A Review of Occlusion-Aware Object Representations in Unsupervised Video Understanding

Natasha Sharan

June 12th, 2025.

1 Introduction

3D object tracking is a valuable mechanism that lends itself to various applications, ranging from medical imaging, to traffic surveillance systems. While supervised object tracking models have demonstrated high performance, they require large volumes of annotated data, the collection of which can be both costly and time-consuming. In contrast, unsupervised approaches eliminate the need for labelled datasets, allowing for greater scalability. However they struggle with persistent object identification, particularly under challenging visual conditions. Objects in video sequences may be blurred by motion, obscured by poor lighting, or occluded by other elements in the scene.

This review focuses on the specific challenge of object occlusion within the context of unsupervised 3D object tracking. In addition to summarizing recent efforts to address occlusion-related limitations, these approaches will be compared to a research proposal aimed at improving unsupervised occlusion-aware object representations. Section 2 will discuss related work, Section 3 will address gaps mentioned in Section 2, and compare it to the proposal, Section 4 will discuss the details of the proposal, and Section 5 will conclude the review.

2 Related Works

This section reviews key developments in occlusion-aware object detection and tracking (2.1), as well as additional studies that provide relevant context(2.2).

2.1 Key Papers

2.1.1 Dynamic GNNs for Occlusion-Aware Tracking

As previously mentioned, object detection is an essential tool utilized across a variety of systems, one such example being traffic surveillance. An important aspect of traffic surveillance involves the detection and tracking of small objects, such as pedestrians and cyclists. However, this feature has persistently suffered from limitations of real-time object tracking induced by challenging visual factors such as motion-blur and occlusion. Traditional object detection methods, Convolutional Neural Networks, as well as earlier versions of YOLO, are limited in their abilities to capture temporal relationships, which are crucial for object tracking across frames. Consequently, these implementations struggle with small objects that move erratically, and are frequently occluded. This is especially prevalent in complex urban traffic conditions, which can lead to unreliable detection and tracking, impacting traffic safety and management. In the paper, ”Interpretable Dynamic Graph Neural Networks for Small Occluded Object Detection and Tracking” by Shahriar Soudeep, Md Abrar Jahin, and M. F. Mridha [12], these challenges are addressed. The research presented in this paper, aims to overcome these limitations by proposing a system that can robustly handle dynamic conditions, thus advancing the capabilities of intelligent transportation systems.

The work introduces the DGNN-YOLO framework, which integrates the YOLO11 object detector with Dynamic Graph Neural Networks for robust tracking. The YOLO11 detection mechanism is utilized for its advanced spatial feature extraction capabilities, allowing for the precise initial detection of small objects. Similarly, the DGNN is employed to capture the evolution of object positions, and extract the temporal features of the scene. Fusing the spatial and temporal feature extraction ensures robust object tracking in cluttered traffic environments.

For each frame, a set of objects are detected by YOLO11, as well as their bounding boxes, the confidence score, the object category, and spatial features. This information is used to construct a dynamic graph, representing the objects and their interactions. The nodes correspond to the detected objects, containing a feature vector consisting of spatial, motion and appearance features. The edges model the relationships between objects and are updated based on the euclidean distance between bounding boxes (proximity), the difference in velocity of objects at a specific time (velocity similarity), and the cosine similarity between feature vectors (appearance similarity). The adjacency matrix encodes the strengths of these relationships. This graph is updated and recalculated dynamically at each frame, ensuring the information is consistently distributed across the nodes. This allows for the mechanism to capture both spatial and temporal dynamics, and ensuring the accuracy of object tracking across frames.

The DGNN-YOLO framework aims to balance the contributions of the detection and tracking tasks, by optimizing two primary loss functions. The detection loss is responsible for evaluating the object detection accuracy, and it is comprised of the sum of a bounding box loss and cross entropy loss. The objective of the tracking loss is to minimize inconsistencies in object identities across frames, which is given by the sum of the squared euclidean distances between nodes, and the cosine similarity of the feature embeddings. The total loss function is the weighted sum of the detection and tracking loss. The YOLO11 component of the model is pretrained, and is further trained and fine tuned on annotated data. Similarly, while the DGNN component learns dynamically, it also relies on ground truth labels, as is evident in the loss functions. Thus it can be noted that the DGNN-YOLO is a supervised model, and requires large amounts of annoted data.

In the evaluation stages, the DGNN-YOLO framework demonstrated significant improvements over existing models such as standard YOLO versions, and Faster R-CNN. It achieved superior results, with a precision of 0.8382, a recall of 0.6875, and a mean average prediction (from 0.5 to 0.95) of 0.6476 on the i2Object Detection Dataset. Without the DGNN component, the performance of YOLO11 was lower, with a precision of 0.8176, exhibiting the necessity of the DGNN and its temporal feature extraction. Similarly, without the spatial features, the performance exhibited a mild decline. These results highlight the importance of both spatial and temporal features for robust object tracking in complex environments. On the other hand, despite its strong performance, the DGNN-YOLO's reliance on visual data causes it to struggle in extreme weather conditions, and it has difficulty differentiating between visually similar or underrepresented classes. The authors have proposed to improve this by integrating sensors such as LiDAR for low-visibility, and exploring semi-supervised learning to improve the detection of rare classes, and reduce dependence on manually labelled data.

