This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/ICCVDM66874.2025.11290539,
2025 6th International Conference on Computer Vision and Data Mining (ICCVDM)

Analyzing the Impact of Depth Features on Point Track Performance

1st Khadijah Alkandary *
Department of Electronic and Electrical Engineering
Brunel University London
London, UK
ORCID: https://orcid.org/0009-0000-0260-0817
Email: khadijah.alkandary@brunel.ac.uk

2nd Ahmet Serhat Yildiz
Department of Electronic and Electrical Engineering
Brunel University London
London, UK
ORCID: https://orcid.org/0000-0002-2957-7394
Email: ahmetserhat.yildiz@brunel.ac.uk

3rd Hongying Meng
Department of Electronic and Electrical Engineering
Brunel University London
London, UK
ORCID: https://orcid.org/0000-0002-8836-1382
Email: hongying.meng@brunel.ac.uk

***Abstract*—The multi-object tracking and segmentation task in urban traffic scenes in for improving autonomous driving, poses ongoing challenges from occlusions, lighting variations, and background noise interferences. We tackle the issue of identity switches by enhancing the existing PointTrack framework, by incorporating raw and monocularly estimated depth information into the color-offset tracking pipeline. By combining depth cues directly into the offset features, our approach strengthens geometric reasoning and leads to improved object association in cases of occlusions and reappearances. On the KITTI multi-object tracking and segmentation dataset, our method reduces identity switching by 21.11% compared to PointTrack baseline, showing increased robustness of tractlet association in challenging scenes. Overall, the approach evaluated notably reduces the occurrence of excessive ID switches, which are a major handicap in real, complicated settings. Numerically, our model performs better by having fewer ID switches while maintaining and in certain cases, enhancing the overall MOTSA score.**

Keywords—KITTI, LiDAR, PointTrack, MOT, RGB camera

I. Introduction

Multi-object tracking (MOT) is a building block in contemporary computer vision systems with applications in autonomous driving, surveillance, and human-computer interaction. The primary goal of MOT is to preserve consistent identities of objects detected in consecutive video frames despite occlusions, lighting variations, and cluttered backgrounds. As high-quality visual data and computational power become widely available, notable advances in tracking algorithms have been achieved. Yet, robustness in real-world scenarios continues to be a central challenge [1] [2] [3] [4].

The purpose of this research is to develop a systematic assessment of incorporating depth information into the PointTrack model for enhancing the performance of multi-object tracking in real environments. The motivation behind it stems from the drawback of traditional convolution-based segmentation approaches, which fail in separating object instances from background clutter efficiently, particularly in dynamic scenes. This background interference deteriorates tracking accuracy and reliability.

Most recent progress on MOT has been based on deep learning-based methods that utilize 2D visual cues alone, including bounding boxes, appearance embeddings, and optical flow. Although largely successful, these methods tend to fail on spatial ambiguities resulting from overlapping objects or dynamic scenes. Researchers have thus begun to integrate 3D information, e.g., depth or Light Detection and Ranging (LiDAR), to improve tracking performance and spatial perception [5], [6]. Most of the current solutions, however, either demand costly hardware equipment or are not scalable, rendering them unsuitable for large scale deployment.

To overcome these limitations, this article proposed starting with the PointTrack framework, which casts tracking as a point-matching problem between RGB images and segmentation masks [7]. In this research, a depth-augmented variant of PointTrack that incorporates depth cues natively into the tracking pipeline, is presented. Firstly, depth-aware offset vectors are developed by lifting 2D spatial coordinates to 3D space through estimated depth values. Secondly, leveraging Depth Anything V2 [8] [9]; a state-of-the-art monocular depth estimation network, to produce dense depth maps from RGB frames, without relying on physical depth sensors. Finally, RGB-D(Depth) feature fusion is conducted to enhance both spatial and appearance representations. The combination of these methods realizes better tracking accuracy and identity preservation on the KITTI benchmark [10], verifying the advantages of depth incorporation in point-based object tracking.

The author's contribution is to investigate two different approaches for incorporating depth information into the PointTrack framework with the intention of better preserving the models spatial and contextual relationships. To validate the efficacy of the proposed methods, thorough experiments were conducted on the KITTI dataset [11], which had realistic challenges like different object scales, motion, and occlusion in outdoor driving environments. The performance was measured using standard metrics such as MOTA (Multiple Object Tracking Accuracy), IDF1, precision, recall, and IDSW. This work offers a fresh perspective on improving robustness and accuracy in challenging environments with depth-aware tracking models.

