PBJDT: Point-Based Joint Detection-and-Tracking

Zhen-Xun Lee[1] and Jian-Jiun Ding [2*]
[1] National Taiwan University, Taiwan
E-mail: r11942118@ntu.edu.tw Tel: +886-937265017
[2*] National Taiwan University, Taiwan
E-mail: jjding@ntu.edu.tw Tel: +886-233669652

***Abstract*—Multiple object tracking is a more and more important technique with the rapid development of autonomous driving. In this work, we introduce a keypoint-based multiple object tracking method aim to address the challenges in complex road environments. It utilizes anchor-free keypoint detection to reduce computational resources while extracting robust features through an improved deep learning model architecture. We implement a background suppression technique to minimize false detections and incorporate the information from adjacent frames to capture object motion characteristics. To address occlusion and overexposure scenarios, we innovatively combine a Kalman filter with the detector and dynamically adjust the tracking strategy based on detection confidence. Experimental results demonstrate that the proposed method can be performed in real time and achieve a MOTA of 92.51% on the KITTI tracking dataset, which outperforms state-of-the-art methods.**

I. INTRODUCTION

Multiple Object Tracking (MOT) is a rapidly evolving and crucial research area within computer vision, with wide-ranging applications in surveillance systems, autonomous driving, human action analysis, and other fields. It aims to simultaneously track the movement trajectories of multiple objects in video sequences. The requirement for efficient and accurate MOT systems has grown exponentially, particularly in autonomous driving, where real-time tracking of numerous vehicles, pedestrians, and other road users is critical for safe navigation. MOT approaches can be broadly categorized based on how they capture object relationships, with tracking-by-detection and tracking-by-attention being prominent paradigms.

There are several well-known tracking-by-detection methods, e.g., CenterTrack [1], FairMOT [2], and SORT [3]. These approaches utilize detection results from adjacent frames to capture object relationships, followed by a refinement stage to manage trajectories. Conversely, prominent tracking-by-attention methods such as TrackFormer [4], ByteTrack [5], and MOTR [6] employ attention mechanisms to capture object relationships across frames.

Furthermore, in the detection step, approaches can be differentiated by their pipeline structure and methodology: some employ pre-defined anchors to detect objects, while others utilize anchor-free methods, directly predicting detection results from the backbone network.

Traditional multiple object tracking methods have predominantly relied on anchor-based mechanisms, wherein multiple predefined anchors are strategically placed within the image for object detection and tracking. While these anchor-based approaches have substantially enhanced detection and tracking accuracy, they are not without limitations. Key challenges include the necessity for precise anchor configuration, high computational demands, and suboptimal performance in tracking small objects and navigating dense scenes.

As the complexity of real-world scenarios escalates, particularly in urban driving environments, these limitations have become increasingly pronounced. Consequently, there is a growing need for novel solutions that can effectively address these challenges, especially in scenarios characterized by occlusion and rapid illumination changes.

To address these challenges, we propose an innovative approach to multiple object tracking. Our method offers more flexible handling of objects under varying illumination conditions and occlusions, reduces computational costs, and demonstrates superior performance in challenging scenarios.

In this paper, we introduce PBJDT (Point-Based Joint Detection and Tracking), a novel framework designed to tackle these MOT challenges. Our main contributions include:

1. An hourglass backbone architecture incorporating multi-scale feature fusion is proposed to enhance the robustness of extracted features.
2. A novel background suppression method is applied to feature heatmaps to mitigate the detection misjudgments caused by background interference.
3. A flexible system that applied Kalman filtering is proposed to enhance the tracking robustness in challenging conditions, particularly in the scenarios characterized by occlusions and rapid illumination changes.

Preliminary results are highly promising, with PBJDT achieving a Multiple Object Tracking Accuracy (MOTA) [17] of 92.51% on the KITTI tracking dataset [16], surpassing existing state-of-the-art methods. This performance underscores the potential of our approach in advancing the field of multiple object tracking.