2.1.2 Semi-Supervised Temporal GNNs for 3D Tracking

[15]

2.1.3 GNNs with Cross-Edge Modality Attention

[1]

2.1.4 Iterative Scene Graph Generation

[6]

2.1.5 Context-Aware Compositional Nets

[13]

2.2 Additional Studies

Advances in unsupervised and self-supervised have allowed for refined object tracking methods, especially in approaching the issue of occlusion-handling. Similarly, dynamic 3D scene understanding can contribute to a more effective understanding of spatial-temporal relations. This section reviews developments in these areas.

2.2.1 Advances in Unsupervised and Self-Supervised Tracking

[5] [11] [16] [17] [18]

2.2.2 Occlusion Handling Mechanisms

[14] [9] [3] [20]

2.2.3 Dynamic 3D Scene Understanding

[2] [7] [4] [10] [19] [8]

3 Gap Analysis

4 Methodology

5 Conclusion

References

[1] Martin Buchner and Abhinav Valada. *3D Multi-Object Tracking Using Graph Neural Networks with Cross-Edge Modality Attention*. 2022. arXiv: 2203.10926 [cs.CV]. URL: https://arxiv.org/abs/2203.10926.

[2] Geonho Cha, Minsik Lee, and Songhwai Oh. "Unsupervised 3D Reconstruction Networks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.

[3] Xingping Dong et al. "Occlusion-Aware Real-Time Object Tracking". In: *IEEE Transactions on Multimedia* PP (Nov. 2016), pp. 1–1. DOI: 10.1109/TMM.2016.2631884.

[4] Vitor Guizilini and Fabio Ramos. "Unsupervised Feature Learning for 3D Scene Reconstruction with Occupancy Maps". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1 (2017). DOI: 10.1609/aaai.v31i1.11039. URL: https://ojs.aaai.org/index.php/AAAI/article/view/11039.

[5] Shyamgopal Karthik, Ameya Prabhu, and Vineet Gandhi. *Simple Unsupervised Multi-Object Tracking*. 2020. arXiv: 2006.02609 [cs.CV]. URL: https://arxiv.org/abs/2006.02609.

[6] Siddhesh Khandelwal and Leonid Sigal. *Iterative Scene Graph Generation*. 2022. arXiv: 2207.13440 [cs.CV]. URL: https://arxiv.org/abs/2207.13440.

[7] Kevin Lai, Liefeng Bo, and Dieter Fox. "Unsupervised feature learning for 3D scene labeling". In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. 2014, pp. 3050–3057. DOI: 10.1109/ICRA.2014.6907298.

[8] Jiaxu Liu et al. "U3DS3: Unsupervised 3D Semantic Scene Segmentation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2024, pp. 3759–3768.

[9] Qiankun Liu et al. *Online Multi-Object Tracking with Unsupervised Re-Identification Learning and Occlusion Estimation*. 2022. arXiv: 2201.01297 [cs.CV]. URL: https://arxiv.org/abs/2201.01297.

[10] Feixiang Lu et al. *Real-time 3D scene reconstruction with dynamically moving object using a single depth camera*. 2018. DOI: https://doi.org/10.1007/s00371-018-1540-8.

[11] Ioannis Papakis, Abhijit Sarkar, and Anuj Karpatne. *GCNNMatch: Graph Convolutional Neural Networks for Multi-Object Tracking via Sinkhorn Normalization*. 2021. arXiv: 2010.00067 [cs.CV]. URL: https://arxiv.org/abs/2010.00067.

[12] Shahriar Soudeep, Md Abrar Jahin, and M. F. Mridha. *Interpretable Dynamic Graph Neural Networks for Small Occluded Object Detection and Tracking*. 2025. arXiv: 2411.17251 [cs.CV]. URL: https://arxiv.org/abs/2411.17251.

[13] Angtian Wang et al. *Robust Object Detection under Occlusion with Context-Aware CompositionalNets*. 2020. arXiv: 2005.11643 [cs.CV]. URL: https://arxiv.org/abs/2005.11643.

[14] Guangming Wang et al. "Unsupervised Learning of Depth, Optical Flow and Pose With Occlusion From 3D Geometry". In: *IEEE Transactions on Intelligent Transportation Systems* 23.1 (2022), pp. 308–320. DOI: 10.1109/TITS.2020.3010418.

[15] Jianren Wang et al. *Semi-supervised 3D Object Detection via Temporal Graph Neural Networks*. 2023. arXiv: 2202.00182 [cs.CV]. URL: https://arxiv.org/abs/2202.00182.

[16] Ning Wang et al. *Unsupervised Deep Tracking*. 2019. arXiv: 1904.01828 [cs.CV]. URL: https://arxiv.org/abs/1904.01828.

[17] Qiangqiang Wu, Jia Wan, and Antoni B. Chan. *Progressive Unsupervised Learning for Visual Object Tracking*. 2021. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Wu_Progressive_Unsupervised_Learning_for_Visual_Object_Tracking_CVPR_2021_paper.html.

[18] Yaochen Xie, Zhao Xu, and Shuiwang Ji. *Self-Supervised Representation Learning via Latent Graph Prediction*. 2022. arXiv: 2202.08333 [cs.LG]. URL: https://arxiv.org/abs/2202.08333.

[19] Chenyangguang Zhang et al. *Open-Vocabulary Functional 3D Scene Graphs for Real-World Indoor Spaces*. 2025. arXiv: 2503.19199 [cs.CV]. URL: https://arxiv.org/abs/2503.19199.

[20] Yubo Zhang, Liying Zheng, and Qingming Huang. "Occlusion-related graph convolutional neural network for multi-object tracking". In: *Image and Vision Computing* 152 (2024), p. 105317. ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2024.105317. URL: https://www.sciencedirect.com/science/article/pii/S0262885624004220.