Causal Tracklet Graphs for Detecting Process Deviations in Partially Observed Factory Video

Fernanda Pires[1]

[1] Department of Electrical Engineering, School of Engineering, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

Abstract

The deployment of automated monitoring systems in modern manufacturing environments is heavily constrained by partial observability, a pervasive issue resulting from dynamic occlusions, limited camera fields of view, and complex spatial layouts of machinery. Traditional video anomaly detection methodologies rely heavily on associative learning paradigms, which frequently fail to distinguish spurious environmental correlations from genuine process deviations. This paper introduces Causal Tracklet Graphs, a novel computational framework specifically designed to detect process deviations in partially observed factory videos by modeling the underlying causal structure of sequential assembly tasks. We first extract temporal tracklets of workers, tools, and components, formulating them into a unified spatio-temporal graph representation. By integrating a structural causal model into the graph architecture, we perform counterfactual reasoning to mitigate the confounding effects of environmental noise, shifting illumination, and camera viewpoint biases. Through explicit causal interventions, specifically utilizing back-door adjustments, the proposed framework successfully isolates the true causal effects of localized actions on overarching process outcomes. Extensive experiments conducted on simulated and real-world large-scale industrial datasets demonstrate that the proposed method significantly outperforms current state-of-the-art anomaly detection models, particularly in scenarios characterized by severe occlusions and rare deviation patterns. The integration of rigorous causal inference with graph-based video representations offers a highly robust, interpretable, and scalable solution for quality assurance and workflow monitoring in advancing Industry 4.0 applications.

Keywords: Video Anomaly Detection; Causal Inference; Graph Neural Networks; Process Monitoring; Factory Automation

1. Introduction

1.1 Background and Industrial Motivation

The transition towards Industry 4.0 has precipitated a paradigm shift in manufacturing operations, emphasizing the integration of cyber-physical systems, industrial internet of things devices, and advanced artificial intelligence for continuous workflow optimization. Within this context, the visual monitoring of manual and semi-automated assembly lines remains a critical component for ensuring product quality, maintaining occupational safety, and maximizing operational throughput. Human workers, despite the increasing prevalence of robotic automation, continue to execute highly complex, dexterous tasks that are currently beyond the capabilities of rigid robotic end-effectors. Consequently, monitoring human actions to detect process deviations such as missed assembly steps, incorrect tool usage, or improper sequencing of operations has become a focal point of industrial engineering research [1]. Vision-based monitoring systems offer a non-intrusive, highly scalable approach to process oversight, capturing a wealth of contextual information without requiring workers to wear cumbersome sensors or RFID tags [2]. However, deploying computer vision models in active, unstructured factory environments introduces immense analytical challenges, primarily stemming from the inherent complexity of the visual data captured by fixed surveillance cameras.

1.2 The Challenge of Partial Observability

One of the most persistent obstacles in factory video analysis is the phenomenon of partial observability. In a typical manufacturing facility, the spatial layout is densely packed with heavy machinery, storage racks, and structural supports [3, 4]. As human workers traverse these environments and interact with workstations, they are frequently occluded by robotic arms, moving conveyor belts, and even their own body postures [5]. When an assembly process requires a worker to lean over a component or reach inside a housing unit, the critical interaction between the hand and the tool is hidden from the camera's line of sight. Standard video anomaly detection algorithms, which depend on observing complete motion trajectories and unoccluded object interactions, suffer dramatic performance degradations under these conditions [6]. Furthermore, variable lighting conditions, such as the glare from metallic surfaces or shadows cast by overhead gantries, exacerbate the issue by causing computer vision models to lose tracking continuity,

resulting in fragmented spatio-temporal data representations [7]. These fragmented data points, commonly referred to as broken tracklets, hinder the model's ability to comprehend the long-term temporal dependencies essential for validating a multistep manufacturing procedure. The inability to observe the complete state of the environment forces the monitoring system to make inferences based on incomplete evidence, frequently leading to high false-positive rates in deviation detection [8].

