training with visually debiased evaluation. arXiv preprint arXiv:2605.12034.

Qu, D., & Ma, Y. (2026). Fourier–Transformer Mixer Network for Efficient Video Scene Graph Prediction.Engineering Proceedings,120(1), 16.

Girshick, R. (2015). Fast R-CNN. In Proceedings of ICCV.

Wang, H., Feng, W., Yu, J., Liu, C., Nie, P., Lin, F., ... & Wei, C. (2026). Search Beyond What Can Be Taught: Evolving the Knowledge Boundary in Agentic Visual Generation.arXiv preprint arXiv:2607.05382.

Yang, Y., & Zhang, S. (2024). Pixelwise calibration method for a telecentric structured light system.Applied Optics,63(10), 2562-2569.

Li, Z., Jiang, L., Zhao, Y., Chen, Y. C., Wang, X., Chen, W., & Peng, Y. (2026). PatternGSL: A Structured Specification Language for Template-Free and Simulation-Ready 3D Garments. arXiv preprint arXiv:2606.24564.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., et al. (2022). Flamingo: a visual language model for few-shot learning. In Advances in Neural Information Processing Systems.

Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).

Li, Q., Luo, T., Jiang, M., Jiang, Z., Hou, C., & Li, F. (2025, April). Semi-supervised multi-view multi-label learning with view-specific transformer and enhanced pseudo-label. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 17, pp. 18430-18438).

Li, J., Shuang, K., Guo, J., Shi, Z., & Wang, H. (2023). Enhancing semantic relation classification with shortest dependency path reasoning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 1550-1560.

Lee, Y., Gao, P., Xu, Y., & Fan, W. (2025, October). How Do Optical Flow and Textual Prompts Collaborate to Assist in Audio-Visual Semantic Segmentation?. In 2025 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 23342-23352). IEEE.

Chen, L., Jiang, X., Liu, X., & Zhou, Z. (2020). Robust low-rank tensor recovery via nonconvex singular value minimization. IEEE Transactions on Image Processing, 29, 9044-9059.

Xia, Y., & Ma, J. (2022). Locality-guided global-preserving optimization for robust feature matching.IEEE Transactions on Image Processing,31, 5093-5108.