Article

MSA-MOT: Multi-Stage Association for 3D Multimodality Multi-Object Tracking

Ziming Zhu 1, Jiahao Nie 1, Han Wu 1, Zhiwei He 1,2,* and Mingyu Gao 1,2

1 The School of Electronic Information, Hangzhou Dianzi University, Hangzhou 310018, China
2 Zhejiang Province Key Laboratory of Equipment Electronics, Hangzhou 310018, China
* Correspondence: zwhe@hdu.edu.cn

Abstract: Three-dimensional multimodality multi-object tracking has attracted great attention due to the use of complementary information. However, such a framework generally adopts a one-stage association approach, which fails to perform precise matching between detections and tracklets, and, thus, cannot robustly track objects in complex scenes. To address this matching problem caused by one-stage association, we propose a novel multi-stage association method, which consists of a hierarchical matching module and a customized track management module. Specifically, the hierarchical matching module defines the reliability of the objects by associating multimodal detections, and matches detections with trajectories based on the reliability in turn, which increases the utilization of true detections, and, thus, guides accurate association. Then, based on the reliability of the trajectories provided by the matching module, the customized track management module sets maximum missing frames with differences for tracks, which decreases the number of identity switches of the same object and, thus, further improves the association accuracy. By using the proposed multi-stage association method, we develop a tracker called MSA-MOT for the 3D multi-object tracking task, alleviating the inherent matching problem in one-stage association. Extensive experiments are conducted on the challenging KITTI benchmark, and the results show that our tracker outperforms the previous state-of-the-art methods in terms of both accuracy and speed. Moreover, the ablation and exploration analysis results demonstrate the effectiveness of the proposed multi-stage association method.

Keywords: 3D multi-object tracking; multimodal; multi-stage association; track management

1. Introduction

Multi-object tracking (MOT) is a key component of autonomous driving and robot navigation systems [1,2], which aims to obtain dynamic information about the environment by associating the objects in consecutive frames. Early multi-object tracking methods [3–5] are based on the images captured by camera and achieves promising results. However, the image information typically degrades in complex scenes (e.g., due to overexposure and insufficient light), leading to limited performance. In contrast, the point cloud data acquired by lidar include depth information of the scene [6], which shows robustness to complex conditions. Therefore, researchers are working on developing 3D multi-object tracking (3D MOT) frameworks [7,8] based on lidar sensors.

Existing 3D MOT methods can be mainly divided into single-modality-based [9,10] and multimodality-based [11–13] methods. Single-modality methods are generally based on lidar sensors for tracking. In contrast, multimodal approaches typically show better performance due to the use of complementary information. Multimodality methods [13–15] generally use 2D and 3D detectors to generate bounding boxes of objects, and then use a feature fusion module to fuse the different modal features provided by the feature extractor. Afterward, the objects are associated with candidate tracks based on similarity to perform tracking. However, most existing methods adopt one-stage association, failing to match the detections and tracklets with great accuracy. The one-stage association method filters
out the low-score detections before input, and thus ignores the real objects contained in the low-score detections, which significantly affects the association accuracy. This method also does not re-track objects that are occluded for a long time, which leads to identity switching of the same object.

To solve the above problems, we propose a novel 3D multimodality tracker, namely, MSA-MOT, which is centered on a multi-stage association method consisting of a hierarchical matching module and a customized track management module. Specifically, to guide accurate association between detections and tracks, we propose a hierarchical matching module. First, 3D bounding boxes are projected to the image and matched with 2D bounding boxes to evaluate the detection reliability. Second, considering that most unreliable detections degrade the association accuracy, we only match the high-reliability detections with all candidate trajectories. Furthermore, the unmatched trajectories in the previous stage are matched with unreliable 3D detections, which increases the utilization of real objects in unreliable 3D detections. Finally, considering that some objects can be detected by camera but not by lidar (e.g., distant objects with low scores), we match the unreliable 2D detections with the remaining trajectories, reducing the number of missed detections of true objects at a distance. Based on this module, reliability information on tracklets is generated. To fully utilize this information to effectively manage the tracklets, we propose a customized track management module. Premature deletion of tracklets and tracking drift are the main reasons for object identity switching. Moreover, the tracks that are matched with 3D detections may not disappear in a short time. Therefore, we set larger maximum missing frames for reliable trajectories than for unreliable tracks, effectively reducing the number of identity switches in tracking.