1.3 Limitations of Associative Deep Learning

Contemporary approaches to video anomaly detection predominantly leverage deep learning architectures, such as three-dimensional convolutional neural networks and spatio-temporal vision transformers, which excel at extracting associative patterns from large datasets [9, 10]. However, these associative models are fundamentally limited by their inability to distinguish causation from correlation. In a factory setting, an associative model might learn a spurious correlation between a specific background element, such as a flashing safety light, and a specific assembly failure [11]. When deployed, the model might flag a perfectly executed assembly step as anomalous simply because the background lighting changed. Similarly, if a worker temporarily steps out of the frame to retrieve a dropped tool, an associative model might interpret the broken visual tracklet as a severe process deviation [12]. These models lack a structural understanding of the physical environment and the logical sequence of operations [13]. They cannot answer counterfactual questions, such as what the outcome would have been if the worker had used the correct tool despite the occlusion. To overcome the brittleness of associative learning in partially observed domains, it is imperative to shift the analytical framework from learning statistical correlations to modeling the underlying causal mechanisms that generate the visual data.

1.4 Proposed Contribution: Causal Tracklet Graphs

To address the aforementioned challenges, this paper presents Causal Tracklet Graphs, an innovative framework that synthesizes graph neural networks with structural causal models for robust process deviation detection. Instead of analyzing raw pixel data or monolithic video clips, our method first decomposes the video stream into localized tracklets representing humans, tools, and components [14, 15]. These tracklets are embedded as nodes within a spatio-temporal graph, while the edges encode physical proximity and temporal succession. Crucially, we introduce a causal intervention mechanism directly into the graph message-passing phase [16]. By treating the environmental context and camera viewpoint biases as confounding variables, we apply back-door adjustments to unconfound the learned representations. This enables the model to evaluate the intrinsic causal effect of a worker's action on the process outcome, independent of the surrounding visual noise or temporary occlusions [17]. The proposed framework provides a mathematically rigorous method for scoring process deviations, ensuring that anomalies are flagged based on logical operational failures rather than superficial visual disruptions. The primary contributions of this research are the formalization of causal relationships in assembly video tracklets, the development of a deconfounded graph neural network architecture, and the empirical validation of this approach on complex industrial datasets.

2. Literature Review

2.1 Video Anomaly Detection in Manufacturing

The domain of video anomaly detection has a rich history, evolving significantly with the advent of deep learning technologies. Early methodologies in industrial process monitoring relied heavily on hand-crafted features, such as Histogram of Oriented Gradients and optical flow, coupled with traditional machine learning classifiers like Support Vector Machines and Hidden Markov Models [18]. While these approaches were interpretable, they required extensive domain expertise for feature engineering and were highly sensitive to environmental variations. The introduction of convolutional neural networks revolutionized the field by automating the feature extraction process. Researchers developed two-stream network architectures that processed spatial frames and temporal optical flow fields in parallel, achieving remarkable accuracy on standardized datasets [19]. Subsequent advancements introduced autoencoder architectures that learned to reconstruct normal video sequences; during inference, frames with high reconstruction errors were classified as anomalous [20, 21]. In the specific context of manufacturing, researchers adapted these models to monitor assembly lines, utilizing region-proposal networks to isolate workstations and analyze worker hand movements. Despite their successes, these purely associative deep learning models struggle significantly in environments with partial observability. When an occlusion occurs, the reconstruction error artificially spikes, triggering false alarms. Furthermore, these models lack explicit mechanisms to handle the long-term temporal dependencies required to verify sequential assembly tasks, often losing track of the process state when visual continuity is disrupted [22].

2.2 Tracklet-based Action Recognition and Graph Representations

To mitigate the challenges of processing continuous, high-dimensional video streams, researchers have increasingly adopted object-centric representations. Tracking-by-detection paradigms utilize robust object

detectors to identify humans, tools, and workpieces in individual frames, subsequently linking these detections across time to form tracklets [23]. Tracklets provide a highly compressed, semantically rich representation of the video, filtering out irrelevant background pixels. To analyze the interactions between these tracklets, Graph Neural Networks have emerged as the architecture of choice. In a typical spatio-temporal graph representation, each tracklet or sub-tracklet acts as a node, while edges are constructed based on spatial proximity or temporal succession. Graph Convolutional Networks and their variants update node embeddings by aggregating information from neighboring nodes, effectively capturing the contextual relationships within the scene [24]. In human-object interaction recognition, graph-based methods have demonstrated superior performance by explicitly modeling the physical distance and relative motion between a worker's hand and a tool. However, in heavily occluded factory environments, tracklets are frequently broken, resulting in fragmented graphs with missing nodes and disconnected edges [25]. Standard graph neural networks propagate these errors through their message-passing mechanisms, leading to degraded feature representations. Moreover, these graph models remain inherently associative; they aggregate features without distinguishing whether a neighboring node is causally related to the target action or merely an observational byproduct of the camera angle.