The remainder of this paper is structured as follows: Section II discusses the related literature on multi-object tracking and segmentation, especially those using depth features. Section III

describes the applied benchmark dataset, namely the KITTI MOTS dataset, and illustrates some examples of RGB and depth images. Section IV describes the proposed approach, including how depth was integrated into the PointTrack framework, and defines the evaluation metrics used in this research, like Multi-Object Tracking and Segmentation Accuracy (MOTSA) and IDSW. Section V reports the experimental results and contains an in-depth discussion of the results. Finally, Section VI concludes the paper and provides directions for future research.

II. Related Works

MOT in urban environments is progressing through the development of point-based tracking, depth estimation, and sensor fusion. This section reviews past, and recent associated works, with an emphasis on the PointTrack framework, depth-aware improvements, and multi-modal fusion methods.

PointTrack [12] proposed a deep learning-based association framework with sparse points and pairwise feature differences, which allows for effective online MOT with real-time performance. PointTrack++ [6] builds on the former with enhanced feature quality and match robustness. Both models achieve acceptable trade-off between speed and accuracy and are suitable for practical applications such as autonomous driving. The authors implement advancements in the same trajectory by incorporating depth information for enhancing identity consistency.

Depth cues greatly improve tracking robustness during occlusions. Approaches such as M3SOT [13] utilize LiDAR for high-accuracy tracking, however, lightweight monocular solutions like Depth Anything [8] and CATNet [22] , provide efficient transformer-based depth estimation. These advancements enable RGB-only trackers to enjoy depth awareness without additional sensors. The research exploits this advantage by injecting raw or predicted depth into PointTrack's feature pipeline for improved spatial reasoning.

Recently, MOT has improved its stability under occlusion and visual similarity by adding depth cues. Khanchi et al. suggested a zero-shot method called DepthMOT that uses monocular depth estimation and adds a Hierarchical Alignment Score to make connections more reliable without needing task-specific training [14]. Sun et al. created ViewTrack, which uses bounding box area and position to figure out relative depth relationships and make a view-adaptive association strategy [15]. Han et al. made GRASPTrack, which uses monocular depth and segmentation to make 3D point clouds and figure out voxel-based 3D IoU and adapt motion filtering when there are occlusions [16]. These works together show how important it is to include depth information, whether it's through explicit 3D geometry or lightweight pseudo-depth cues, to improve MOT performance.

FusionTrack [17] and Emberds [18] illustrate the value of fusing camera and LiDAR data through multi-stage association pipelines in eliminating ID switches. Graph-based approaches such as MCTrack and ReST [19] also exhibit better temporal consistency in multi-camera settings. Such systems, however, are demanding in terms of hardware. The proposed solution, tested on the KITTI MOTS dataset [10], offers a light and versatile alternative, with robust tracking performance, using optional depth augmentation.

III. Benchmark Dataset

The KITTI MOTS dataset [20] is a seminal and highly popular benchmark for Multi-Object Tracking and Segmentation evaluation in autonomous driving research. Proposed by Voigtlaender et al. [20], it builds upon the original KITTI tracking dataset by adding pixel-level instance annotations, with particular emphasis on the car class. The dataset consists of 21 training and 29 testing sequences, recorded with high-resolution stereo RGB cameras and a Velodyne 3D LiDAR scanner. As an alternative to conventional bounding-box annotations, KITTI MOTS uses instance masks to better deal with overlapping objects and occlusions. This makes it particularly valuable for benchmarking state-of-the-art tracking systems that demand fine-grained spatial precision and robust identity preservation in urban driving environments. The dataset includes 8008 frames from 21 videos.

IV. Methodology

In the original PointTrack [7] framework, the model takes two modalities as input: RGB images and segmentation maps. The pipeline works by separating the object (e.g., car) into two different 2D point clouds. Foreground, which is the object itself, and Background, which is the environment around it. In the Foreground, it samples a fixed number of points (e.g., 500). From every point, the model extracts the 2D offset from the centre of the object (denoted by a blue dot) and each pixel's RGB colour. In the Background, after sampling, it extracts the offset, RGB colour, and a category embedding that represents object-level information.