TABLE I. SUMMARY OF MOT PROCESSES FOR DIFFERENT TRACKING PARADIGMS

Method \ Process	Feature Extraction	Correlation	Refinement
Tracking-by-detection	Uses DNNs to extract features from detected objects	Correlates detected instances with existing tracklets	Applies Re-ID techniques to refine tracking trajectories
Tracking-by-attention	Employs attention modules to extract and focus on relevant features	Correlates tokens (which can represent objects or parts of objects)	Refines input features based on feature correlation

II. RELATED WORKS

Existing literature on multiple object tracking (MOT) can be categorized according to their respective tracking paradigms.

A. Tracking-by-Detection

Tracking-by-detection has emerged as a widely adopted paradigm in multiple object tracking. This approach first detects objects independently in each frame, then links these detections across frames to form trajectories. The tracking-by-detection framework gained prominence with the advent of powerful object detectors such as the Faster R-CNN [7], YOLO [8], the SSD [9], and the CenterNet [10].

Tracking-by-detection methods can be broadly classified into two main categories: one-stage and two-stage methods. Two-stage methods bifurcate the tracking task into object detection and data association, while one-stage methods simultaneously detect object information and learn inter-frame relationships. For instance, CenterTrack [1] exemplified a classic one-stage framework, employing point-based detection for concurrent detection and tracking. For refinement, some methods incorporate additional modules to enhance trajectory management. FairMOT [2] utilized Re-identification (Re-ID) to address id switch issues, while SORT [3] employed the Hungarian algorithm for data association.

B. Tracking-by-Attention

Tracking-by-attention leverages attention mechanisms to enhance the tracking process. This approach enables the model to focus on salient parts of the input data, thereby improving the accuracy and robustness of object tracking, particularly in scenarios involving occlusions and complex object interactions.

Notable examples in this category include:

1. TransTrack [11], which integrated transformers with a traditional detection-based tracking framework.
2. TrackFormer [4], which introduced a transformer-based architecture for end-to-end object tracking, utilizing attention mechanisms to directly model inter-frame object interactions.
3. ByteTrack [5], which proposed a simple yet effective association method that exploits both high and low confidence detections through Kalman filtering, significantly enhancing tracking performance in crowded scenes.

While tracking-by-attention methods offer improved handling of occlusions and complex object interactions, they face several challenges. These include high computational complexity, substantial requirements for training data, and performance limitations in extremely dense scenes.

C. Summary and Other Approaches

We summarize the key processes of tracking-by-detection and tracking-by-attention methods in Table I, highlighting their approaches to feature extraction, correlation, and refinement. Tracking-by-detection and tracking-by-attention are two predominant paradigms in MOT.

Tracking-by-regression: This method directly regresses the object's position in subsequent frames, often using neural networks trained on sequential data.

Tracking-by-segmentation: This approach combines instance segmentation with tracking, providing more detailed object information.

Hybrid approaches: These methods integrate multiple techniques to achieve enhanced performance. For example, some methods combine detection-based tracking with attention mechanisms or incorporate segmentation information into the tracking pipeline.

These diverse approaches reflect the ongoing innovation in the field of multiple object tracking, each offering unique strengths in addressing the challenges of complex tracking scenarios.

III. PROPOSED METHOD

Regarding illumination changes, when vehicles pass through tunnels or shaded areas, images captured by dashcams may be overexposed or underexposed, leading to misjudgments by detectors. The image information in these scenes is compromised due to abrupt lighting changes. On the other hand, object occlusion is commonplace in everyday scenarios, where target objects may be partially obscured by background elements or other objects, or become difficult to detect due to changes in camera perspective.

In this section, we introduce our novel approach, 'Point-based Joint Detection and Tracking' (PBJDT). The overall pipeline is shown in Fig. 3. This method is designed to address the challenges of object detection and tracking in scenarios characterized by rapid illumination changes and frequent object occlusions. PBJDT integrates three key components to enhance detector robustness in these challenging environments:

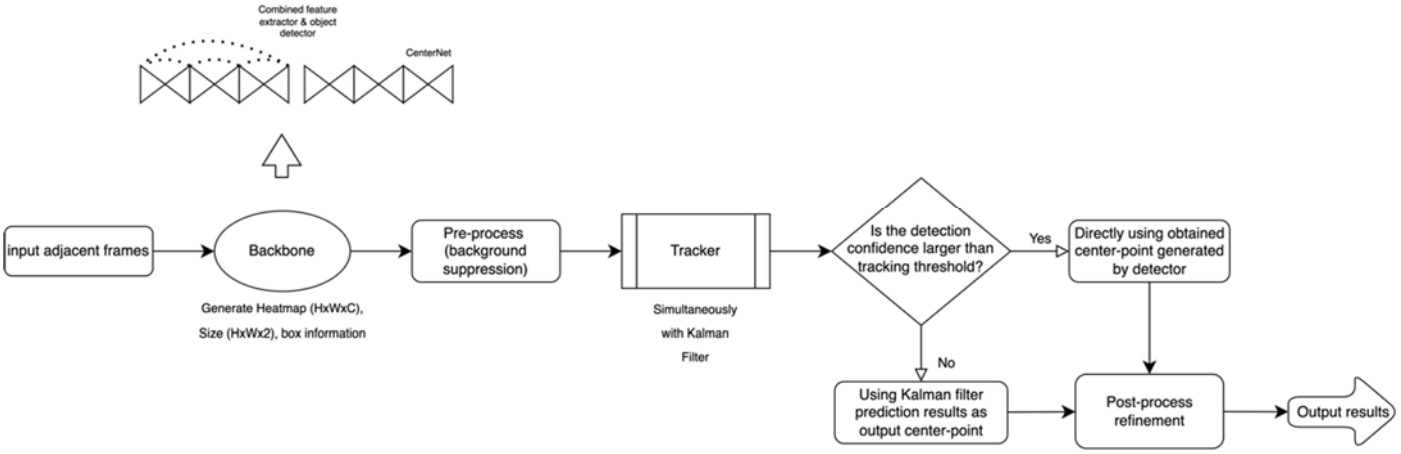

Fig. 1 The overall pipeline of the proposed PBJDT.

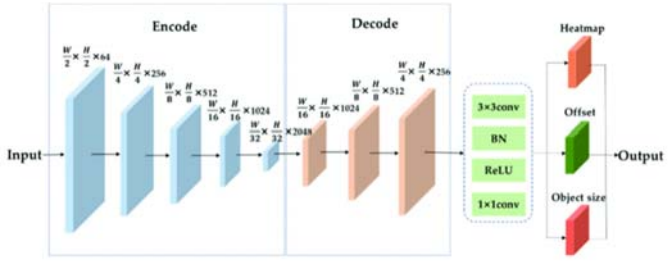

Fig. 2 The proposed hourglass network architecture

1. Multi-scale feature fusion strategies
2. Background suppression on feature heatmaps
3. Kalman filter-aided tracking

The synergy of these components enables our method to maintain high performance in complex and dynamic scenes, particularly those encountered in urban driving environments.

A. Backbone and Detection Network

Our backbone network employs a sophisticated architecture designed for robust feature extraction and efficient object detection. The network structure utilizes a 6x ResNet-50 stacked hourglass network, incorporating skip-connections and multi-scale feature fusion to enhance the robustness of our extracted features.

The hourglass architecture is specifically designed to capture and consolidate information across multiple scales, which is crucial for accurate object localization and tracking. Each hourglass module consists of a downsampling path, an upsampling path, along with bottleneck and skip connections. This design enables our backbone to identify target keypoints (center-points) under different resolutions through consecutive down-sampling and up-sampling processes. The obtained keypoints are then propagated through skip-layers in each layer, and globally combined to synthesize our final center-point heatmap.

For feature fusion, we employ a Feature Pyramid Network (FPN) structure to fuse features at different scales, analogous to the SIFT [12] concept. The skip-connections play a vital role in combining low-level spatial details with high-level semantic information, further enhancing the network's ability to detect objects across various scales and complexities.

The output of our network generates a center-point heatmap of target objects with 1/4 resolution of the input image. This resolution maintains a balance between computational efficiency and detection precision. Additionally, the network outputs a series of box information including center-point coordinates, object length and width, detection confidence, and object IDs. Thus, we could obtain offset through computing adjacent frames' box information.