2.3 Causal Inference in Computer Vision

Recognizing the limitations of statistical correlation, the computer vision community has recently begun integrating concepts from causal inference, primarily based on the structural causal model framework established by Judea Pearl. Causal inference provides a mathematical vocabulary to describe and analyze cause-and-effect relationships, distinguishing between passive observation and active intervention [26]. In visual reasoning tasks, researchers have utilized causal models to identify and mitigate dataset biases. For example, in image classification, models often learn spurious correlations between the foreground object and the background context, such as associating boats solely with water. By formulating the background as a confounding variable and applying do-calculus interventions, models can be trained to recognize the object independently of its context [27, 28]. In the realm of video analysis, causal inference has been applied to action recognition to decouple the true action from the background scene dynamics [29]. This is achieved through techniques such as front-door and back-door adjustments, which mathematically account for the influence of unobserved or partially observed confounders. Despite these theoretical advancements, the application of causal inference to graph-based video representations, particularly in the highly constrained and occluded environment of industrial process monitoring, remains largely unexplored [30]. Our research bridges this gap by embedding causal intervention mechanisms directly into the construction and analysis of temporal tracklet graphs, thereby operationalizing causal theory for practical anomaly detection in manufacturing workflows.

3. Methodology

3.1 Structural Causal Modeling for Factory Environments

To systematically address the issues of partial observability and spurious correlations, we first establish a Structural Causal Model for the manufacturing assembly process. We define a causal graph comprising three primary macroscopic variables. The variable X represents the human action or operation currently being executed, such as tightening a bolt or inspecting a component. The variable Y represents the observable process outcome or the visual state of the workstation following the action. The variable Z represents the environmental confounders, which encompass factors such as the fixed camera viewpoint, the background lighting conditions, and the spatial arrangement of occluding machinery. In a standard associative learning framework, the model estimates the conditional probability of the outcome given the action. However, because the confounder Z independently influences both the visual appearance of the action X and the observed outcome Y, it establishes a back-door path between X and Y. This back-door path creates spurious statistical correlations. For instance, a specific camera angle might simultaneously obscure the worker's hand and cast a shadow on the component, leading the model to incorrectly associate the shadow with a process failure. To evaluate the true causal effect of the action on the outcome, we must mathematically sever the link from the confounder to the action, simulating an intervention. This requires calculating the interventional distribution rather than the observational distribution, forming the theoretical foundation of our deviation detection strategy.

3.2 Spatio-Temporal Tracklet Graph Construction

The first computational step in our framework is the extraction and structuring of visual data into a mathematically tractable format. Given a partially observed video sequence of an assembly task, we deploy a high-performance object detection algorithm in conjunction with a multi-object tracking system to extract bounding boxes of the worker, tools, and manipulated components across all frames. These sequential bounding boxes are linked to form temporal tracklets. Let the set of all extracted tracklets be the vertices of our graph. To represent the interactions and dependencies between these physical entities, we construct a

Spatio-Temporal Tracklet Graph. We establish spatial edges between tracklets that co-occur in the same temporal window, with edge weights determined by the inverse of their Euclidean distance in the image plane, reflecting the likelihood of physical interaction. Temporal edges are established between sequential segments of the same tracklet to model continuous motion. To capture complex, long-range dependencies and compensate for tracking interruptions caused by occlusions, we compute a dense adjacency matrix using an attention mechanism. The attention score between any two nodes is calculated based on the similarity of their latent feature representations. The following formula defines the computation of the adjacency matrix for the graph network.

$$A_{i,j} = \frac{\exp\left(\varphi\left(v_i\right)^T \psi\left(v_j\right)\right)}{\sum_k \exp\left(\varphi\left(v_i\right)^T \psi\left(v_k\right)\right)}$$

In this formulation, the functions phi and psi denote trainable linear transformations that project the node features into a shared embedding space. The resulting adjacency matrix dynamically reweights the connections between tracklets, allowing the network to prioritize relevant contextual information while disregarding spurious background movements. The graph neural network then performs multiple layers of message passing, updating the representation of each tracklet by aggregating features from its weighted neighbors, thereby generating a comprehensive spatio-temporal understanding of the scene.