In the first depth integration approach shown in Fig. 1, a novel modality-depth-in is added to the current RGB and segmentation features. As depicted in the purple box (Step 2), depth is captured with a dedicated depth sensor, which subsequently undergoes a max pooling post-processing operation to enhance the data. The integration takes place at two positions in the pipeline:

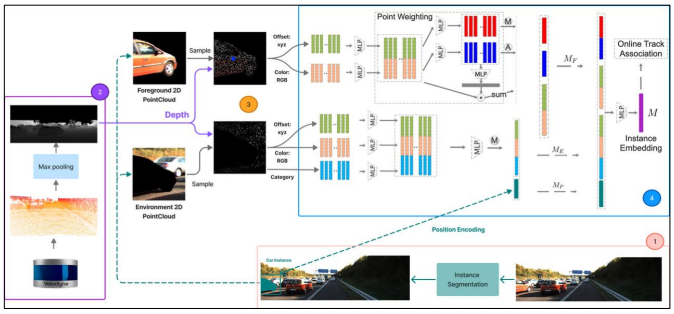

Figure 1: Depth-in integration approach. Depth data from a dedicated sensor is enhanced via max pooling and integrated with RGB and segmentation features at two points in the pipeline.

The following steps were followed to apply Max pooling on the 2D point cloud to produce the Depth image map on the right.

- First, in the Foreground, the 2D offset is expanded to 3D (x, y, z), where the z-axis is depth, creating what the authors refer to as Offset: (x, y, z).

- Second, in the Background, the same strategy is employed. Offset: (x, y, z) is calculated from environmental samples, thus augmenting the model's spatial comprehension.
- The depth features are then stacked along with RGB, and offsets features before feeding them to the positional encoding block.

In the second approach Fig. 2, further improvement is made to the pipeline in two key aspects.

- First, sensor-based depth capture is replaced with Depth Anything V2, a depth estimation framework that produces depth maps using no specialized hardware. This renders the pipeline viable for datasets without depth information, e.g., nuScenes [13] [21].
- Second, the idea of depth incorporation beyond offsetting values is generalized. Depth is directly incorporated into the colour channels, forming RGB-D inputs where the D channel plays a role analogous to pixel opacity or light intensity. Subsequently this four-channel representation is used both for the Foreground, and Background, preserving the framework of the original pipeline, but enhancing it with more accurate spatial and appearance information.
- The estimated depth features stacked in a manner similar to that of the previous experiment (along with RGB, and offsets features).

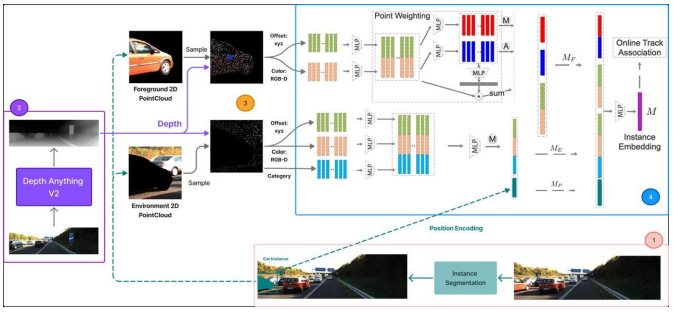

Figure 2: Enhanced depth integration using Depth Anything V2 and RGB-D inputs. Estimated depth replaces sensor-based input and is embedded as a fourth channel for both Foreground and Background, enriching spatial and appearance cues.

A. Integrating Depth Data in the PointTrack Template

The Equation (1) shows how 3D point cloud data (X, Y, Z) from a LiDAR sensor is projected onto a 2D image (x, y). The process involves aligning the LiDAR coordinate system to the camera coordinate system using rotation (R) and translation (T).

$$\begin{vmatrix} x \\ y \\ 1 \end{vmatrix} = \frac{1}{Z_c} K \left(R \begin{vmatrix} X \\ Y \\ Z \end{vmatrix} + T \right) \text{ and } K = \begin{vmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{vmatrix} \quad (1)$$

The camera's intrinsic matrix (K), which includes parameters like focal length (f_x, f_y), is then applied to map the 3D points into pixel coordinates.

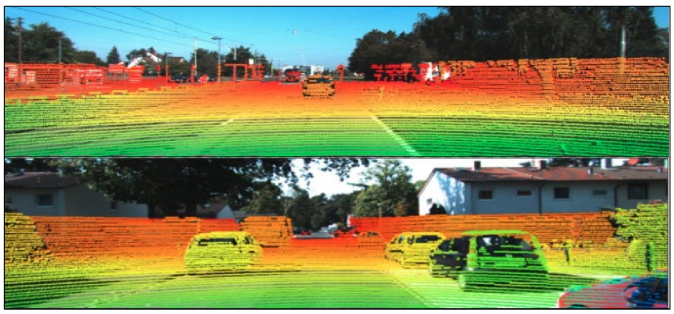

Figure 3 showcase example images of 2D representation of LIDAR point.