To demonstrate the effectiveness and advancement of the proposed tracking method, we conduct a series of comparison experiments and ablation experiments on the challenging KITTI dataset [16]. As shown in Figure 1, the proposed MSA-MOT method achieves the highest Higher-Order Tracking Accuracy (HOTA), while running at a high speed of 130 frames per second (FPS), which verifies the effectiveness of our method.

Figure 1. Comparison results of MSA-MOT with previous state-of-the-art methods. ▲ and ★ represent camera-based and lidar-based methods, respectively. In addition, ★ represents multimodality-based methods. A red outline indicates that the method uses the same detector as our method. A higher value indicates better performance. The results show that MSA-MOT achieves accurate tracking with great speed. The results are presented in detail in Table 1.
To summarize, our main contributions are as follows:

- We propose a novel tracking method MSA-MOT for 3D MOT in complex scenes, in which we improve the association scheme by utilizing multi-stage association, and, thus, achieve precise tracking over a long period of time.
- In the multi-stage association method, the proposed hierarchical matching module successively associates the high- and low-reliability detections, alleviating the long-standing problem of incorrect association. In addition, a customized track management module is proposed for managing tracklets based on the information provided by the matching module, effectively addressing the severe identity switch in tracking.
- Extensive experiments are conducted on the challenging KITTI benchmark. The results show that MSA-MOT achieves state-of-the-art performance (78.52% on HOTA, 97.11% on sAMOTA, and 130 FPS), which demonstrates the effectiveness of our novel multi-stage association method.

The remainder of the paper is structured as follows: Section 2 discusses the related work. Section 3 details the structure of our proposed framework and analyzes each module. Section 4 presents the experiments and results analysis. Section 5 discusses the conclusions and future plans of this study.

2. Related Work

2.1. 2D MOT

Recent research on 2D MOT has mainly focused on two paradigms: tracking by detection and joint detection and tracking. The tracking by detection [17–19] paradigm is based on the detections provided by a detector and uses filters for state estimation such as Kalman [20]. Then, the detections and tracks are associated based on similarity to perform tracking. The joint detection and tracking [21,22] paradigm performs detection and tracking at the same time, and the same backbone network is used to extract features for both detection and tracking.

2.2. Single-Modality 3D MOT

With the rapid development of 2D MOT and 3D detectors [23–25], lidar-based 3D MOT has received great attention. For example, inspired by SORT [5], Weng et al. [7] developed a simple tracking method based on the Kalman filter, which uses the 3D intersection over union (3D IoU) and Hungarian algorithm [26] to perform tracking. Chiu et al. [27] proposed the replacement of the 3D IoU with the Mahalanobis distance [28], and initialized the covariance in the Kalman filtering process with the statistics of the training set. Zhai et al. [29] proposed a scene flow estimation network for obtaining implicit motion information, and then tracked the object of interest by an identity propagation strategy. Moreover, similar to CenterTrack [30], which is a 2D MOT tracker, Yin et al. [31] proposed representing each object by the center of its bounding box, and used speed estimation to achieve tracking without filtering. Wu et al. [32] proposed a tracker based on a data association scheme guided by prediction confidence. Kim et al. [33] explored the impact of geometric relationships between objects for 3D multi-object tracking based on graph neural networks. In addition, various methods [34–36] use other kinds of sensors to perform tracking.