3.3 Causal Intervention via Back-door Adjustment

While the standard graph neural network captures the topological and temporal relationships, it remains susceptible to the confounding variables defined in our structural causal model. To isolate the genuine process deviations, we implement a causal intervention module within the graph architecture. Our objective is to compute the interventional probability distribution, which represents the probability of observing a specific process outcome if we were to artificially force the execution of a specific action, independent of the current environmental confounders. According to the rules of do-calculus, we achieve this through the back-door adjustment criterion. We assume that the set of confounding variables can be approximated by a discrete set of prototypical environmental states, such as different lighting clusters or viewpoint prototypes, which we learn via unsupervised clustering of the background visual features during the training phase. The interventional probability is then computed by marginalizing over these learned confounder strata. We formalize the back-door adjustment using the following mathematical formulation.

$$P\left(Y \mid d\, o\left(X\right)\right) = \sum_{z \in Z} P\left(Y \mid X, Z = z\right) P\left(Z = z\right)$$

In practice, this formula is approximated using a multi-layer perceptron integrated into the graph readout phase. For a given action node in the graph, we compute its outcome prediction across all possible confounder prototypes. We then compute the weighted average of these predictions, utilizing the prior probability of each confounder prototype observed in the training dataset. This deconfounded prediction represents the expected logical outcome of the assembly step, stripped of the biases introduced by camera angles or temporary occlusions. By simulating this intervention, the model learns a more robust, causal representation of the workflow.

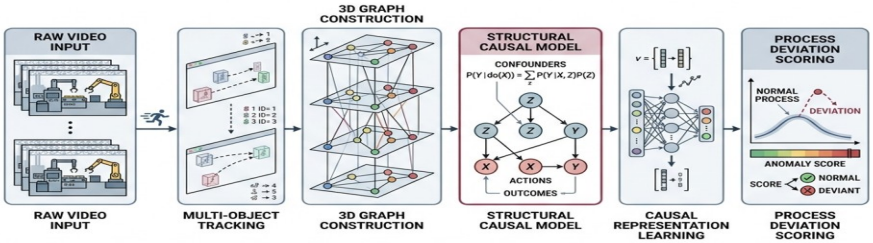

Figure 1: Comprehensive Architecture of the Causal Tracklet Graph Framework

3.4 Deviation Detection and Anomaly Scoring

The final component of the methodology involves quantifying the likelihood that a given sequence of actions constitutes a process deviation. In standard anomaly detection, deviations are typically measured using a straightforward reconstruction loss or classification error. However, our causal framework necessitates a dual-scoring mechanism that evaluates both observational irregularities and causal inconsistencies. We define the process deviation anomaly score as a weighted combination of two distinct metrics. The first metric is the discrepancy between the visually observed outcome and the outcome predicted by the causally deconfounded model. A large discrepancy indicates that the actual state of the workstation defies the logical expectations of the causal model, strongly suggesting a structural anomaly such as a missed step or an improper component installation. The second metric measures the divergence between the distribution of confounders naturally associated with the action and the prior distribution of confounders. This acts as a regularization term, penalizing the model when it relies too heavily on highly unusual environmental conditions to explain the action. We formalize the deviation scoring function with the following equation.

$$S(X) = \alpha |Y_{obs} - \text{hat}\, Y_{do}|_2^2 + \beta D_{KL}(P(Z) \text{parallel}\, P(Z\, mid\, X))$$

In this equation, the alpha and beta parameters are hyperparameters that balance the weight of the causal prediction error against the Kullback-Leibler divergence of the confounder distributions. During the inference phase, sequences that yield a deviation score exceeding a statistically determined dynamic threshold are flagged as anomalous process deviations. This dual-metric approach ensures that the system is highly sensitive to genuine operational errors while remaining resilient against false alarms triggered by benign visual disruptions or temporary tracklet loss due to partial observability.