B. Evaluation SetUp

A comparison was made on two modalities to test performance on three methods: one with ground truth (GT) depth, one with estimated depth from a model like Depth Anything V2, and one with just RGB images. The GT depth method provides an upper-bound baseline with the most precise spatial information. The estimated depth method, on the other hand, employs monocular depth prediction. To estimate 3D structure without the need for specialized sensors, making it a more scalable alternative. The RGB-only baseline omits depth information completely, depending only on visual features.
A comparison between these configurations enables the evaluation of how the quality of depth-ground truth and estimated affects overall performance compared to using RGB input alone.

C. Evaluation Metric

MOTS evaluation involves measures that are an extension of classic MOT to cover segmentation quality. Two important measures are: MOTSA, which measures tracking performance by integrating the quality of segmentation masks and the assignment of predicted instances to ground truth objects. It rewards false positives (FP), false negatives (FN), and IDSW, and it is a holistic mask-based extension of the classic MOTA metric. The formula for MOTSA is:

$$MOTSA = 1 - \frac{(FN+FP+1)}{GT} \quad (2)$$

where FN, FP, IDSW, and GT are the number of false negatives, false positives, identity switches, and ground truth objects, respectively [20] [14]. IDSW calculates the rate at which IDs of tracked objects are wrongly exchanged during tracking, echoing the system's capacity to ensure stable identities throughout time. The lower IDSW, the better the performance and reliability, particularly in surveillance or autonomous navigation applications. The normalized IDSW is given as:

$$IDSW = 1 - (1 - IDSW_{rate})^{\frac{1}{(GT*(GT-1))}} \quad (3)$$

where IDSW rate, is the rate of identity switches, and GT is the number of ground truth objects [20] [14].

V. Results

A. Experimental Setups

In our first experiment, we trained PointTrack for 300 epochs starting from scratch, using the learning rate 5x10-6. The segmentation network was not affected in this experiment.

In the second experiment, 20 epochs training of PointTrack was enough to produce better results. Depth Anything v2 default open-sourced weights.

B. *Comparative Results*

Table 1 illustrates that the incorporation of depth information in the PointTrack framework brings about steady improvements on KITTI sequences, especially in terms of IDSW reduction. Taking sequence 0010, for instance, IDSW reduced remarkably from 6.10 to 2.00, indicating better object association, with MOTSA increasing from 96.16 to 96.84.

TABLE 1. NORMALIZED RESULTS OF PIONTTRACK WITH RGB-ONLY AND GT-DEPTH IN KITTI DATASE

Videos-Sequences	Methods	MOTSA (AVG)	IDSW (AVG)
0008	RGB-Only	97.05	5.7
	GT-Depth	97.08	5.40
0010	RGB-Only	96.16	6.10
	GT-Depth	96.84	2.00
0014	RGB-Only	94.25	1.40
	GT-Depth	94.27	1.30
0006	RGB-Only	97.77	0.00
	GT-Depth	97.77	0.00

Other sequences experience modest improvements with MOTSA improving slightly and IDSW reducing steadily or holding constant. The findings emphasize improved tracking robustness afforded by depth integration, particularly in occlusion or complicated motion scenes.

Table 2 contrasts tracking performance on two KITTI sequences with three approaches: prior to depth integration, with default sensor-based depth, and with estimated depth. Although the MOTSA scores are similar among approaches, estimated depth yields the highest MOTSA (91.40 and 97.69) and significantly fewer IDSW than the baseline and default depth.

TABLE 2. NORMALIZED RESULTS OF PIONTTRACK WITH RGB-ONLY ,GT-DEPTH AND ESTIMATED DEPTH IN KITTI DATASET

Videos-Sequences	Methods	MOTSA (AVG)	IDSW (AVG)
0002	RGB-Only	90.51	10.70
	GT-Depth	90.42	11.50
	Estimated Depth	91.40	8.20
0007	RGB-Only	97.63	9.50
	GT-Depth	97.46	13.40
	Estimated Depth	97.69	6.80

For example, in sequence 0002, estimated depth lowers IDSW from 10.70 (before depth) and 11.50 (default depth) to 8.20. Likewise, in sequence 0007, IDSW drops from 9.50 and 13.40 to 6.80. This demonstrates that the incorporation of estimated depth maps enhances object association and tracking robustness more so than sensor-based depth in these instances.