In addition to the above works, some researchers are committed to using multiple cameras to estimate 3D bounding box information for 3D MOT. For example, Hu et al. [9] proposed the use of quasi-dense similarity learning to identify various objects with appearance characteristics, and then used a 3D bounding box depth sorting trial method for robust instance association. Marinello et al. [37] proposed the combination of triplet embedding and motion characterization for 3D object tracking.

2.3. Multimodality 3D MOT

Single-modality methods generally reduce the ability to obtain information in complicated situations, e.g., cameras are susceptible to interference from light intensity, and the point cloud data acquired by lidar are sparse when the object is at a distance. There-
fore, multimodality methods for 3D MOT have recently attracted wide attention due to the sufficient information provided by sensors. For example, Zhang et al. [15] designed a sensor-agnostic multimodal framework for 3D MOT, which focuses on enabling joint optimization for the basic feature extractor of each mode and the adjacency estimator of the cross mode. Weng et al. [38] obtained both apparent and motion features from 2D and 3D space, and proposed a feature interaction mechanism based on a graph neural network. In addition, Zeng et al. [39] fused the point cloud with the corresponding image and mapped the lidar and camera features to a birds-eye-view using a 3D backbone. Chiu et al. [13] designed a probabilistic lidar and camera-based multi-object tracking system, and proposed an affinity combined with the Mahalanobis distance [28] and feature distance. Huang et al. [11] proposed the generation of bounding boxes and association scores from cameras and lidar data at the same time, and used a simple multiscale feature fusion scheme to estimate appearance affinities for tracking. Moreover, Gautam et al. [14] proposed the use of the deep learning model for correlation, in combination with the interacting multiple model (IMM) filter for state estimation. Koh et al. [40] proposed the use of the graph neural network (GNN) to associate objects based on spatiotemporal features, and combined rule-based edge pruning and attention-based edge control to improve the tracking performance. In addition, Nabati et al. [41] designed an end-to-end network for 3D MOT based on radar and camera sensor fusion, and they greedily used depth, velocity, and 2D displacement information to associate objects. However, the above methods typically associate all the detections and trajectories based on the similarity in one stage, failing to perform accurate matching between detections and tracklets.

Various methods have been developed for solving this problem. For example, Kim et al. [42] used 3D and 2D detectors to obtain multimodal detections. Then, they tracked the objects based on Kalman and a two-stage data association module. However, this method ignores the effects of low-score detections and track management on tracking performance, and, thus, cannot achieve long-term robust tracking. In addition, Wang et al. [43] proposed a deep association mechanism that establishes both 2D and 3D trajectories, and used the differences between the two types of tracks for tracking. However, this method fails to make full use of tracklet information. In this paper, we propose a better solution, namely, the multi-stage association approach MSA-MOT.

3. Methods

3.1. Overall Framework

We propose a multi-stage association tracker, MSA-MOT, for 3D multi-object tracking. The overall framework is shown in Figure 2, which includes a detection module, hierarchical matching strategy, and customized track management module. Specifically, in the detection module, to obtain multimodal bounding boxes, we use 2D and 3D detectors based on camera and lidar sensors, respectively. After that, the proposed hierarchical matching strategy is used to achieve accurate association, which includes four stages, as shown in Figure 3. First, the 3D bounding boxes are projected to the image dimension, and data matching with 2D detections is performed to judge the reliability of objects. Second, the reliable 3D detections are associated with candidate trajectories. Third, the unreliable 3D detections are associated with the unmatched trajectories. Finally, the remaining trajectories are projected to 2D and associated with the unreliable 2D detections. After these two modules, the customized track management module aims to initialize and update tracks, while focusing on efficiently managing the missing tracks. Specifically, based on the reliability of the tracks provided by the matching module, we set a larger maximum number of missing frames for reliable tracks than for unreliable tracks.
3.2. Hierarchical Matching Module