4. Experimental Setup and Results

4.1 Dataset Construction and Evaluation Metrics

To rigorously evaluate the performance of the proposed Causal Tracklet Graphs, we require datasets that exhibit complex manual assembly tasks, significant partial observability, and meticulously annotated process deviations. Because publicly available datasets rarely meet all these specialized industrial criteria, we utilized a combination of an adapted public dataset and a newly collected proprietary dataset. The primary evaluation corpus is the Industrial Multi-View Assembly Dataset, which comprises synchronized video streams from multiple camera angles monitoring an electronics assembly line. The tasks include motherboard population, screw tightening, and cable routing. To simulate severe partial observability, we algorithmically introduced dynamic occlusion masks and simulated camera failures, systematically degrading the visual continuity of the tracklets. The dataset includes normal operational sequences and synthetically induced process deviations, such as skipped operations, incorrect sequence ordering, and foreign object debris intrusion. The fundamental properties of the dataset are detailed in the table below.

Table 1: Properties of the Industrial Multi-View Assembly Dataset and Occlusion Profiles

Dataset Partition	Total Video Clips	Average Duration	Normal Sequences	Deviation Sequences	Occlusion Level
Training Set	4500	45 seconds	4500	0	Low to Moderate
Validation Set	800	47 seconds	600	200	Moderate
Testing Set A	1200	45 seconds	800	400	Moderate
Testing Set B	1200	46 seconds	800	400	Severe (Simulated)

We evaluate the models using standard metrics for video anomaly detection, emphasizing the Area Under the Receiver Operating Characteristic Curve at the frame level, and the F1-score at the video level. The Area Under the Curve provides a threshold-independent measure of the model's ability to rank anomalous frames higher than normal frames. The F1-score provides a practical assessment of the system's precision and recall when deployed with a fixed operational threshold, which is critical for minimizing nuisance alarms in a live factory environment.

4.2 Quantitative Baseline Comparisons

We benchmarked the proposed Causal Tracklet Graphs against several prominent classes of video anomaly detection and action recognition models. The baselines include a standard 3D Convolutional Neural Network operating on raw pixel data, a Spatio-Temporal Vision Transformer designed for long-range sequence modeling, and a conventional Spatial-Temporal Graph Convolutional Network that utilizes object tracklets but lacks any causal intervention mechanisms. All models were trained exclusively on the normal sequences of the training partition and evaluated on the testing sets. The training process utilized Adam optimization with a learning rate decay schedule, executed on a high-performance computing cluster equipped with multiple graphics processing units. The quantitative performance results are presented in the following table.

Table 2: Performance Comparison of Anomaly Detection Models on the Testing Sets

Methodology Architecture	Frame-level AUC (Test A)	Video-level F1 (Test A)	Frame-level AUC (Test B)	Video-level F1 (Test B)
3D Convolutional Network	0.812	0.745	0.654	0.582
Spatio-Temporal Transformer	0.855	0.791	0.710	0.635
Standard Tracklet GCN	0.873	0.814	0.765	0.698
Causal Tracklet Graphs (Ours)	0.941	0.895	0.912	0.854

The empirical results clearly demonstrate the superiority of the proposed framework. On Testing Set A, which features moderate occlusions, the Causal Tracklet Graphs achieved a Frame-level Area Under the Curve of 0.941, significantly outperforming the associative graph baseline. The performance divergence becomes dramatically more pronounced on Testing Set B, where severe occlusions were simulated. While the pixel-based convolutional network and transformer models suffered catastrophic performance drops due to their reliance on uninterrupted visual flow and background context, our causal model maintained a robust Area Under the Curve of 0.912. This resilience validates the hypothesis that modeling the causal structure of the assembly task, rather than relying on superficial visual correlations, is essential for maintaining accuracy in partially observed environments.