C. *Qualitative Results:*

Figure 4 highlights the output of the PointTrack framework without depth integration on a KITTI sequence. Throughout the three frames, the identity switch is stable, as indicated by the red boxes marking the vehicles. The same ID assignment is maintained for the marked cars throughout the sequence, which demonstrates stable tracking performance despite the lack of depth cues. This suggests that, in some cases, PointTrack can sustain consistent object identities over time, even in the absence of extra depth information

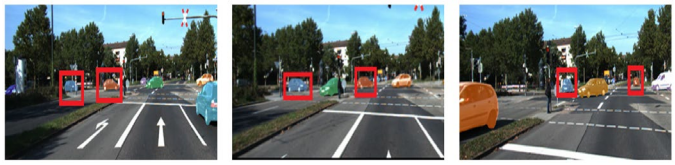

Figure 4: Output of PointTrack with RGB-Only on Kitti dataset

Figure 5 highlights the results achieved by the PointTrack framework with default depth data on a KITTI sequence. Despite the use of depth information, the figure demonstrates the persistence of identity switches, as indicated by the red boxes. The IDs assigned to the same vehicles differ over the three frames, suggesting that even with the original depth data, the model finds it difficult to maintain object identities consistently in such scenarios. The observation infers that while default depth provides some geometric cues, it may not be sufficient to resolve ID assignment issues, especially in complex scenes with objects that are similar in appearance or undergo partial occlusions.

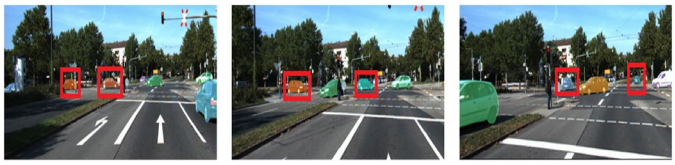

Figure 5: Output of PointTrack with GT-Depth on Kitti dataset.

Figure 6 highlights the output of the PointTrack framework with estimated depth on a KITTI sequence. Unlike in previous setups, the ID assignments are consistent throughout the three frames, as shown by the red boxes. The model is capable of maintaining the same identities for the tracked vehicles, which suggests that the incorporation of estimated depth enhances object association and diminishes identity switches. This further strengthens the successful notion of employing depth estimation techniques in improving the stability of tracking, particularly in difficult situations where consistent tractlets are vital.

Figure 6: Output of PointTrack with estimated depth on Kitti dataset

VI. Conclusions

This research pushes the state of the art in multi-object tracking and segmentation by exhaustively benchmarking sophisticated trackers and augmenting their detection phases with advanced approaches. It offers important insights into the impact of various detection approaches on overall tracking performance and shows that the improvement of backbone networks greatly increases tracker accuracy and reliability. A pivotal contribution is the innovative incorporation of depth information into the PointTrack framework, which augments object discrimination in complicated scenes where 2D appearance features alone are inadequate. Depth integration significantly decreases identity switches and improves tracking stability, particularly when employing estimated depth maps that convey richer spatial awareness than default sensor-based depth. These findings validate the successful utilization of depth as instrumental for robust and accurate tracking in autonomous driving environments.

For future research, endeavours will be directed towards continued optimization of depth estimation methods, better integration with tracking pipelines, and investigating sophisticated backbone architectures for better precision and fewer identity switches. Furthermore, model parameter tuning and testing with medium-sized networks could potentially trade off accuracy for real-time processing requirements, leading to more stable and efficient multi-object tracking systems. Building on this output, future research could also explore the use of temporal depth consistency for additional reduction of tracking errors, domain adaptation methods for generalizing models to varied environments, and sensor fusion with LiDAR and radar data for improving depth accuracy and robustness under adverse weather or lighting conditions.

Acknowledgment

Khadijah Alkandary and Ahmet Serhat Yildiz express their heartfelt gratitude to their family for their endless support, patience, and encouragement throughout this research journey. Ahmet Serhat Yildiz's Ph.D. is sponsored by the Ministry of National Education of Türkiye.

References

[1] A. Milan, L. Leal-Taixe, I. Reid, S. Roth and K. Schindler, "MOT16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831,* 2016.