3.2.1. First Stage of Matching

The first stage aims to prepare for the subsequent stages by determining the reliability of detections. For reliability judgment, a previously proposed method [44] is used to set a threshold based on the confidence provided by the detector. However, for this method, the score must be set manually, which is troublesome in practice. To solve this problem, we note that the objects detected by both 2D and 3D detectors are reliable. As a result, a strategy for reliability judgment based on multimodality information is proposed. Specifically, we obtain a set of 3D detections \(I_{3d}\) and 2D detections \(I_{2d}\) from 3D and 2D detectors, respectively, as expressed in Equations (1)–(3).

\[
I_{3d} = \left\{ I_{3d}^1, I_{3d}^2, \ldots, I_{3d}^n \right\} \tag{1}
\]

\[
I_{3d}^i = \{ x, y, z, \theta, l, w, h \}, \; i \in \{1, 2, \ldots, n\} \tag{2}
\]

\[
I_{2d} = \left\{ I_{2d}^1, I_{2d}^2, \ldots, I_{2d}^m \right\} \tag{3}
\]

where \(n\) and \(m\) denote the numbers of 3D detections and 2D detections, respectively; \((x, y, z)\) denotes the center coordinates of the corresponding 3D objects; \(\theta\) denotes the yaw angle; \(l, w, \) and \(h\) denote the length, width, and height, respectively, of the bounding box. For each 3D bounding box, we replace the polygon box with a precise rectangular box after prediction. \(\hat{T}_{t-1}\) indicates the candidate tracks of the previous frame. \(T_{t-1}\) denotes the trajectories of \(T_{t-1}\) after prediction. \(\hat{1}_{t-T}\) and \(\hat{1}_{t-T}^{un}\) indicate unmatched tracks in the second and third stages, respectively. Moreover, \(T_t\) denotes the output trajectories at the current frame.
projecting to 2D. The projected 2D detection set is represented as $I_{3d}^{proj}$, corresponding to $I_{3d}$. We obtain two sets of 2D detections ($I_{3d}^{proj}$ and $I_{2d}$). Each bounding box in 2D is denoted by $B_{2d}$ as expressed in Equation (4).

$$B_{2d} = (X_1, X_2, Y_1, Y_2)$$

where $(X_1, Y_1)$ and $(X_2, Y_2)$ denote the upper-left and lower-right coordinates, respectively, of the bounding box. Then, we use the Euclidean distance based on $(X_1, X_2, Y_1, Y_2)$ and the greedy algorithm [45] to associate the two sets of 2D detections, as expressed in Equation (5).

$$Output = \text{Greedily} \left( L_2(I_{3d}^{proj}, I_{2d}) \right)$$

where Greedily (·) denotes the greedy algorithm. $L_2$ (·) denotes the Euclidean distance.

Finally, three sets of detections are output for subsequent stages: 3D objects that match with 2D detections, which are considered reliable and denoted $I_{3d}^{rel}$; 3D objects that fail to match, which are referred to as unreliable detections and denoted $I_{3d}^{un}$; unmatched 2D detections, which are denoted $I_{2d}^{un}$.

3.2.2. Second Stage of Matching
To reduce the impact of unreliable 3D detection and, thus, increase the association accuracy, we only associate the candidate trajectories with reliable 3D detections $I_{3d}^{rel}$ in this stage. Specifically, we first use the constant-velocity motion model and Kalman filter to obtain the prediction state $\hat{T}_{t-1}$ of the previous trajectory set $T_{t-1}$. The prediction state is defined in Equation (6):

$$\hat{T}_k^t = \{ \hat{x}, \hat{y}, \hat{z}, \theta, \hat{l}, \hat{w}, \hat{h}, \hat{v}_x, \hat{v}_y, \hat{v}_z \}$$

where $k$ denotes one of the predicted trajectories, $(\hat{x}, \hat{y}, \hat{z})$ denotes the predicted center coordinates, and $\hat{v}_x, \hat{v}_y$, and $\hat{v}_z$ denote the velocity of motion in 3D space. Then, to accurately calculate the similarity between detections and trajectories, we propose a new affinity metric that considers the coordinates and rotation angles of the objects, which is defined in Equation (7):

$$A = L_2(P_{det}, P_{pred}) \times (2 - \cos(\alpha_{det}, \alpha_{pred}))$$

where $L_2$ denotes the Euclidean metric; $P_{det}, P_{pred}$ denote the coordinates of the diagonal of the detection and the track, respectively; $\alpha$ denotes the yaw angle of the corresponding bounding box.