The proposed loss function includes:

1. Heatmap Loss: It uses the Focal loss (a modified cross-entropy loss for imbalanced data) for optimizing the center-point prediction, aiming for more precise detection. Our heatmap focal loss formulation:

$$L_{hm} = \begin{cases} -\frac{1}{N} * \sum(1-p_{ij})^{\alpha} * \log(p_{ij}) \ , \ if \ y_{ij} = 1 \\ -\frac{1}{N} * \sum(1-y_{ij})^{\beta} * p_{ij}^{\alpha} * \log(1-p_{ij}) \ , \ otherwise \end{cases} \quad (1)$$

where p_{ij} is the prediction heatmap response, y_{ij} is the ground truth, and both α, β are the hyper-parameter of the focal loss. The key design consideration is to address the class imbalance in heatmap prediction.

2. Offset Loss: It uses the L_1 loss to optimize the precise location of center-points.

$$L_{off} = 1/N * \sum\left|\hat{O}_k - O_k\right| \quad (2)$$

where $\hat{O}_k$ is the predicted offset and O_k is the ground truth offset.

3. Size Loss: It uses the L_1 loss to optimize the width and height of bounding boxes.

$$L_{size} = 1/N * \sum|\hat{S}_k - S_k| \quad (3)$$

where $\hat{S}_k$ is the predicted size and S_k is the ground truth size.

4. Tracking Offset Loss: It also uses the L_1 loss to optimize object position changes between consecutive adjacent frames.

$$L_{track} = 1/N * \sum|\hat{t}_k - t_k| \quad (4)$$

where $\hat{t}_k$ is the predicted tracking offset and t_k is the ground truth tracking offset.

We all choose the L_1 loss to optimize above three regression tasks due to the property of more robust to outliers. Therefore, our total loss is a weighted sum of these components:

$$L_{total} = \lambda_{hm}L_{hm} + \lambda_{off}L_{off} + \lambda_{size}L_{size} + \lambda_{track}L_{track} \quad (5)$$

where λ_{hm}, λ_{off}, λ_{size}, λ_{track} are weight coefficients to balance the contribution of different loss terms. We combine detection and tracking optimization along with careful tuning of weight coefficients for optimal performance.

B. Background Suppression

To mitigate some misjudgments caused by background interference, we implement a background suppression method on the center-point heatmap. This process involves:

1. Applying a sigmoid activation function to the heatmap to enhance response contrast, pushing background responses towards zero.
2. Implementing a confidence threshold:
   - If the detection confidence score exceeds the threshold, we consider it a target object.
   - Otherwise, we reference adjacent frames' center-point heatmaps to determine if the detection is a false positive or true positive.
3. Multiplying the heatmap with a sigmoid function kernel:
   - This pushes the heatmap response in regions adjacent to target objects towards 1 (with the center-point exactly at 1).
   - Simultaneously, it suppresses background regions towards zero.

It's worth noting that our method's effectiveness may be slightly limited when applied to driving recording data, where the background exhibits relative motion. In such cases, while the background response may not approach zero entirely, our method still maintains a significant suppression effect.

C. Kalman Filter Module

To address challenges posed by rapid illumination changes and occlusion scenarios that are difficult even for human vision to classify, we incorporate a Kalman filter module. This module is specifically designed to track object trajectories in frames where objects are overexposed due to passing trees, entering tunnels, or similar situations.

While modern detectors typically achieve impressive performance with high accuracy rates, we've developed a mechanism to determine whether each frame can be reliably processed by the detector alone or requires the involvement of the Kalman filter module for tracking. Our decision criterion is primarily based on the detection confidence score, which indicates the detector's performance for a given frame.

Therefore, we apply the Kalman filter [13] to track our center-point coordinates in each frame while our designed detector operates simultaneously, anytime if our detection confidence is lower than tracking threshold, Kalman filter could get involved in our final output center-point. This parallel processing allows for robust tracking even when detection becomes unreliable. When the detection confidence falls below a predetermined tracking threshold, the Kalman filter's predictions are incorporated into our final output center-point. This adaptive approach ensures optimal use of both detection and tracking information.