[2] G. Bhat, M. Danelljan, L. V. Gool and R. Timofte, "Learning discriminative model prediction for tracking," *Proceedings of the IEEE/CVF international conference on computer vision,* pp. 6182--6191, 2019.

[3] P. Chu and H. Ling, "Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking," *Proceedings of the IEEE/CVF international conference on computer vision,* vol. 2019, pp. 6172--6181.

[4] W. Zhang, H. Zhou, S. Sun, Z. Wang, J. Shi and C. C. Loy, "Robust multi-modality multi-object tracking," *Proceedings of the IEEE/CVF international conference on computer vision,* pp. 2365-2374, 2019.

[5] B. Shuai, A. Berneshawi, X. Li, D. Modolo and J. Tighe, "SiamMOT: Siamese Multi-Object Tracking," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 12372-12382, 2021.

[6] Z. Xu, W. Zhang, X. Tan, W. Yang, X. Su, Y. Yuan, H. Zhang, S. Wen, E. Ding and L. Huang, "Pointtrack++ for effective online multi-object tracking and segmentation," *arXiv preprint arXiv:2007.01549,* 2020.

[7] X. Zhou, V. Koltun and P. Krahenbuhl, "Tracking objects as points," *European conference on computer vision,* pp. 474--490, 2020.

[8] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng and H. Zhao, "Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 10371-10381, 2024.

[9] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng and H. Zhao, "Depth anything v2," *Advances in Neural Information Processing Systems,* vol. 37, pp. 21875-21911, 2024.

[10] A. Geiger, P. Lenz and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," *2012 IEEE conference on computer vision and pattern recognition,* pp. 3354--3361, 2012.

[11] A. Geiger, P. Lenz, C. Stiller and R. Urtasun, "Vision meets robotics: The kitti dataset," *The international journal of robotics research,* vol. 32, pp. 1231--1237, 2013.

[12] Z. Xu, W. Zhang, X. Tan, W. Yang, H. Huang, S. Wen, E. Ding and L. Huang, "Segment as points for efficient online multi-object tracking and segmentation," *European conference on computer vision,* pp. 264--281, 2020.

[13] J. Liu, Y. Wu, M. Gong, Q. Miao, W. Ma, C. Xu and C. Qin, "M3SOT: Multi-frame, multi-field, multi-space 3D single object tracking," *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 38, no. 4, pp. 3630--3638, 2024.

[14] M. Khanchi, M. Amer and C. Poullis, "Depth-Aware Scoring and Hierarchical Alignment for Multiple Object Tracking," *arXiv preprint arXiv:2506.00774,* 2025.

[15] H. Sun, Y. Li, G. Yang, Z. Su and K. Luo, "View adaptive multi-object tracking method based on depth relationship cues," *Complex & Intelligent Systems,* vol. 11, no. 2, p. 145, 2025.

[16] X. Han, P. Fang, Y. Tian, J. Yu, X. Cai, D. Roggen and P. Birch, "GRASPTrack: Geometry-Reasoned Association via Segmentation and Projection for Multi-Object Tracking," *arXiv preprint arXiv:2508.08117,* 2025.

[17] W. Zeng, J. Fan, X. Tian, H. Chu and B. Gao, "FusionTrack: An Online 3D Multi-object Tracking Framework Based on Camera-LiDAR Fusion," *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* pp. 4920--4925, 2024.

[18] S. Kumar, M. Hassan, M. Escudero-Vinolo, A. Hannan, A. Manzoor, A. Godinho, I. M. Pires and P. J. Coelho, "Multi-Modal Tracking Using LiDAR and Visual Signals," *2024 11th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI),* pp. 371--378, 2024.

[19] Z. Song, L. Liu, F. Jia, Y. Luo, C. Jia, G. Zhang, L. Yang and L. Wang, "Robustness-aware 3d object detection in autonomous driving: A review and outlook," *IEEE Transactions on Intelligent Transportation Systems,* 2024.

[20] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger and B. Leibe, "Mots: Multi-object tracking and segmentation," *Proceedings of the ieee/cvf conference on computer vision and pattern recognition,* pp. 7942--7951, 2019.

[21] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 11621-11631, 2020.

[22] M.-J. Kwon, I.-J. Yu, S.-H. Nam, and H.-K. Lee, "CAT-Net: Compression Artifact Tracing Network for Detection and Localization of Image Splicing,"in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 375–384..