Then, we match the 3D detections with all the candidate tracks based on the proposed affinity and greedy algorithm. Finally, the unmatched trajectories $\hat{T}_{un}^{t-1}$ are used as input for the next stage.

3.2.3. Third Stage of Matching
In this stage, we focus on the detections $I_{3d}^{un}$ output from the first stage. Because their existence is reasonable, we believe that the unreliable detections $I_{3d}^{un}$ still include some real objects, such as heavily obscured objects that can be detected only by the 3D detector. Therefore, considering the components of $I_{3d}^{un}$, we still use it to recover the real objects in unreliable detections. Specifically, we greedily match $I_{3d}^{un}$ with $\hat{T}_{un}^{t-1}$ based on the proposed affinity to further achieve accurate associations. Finally, the tracks that are unmatched again are named $\hat{T}_{un}^{t-1}$.

3.2.4. Fourth Stage of Matching
Finally, to obtain more accurate trajectories, we use the unmatched 2D detections $I_{2d}^{un}$ provided by the first stage to recover the objects that only detected by the camera sensor. Specifically, $\hat{T}_{un}^{t-1}$ failing several times to match in the third stage does not mean that all these tracklets have left the field of view. For example, for objects at long distances, because
the point cloud data are sparse but the image information is relatively rich, it is likely that the 3D detector cannot detect them, but the 2D detector can. Therefore, to make full use of 2D information, we match the unmatched 2D detections $\hat{z}_{2d}$ with $\hat{z}_{1d}$ based on the Euclidean distance and greedy algorithm. Finally, the matching result is used as a new message for those tracks.

### 3.3. Customized Track Management Module

The track management module aims to initialize, update, and delete the candidate tracks. For the track initialization, we set the minimum hitting frames commonly used in 3D MOT to determine whether a new trajectory needs to be initialized. Then, for the update process, we use the 3D detections that match the candidate 3D tracks to update the status of the corresponding trajectories based on Kalman filtering, and the updated tracks are denoted $T_i$, which is defined in Equation (8).

$$T_i^s = \{ \hat{x}, \hat{y}, \hat{z}, \hat{\dot{x}}, \hat{\dot{y}}, \hat{\dot{z}}, \hat{\ddot{x}}, \hat{\ddot{y}}, \hat{\ddot{z}} \}$$

where $s$ represents one of the updated tracks and the variables with “.” correspond to the updated value. For the trajectories that only match with $\hat{z}_{2d}$, considering that it is difficult to accurately transform 2D bounding boxes to the 3D coordinate system, we only use the prediction results obtained by Kalman filtering as the new states of the trajectories and add them into $T_i$. Finally, we delete candidate trajectories based on the reliability of the tracks to reduce the identity switching in the tracking process. We consider two factors that may lead to identity switching as shown in Figure 4: candidate trajectories are deleted prematurely, failing to be re-tracked later, and tracks drift due to predictions that are too long, which cannot be accurately matched again.

**Figure 4.** Illustration of existing problems in track management module. Different colored bounding boxes indicate different objects. If the candidate trajectory is deleted early, a new trajectory will be generated, resulting in identity switching of the object. In addition, if the trajectory is retained for too long, the trajectory will drift and, thus, generate an incorrect association.
To alleviate these problems, we note that objects detected by 3D detectors may exist for a long time, while objects recognized only by camera-based detectors may leave from view more quickly. Therefore, we propose a new strategy based on the information provided by the matching module for reducing identity switches. Specifically, we regard trajectories that match with 3D detections as reliable and, thus, set a large maximum missing frames ($M^F$) for them. A track is deleted when the number of unmatched consecutive frames of the track is higher than $M^F$. In addition, for trajectories that only match 2D detections, we set $M^F$ to smaller values to reduce interference with reliable trajectories, as defined in Equation (9). Experiments show that this strategy can effectively reduce the number of redundant tracks and improve the tracking accuracy.