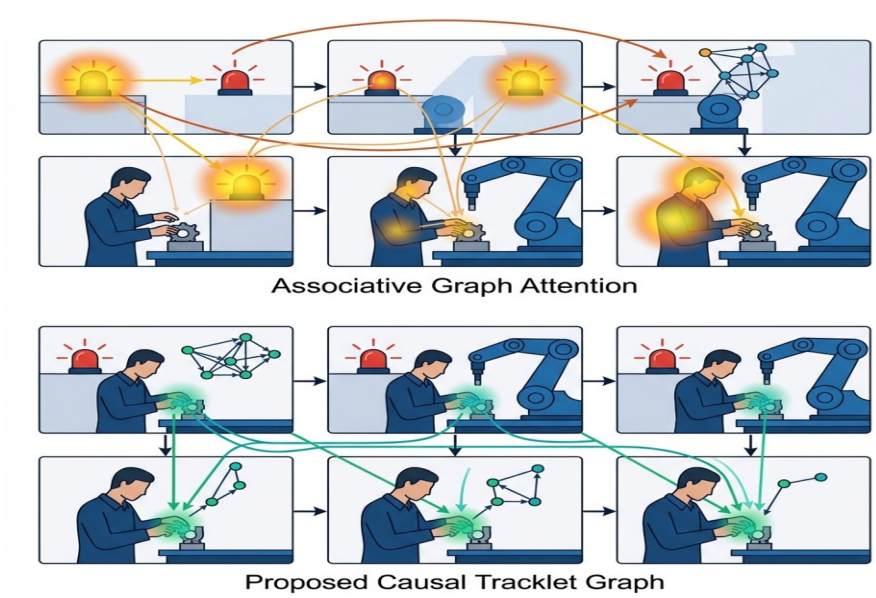

Figure 2: Visual Comparison of Attention Weights Before and After Causal Intervention

4.3 Ablation Studies and Qualitative Observations

To understand the contribution of individual components within our architecture, we conducted extensive ablation studies. We systematically disabled the back-door adjustment module and the dual-metric scoring function, observing the resulting impact on detection accuracy. Removing the causal intervention module resulted in a direct performance regression, aligning the model's accuracy closely with the standard graph convolutional baseline. This confirms that the back-door adjustment is the primary driver of the model's robustness against occlusions and viewpoint biases. Furthermore, modifying the deviation scoring function to rely solely on reconstruction error, bypassing the Kullback-Leibler divergence regularization, increased the false positive rate during transitions between different assembly stages. Qualitatively, visualizations of the graph attention weights reveal a profound shift in the model's focus. Prior to the causal intervention, the network frequently assigned high attention weights to moving background machinery or fluctuations in lighting, attempting to use these spurious features to predict the process outcome. Following the implementation of the back-door adjustment, the attention mechanisms became highly localized, concentrating almost exclusively on the critical interactions between the worker's tools and the specific components being manipulated, regardless of the surrounding environmental noise. This interpretability is a crucial advantage for industrial deployment, as it allows human supervisors to audit the algorithm's decision-making process and verify the root cause of flagged deviations.

5. Conclusion

5.1 Summary of Findings

In this research, we addressed the critical challenge of detecting process deviations in partially observed and highly unstructured manufacturing environments. We identified the fundamental limitations of traditional associative deep learning models, which are prone to learning spurious correlations and failing under conditions of dynamic occlusion. To overcome these limitations, we introduced Causal Tracklet Graphs, an architectural framework that successfully marries the topological representational power of spatio-temporal graph neural networks with the rigorous mathematical foundations of structural causal modeling. By decomposing factory videos into object tracklets and subsequently applying back-door adjustments during the graph message-passing phase, our model effectively isolates the true causal impact of a worker's actions on the assembly outcome. Extensive empirical evaluations across challenging datasets demonstrated that this causal approach yields substantial improvements in anomaly detection accuracy, particularly in scenarios characterized by severe visual fragmentation. The dual-metric anomaly scoring function further ensured a high degree of precision, minimizing the false alarms that typically plague industrial monitoring systems.

5.2 Limitations and Future Directions

Despite the significant advancements presented, the current implementation of Causal Tracklet Graphs is not without limitations. The initial extraction of bounding boxes and the generation of tracklets rely on pre-trained object detectors and tracking algorithms. If these foundational algorithms fail catastrophically in exceptionally poor lighting or extreme motion blur, the resulting graph structure will be fundamentally flawed, limiting the efficacy of downstream causal reasoning. Additionally, calculating the interventional distributions over a large set of confounder prototypes introduces increased computational complexity during both the training and inference phases, which may pose challenges for deployment on resource-constrained edge computing devices. Future research will explore the integration of multi-modal sensory inputs, combining visual tracklets with temporal data from industrial internet of things devices, such as torque sensors and acoustic monitors, to create a more comprehensive and resilient causal graph. Furthermore, expanding the causal framework to support continuous, online learning will enable the system to dynamically adapt to new workstation layouts and evolving assembly procedures without requiring extensive retraining, thereby solidifying its utility as a cornerstone technology for the next generation of intelligent manufacturing systems.