Moreover, we assume our whole system is linear system given that our input video is processed at 15 frames per second (fps), the center-point trajectory between adjacent frames can be approximated as linear. This assumption simplifies our model and allows for efficient computation. Based on our linear system assumption, we employ a Linear Kalman Filter rather than more complex nonlinear variants like the Extended Kalman Filter (EKF) [14] or Unscented Kalman Filter (UKF) [15], balancing computational efficiency with tracking accuracy with this choice.

Our Kalman filter state vector includes the center-point coordinates and their velocities. The measurement model directly relates to the detected center-point coordinates when available. For update mechanism, in frames where detection confidence is high, we use the detector's output to update the Kalman filter's state. When detection is unreliable, we rely on the Kalman filter's predictions to maintain tracking continuity.

D. Post Process Refinement

To tackle occlusion challenges in complex scenes, we've implemented an enhanced frame referencing strategy. Unlike conventional methods that typically consider only three adjacent frames before and after the current frame, our approach expands this window to five frames in each direction. This extended temporal context allows us to extract more comprehensive reference data for our tracking process.

By inputting box information from a wider range of adjacent frames, we improve our Kalman filter module's accuracy in estimating center-point trajectories. This expanded temporal window enables our system to better capture long-

range dependencies and infer object positions even during temporary occlusions. It significantly enhances the performance in challenging environments characterized by frequent occlusions and complex object interactions. It allows for more robust and consistent multiple object tracking in scenarios where traditional methods might falter due to limited temporal context.

Furthermore, in some scenes where the detection confidence is just slightly above the tracking threshold, there may be some misjudgments. Therefore, we use spatial continuity to check if the objects' motion is reasonable. Typically, objects do not exhibit large movements in adjacent frames. By examining the movement of each object, we can determine whether there may be a false alarm.

IV. Experiments and Results

A. Datasets and Metrics

Datasets. The KITTI tracking dataset [16] is used in our experiments. The KITTI dataset is a comprehensive benchmark for various computer vision tasks, including tracking, and contains annotated images of urban scenes captured from a moving vehicle. This dataset is essential for evaluating object detection and tracking in real-world driving scenarios.

Metrics. Our evaluation metric is Multiple Object Tracking Accuracy (MOTA) [17], which provides a comprehensive measure of tracking performance across each dataset. MOTA encapsulates three key aspects of tracking: detection accuracy, false alarms, and identity preservation. The MOTA [23] metric is defined by the following equation:

$$MOTA = 1 - \frac{\sum_t FN_t + FP_t + IDSW_t}{\sum_t GT_t} \quad (6)$$

where FN_t (False Negatives) represents the number of ground truth objects that were not detected, FP_t (False Positives) represents the number of detected objects that do not correspond to any ground truth object, $IDSW_t$ (ID Switches) represents the number of times that a tracked object changes its identity label, and GT_t represents ground truth, the total number of ground truth objects across all frames.

B. Results

The performance of the proposed and other MOT methods on is shown in Table II, which show that the proposed PBJDT outperformed the state-of-the-art (SOTA) method UCMCTrack [18] with 2.1% improvements in MOTA. This improvement is largely attributed to our background suppression technique, designed hourglass network and adaptive Kalman filter, which together enhance the system's robustness in complex scenarios.

Fig. 3 shows the visualization results that PBJDT can handle illumination changes problem rather than previous similar framework CenterTrack [1].

TABLE II. THE PERFORMANCE ON KITTI FOR MULTIPLE OBJECT TRACKING

Methods (Year)	**MOTA**
3DT (2017) [19]	84.52
MOTS (2019) [20]	84.83
JRMOT (2020) [21]	85.70
CenterTrack (2020) [6]	89.44
SRK ODESA (2020) [22]	90.03
QD-3DT (2022) [23]	86.41
EagerMOT (2021) [24]	87.82
DEFT (2021) [25]	88.95
OC-SORT (2023) [26]	90.3
UCMCTrack (2024) [18]	90.4
Proposed PBJDT	**92.51**

V. Conclusions

In this work, we introduce PBJDT, a point-based joint detection-and-tracking MOT framework. It features a specially designed hourglass network backbone, a background suppression technique, and an adaptive Kalman filter, significantly improving tracking performance in scenarios with rapidly changing illumination and occlusion.

References

[1] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *European Conference on Computer Vision*, pp. 474-490, 2020.

[2] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *Int. J. Computer Vision*, vol. 129, pp. 3069-3087, 2021.

[3] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *IEEE Int. Conf. Image Processing*, pp. 3464-3468, 2016.

[4] T. Meinhardt, A. Kirillov, L. Leal-Taixé, and C. Feichtenhofer, "Trackformer: Multi-object tracking with transformers," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 8844-8854, 2021.

[5] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-object tracking by associating every detection box," in *European Conf. Computer Vision*, pp. 1-21, 2022.

(a) The results of CenterTrack

(b) The results of the Proposed PBJDT

Fig. 3 Key frame comparison with CenterTrack [1]. The upper three frames show the results of CenterTrack, while the lower three frames show the results of the proposed PBJDT. It can be seen that PBJDT effectively handles illumination changes and tracks the target object with high precision.

[6] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "MOTR: End-to-end multiple-object tracking with transformer," in *European Conf. Computer Vision*, pp. 659-675, 2022.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137-1149, June 2016.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 779-788, 2016.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European Conf. Computer Vision*, pp. 21-37, 2016.

[10] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 6569-6578, 2019.

[11] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "Transtrack: Multiple-object tracking with transformer," *arXiv preprint arXiv:2012.15460*, 2020.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60, issue 2, pp. 91-110, 2004.

[13] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Engineering*, vol. 82, no. 1, pp. 35-45, 1960.

[14] A. Gelb, *Applied Optimal Estimation*, MIT Press, 1974.

[15] S. J. Julier and J. K. Uhlmann, "New extension of the Kalman filter to nonlinear systems," in *Int. Symp. Aerospace/Defense Sensing, Simulation, and Controls*, vol. 3068, pp. 182-193, 1997.

[16] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 3354-3361, 2012.

[17] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image and Video Processing*, vol. 2008, article 246309, 2008.

[18] K. Yi, K. Luo, X. Luo, J. Huang, H. Wu, R. Hu, and W. Hao, "Ucmctrack: Multi-object tracking with uniform camera motion compensation," in *AAAI Conf. Artificial Intelligence*, vol. 38, no. 7, pp. 6702-6710, 2024.

[19] W. Choi and S. Savarese, "Multiple target tracking in world coordinate with single, minimally calibrated camera," in *European Conf. Computer Vision*, pp. 553-569, 2017.

[20] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, "MOTS: Multi-object tracking and segmentation," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 7942-7951, 2019.

[21] A. Shenoi, M. Patel, J. Y. Gwak, P. Goebel, A. Sadeghian, H. Rezatofighi, R. Martın-Martın, and S. Savarese, "JRMOT: A real-time 3d multi-object tracker and a new large-scale dataset," in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pp. 10335-10342, 2020.

[22] D. Mykheievskyi, D. Borysenko, and V. Porokhonskyy, "Learning local feature descriptors for multiple object tracking," in *Asian Conf. Computer Vision*, pp. 1-18, 2020.

[23] H. N. Hu, Y. H. Yang, T. Fischer, T. Darrell, F. Yu, and M. Sun, "Monocular quasi-dense 3D object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, issue 2, pp. 1992-2008, 2022.

[24] A. Kim, A. Osep, and L. Leal-Taixe, "EagerMOT: 3D Multi-object tracking via sensor fusion," in *IEEE Int. Conf. Robotics and Automation*, pp. 11315-11321, 2021.

[25] M. Chaabane, P. Zhang, J. R. Beveridge, and S. O'Hara, "DEFT: Detection embeddings for tracking," *arXiv preprint arXiv: 2102.02267*, 2021.

[26] J. Cao, J. Pang, X. Weng, R. Khirodkar and K. Kitani. "Observation-centric sort: Rethinking sort for robust multi-object tracking," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 9686-9696, 2023.