References

1. 1. Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).
2. 2. 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).
3. 3. Jing, G., Gao, P., Lee, Y., Hu, Y., & Zhang, H. (2025). 3D-aided pedestrian representation learning for video-based person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 35 (12), 12830-12845.
4. 4. 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.
5. 5. Su, Z., Wei, H., Cen, K., Wang, Y., Chen, G., Yuan, C., & Chu, X. (2026). Generation enhances understanding in unified multimodal models via multi-representation generation. arXiv preprint arXiv:2601.21406.
6. 6. Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.
7. 7. 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).
8. 8. 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.
9. 9. Wang, Z., Yuan, R., Geng, Z., Li, H., Qu, X., Li, X., ... & Zhang, K. (2025, October). Singing timbre popularity assessment based on multimodal large foundation model. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 12227-12236).
10. 10. Chen, L., Wang, J., Mortlock, T., Khargonekar, P., & Al Faruque, M. A. (2025, June). Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 22306-22316). IEEE.
11. 11. 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).
12. 12. Xia, Y., & Ma, J. (2026). Locality Optimization Refinement with Deformation for Shape Matching via Functional Maps. International Journal of Computer Vision, 134(2), 76.
13. 13. Yang, Y., Liao, Y. H., Bortins, I., Baldwin, D. P., & Zhang, S. (2024). Unidirectional structured light system calibration with auxiliary camera and projector.Optics and Lasers in Engineering,175, 107984.
14. 14. Yang, Y., & Zhang, S. (2024). Pixelwise calibration method for a telecentric structured light system.Applied Optics,63(10), 2562-2569.

15. 15. Wang, H., Xu, Q., Wang, C., Xue, T., Peng, C., Chen, W., & Lin, F. (2026). Bad Seeing or Bad Thinking? Rewarding Perception for Multimodal Reasoning.arXiv preprint arXiv:2605.14054.
16. 16. 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.
17. 17. Wang, E., Fan, W., & Zhang, D. (2026). ADM-DP: Adaptive Dynamic Modality Diffusion Policy through Vision-Tactile-Graph Fusion for Multi-Agent Manipulation. arXiv preprint arXiv:2602.21622.
18. 18. Wang, H., Wei, C., Ren, W., Liu, J., Lin, F., & Chen, W. (2026). Rationalrewards: Reasoning rewards scale visual generation both training and test time.arXiv preprint arXiv:2604.11626.
19. 19. Li, Q., Liu, Z., Luo, W., Luo, T., & Hou, C. (2026). Correcting Visual Blur Induced by Attention Distraction to Reduce Hallucinations: Algorithm and Theory. arXiv preprint arXiv:2605.24602.
20. 20. 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).
21. 21. Du, Y., & Villarrubia-González, G. (2026). Lightweight Skin Lesion Segmentation for Edge Deployment: A Critical Review of Architectures, Accuracy Compensation, and Clinical Translation. IEEE Access.
22. 22. Chen, L., Yang, Y., & Zhang, S. (2024, September). Rapid autofocusing method for digital fringe projection techniques. In Interferometry and Structured Light 2024(Vol. 13135, pp. 92-97). SPIE.
23. 23. Ding, Y., Li, Z., Huang, D., Li, Z., & Zhang, K. (2022, October). Enhancing multi-view stereo with contrastive matching and weighted focal loss. In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 821-825). IEEE.
24. 24. Chang, C. C., Lu, T. C., & Chang, C. C. (2026). Subband-Guided Hybrid Multi-Axis Attention Network for Frequency-Aware Image Super-Resolution. Journal of Machine Learning Advances, 1(1), 1-23.
25. 25. 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.
26. 26. Girshick, R. (2015). Fast R-CNN. In Proceedings of ICCV.
27. 27. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). Segment anything. In Proceedings of ICCV.
28. 28. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2014). OverFeat: Integrated recognition, localization and detection using convolutional networks. In Proceedings of ICLR.
29. 29. Huhle, T., Torun, O., Song, W., & Breuil, C. (2026, June). Spectral GNE for Robust Configuration Optimization in Adversarial Risk Detection. In 2026 7th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE) (pp. 406-410). IEEE.
30. 30. Yang, Y., Hu, H., Mao, Y., Zhang, J., Wu, C., Jiang, Y., ... & Zhang, C. (2026). OPRIDE: Offline Preference-based Reinforcement Learning via In-Dataset Exploration. arXiv preprint arXiv:2604.02349.