$$M^F = \begin{cases} H_{\text{Frames}}, \hat{T}_{t-1}^{i} \otimes I_{3d}^m, & \hat{T}_{t-1}^{i} \in \hat{T}_{t-1}^{-1} \\ L_{\text{Frames}}, \hat{T}_{t-1}^{i} \otimes I_{2d}^n, & \hat{T}_{t-1}^{i} \in \hat{T}_{t-1}^{-1} \end{cases}$$  

(9)

where $H_{\text{Frames}}, L_{\text{Frames}}$ represent high and low frames, respectively; $\hat{T}_{t-1}^{i}$ represents a candidate track; $I_{3d}^m, I_{2d}^n$ denote one of the 3D and 2D detections; $\otimes$ denotes successful association.

4. Experiments

4.1. Dataset

We evaluate our method on the challenging KITTI [16] benchmark. KITTI provides both image and point cloud data, collected by a front camera and Velodyne HDL-64E lidar scanner, respectively. The dataset contains 21 training sequences and 29 testing sequences. We follow GNN3DMOT [38] to separate the training sequence. The sub validation set consists of 10 sequences, and the training set consists of 11 sequences. Because the KITTI dataset only provides ground truth labels for the training/validation split, all the ablation experiments in the paper are performed on the validation set.

4.2. Evaluation Metrics

To fully demonstrate the effectiveness of MSA-MOT, we evaluate both the 2D and 3D MOT performances. For 2D MOT, the tracking results are obtained by projecting the 3D bounding boxes to the image plane. We use widely used metrics [46,47] to evaluate the performance of 2D MOT, e.g., HOTA, Association Accuracy (AssA), Multi-Object Tracking Accuracy (MOTA), and Identity Switch (IDSW). For 3D MOT evaluation metrics, we use the scaled Accuracy Multi-Object Tracking Accuracy (sAMOTA) and the Averaged Multi-Object Tracking Accuracy (AMOTA) proposed in AB3DMOT [7]. sAMOTA is defined in Equation (10):

$$s\text{AMOTA} = \frac{1}{L} \sum_{r \in \{0, 0.1, \ldots, 1\}} \text{sMOTA}_r$$  

(10)

$$\text{sMOTA}_r = \frac{\max(0, 1 - \frac{\text{IDS}_r + \text{FP}_r + \text{FN}_r - (1 - r) \times \text{GT}_r}{r \times \text{GT}_r})}{\text{GT}_r}$$  

(11)

where $r$ is the recall value (confidence threshold), $L$ is the number of recall values, and $\text{IDS}_r, \text{FP}_r, \text{FN}_r, \text{GT}_r$ represent the numbers of identity switches, false positives, false negatives, and ground truths, respectively.

4.3. Implementation Details

All the experiments are implemented using Python 3.7 with an Intel Core i5 11400F 2.6 GHz CPU and 16 GB of RAM. We use PointGNN [48] as a 3D detector to predict 3D bounding boxes and use RRC [49] for 2D images. The affinity thresholds for 2D and 3D are set to 25 and 7, respectively. For tracklets with high reliability, the maximum number of missing frames is 11, and the maximum number of missing frames for the remaining tracklets is set to 3. In the 3D MOT performance evaluation, we set $\text{IoU} = 0.25$ to conduct a fair comparison with other works.
4.4. Comparison with the State-of-the-Art Methods

4.4.1. Quantitative Comparison

2D MOT: Table 1 shows the comparison results of the proposed method for the car class of the KITTI MOT testing set. To demonstrate the effectiveness of our method, we compare MSA-MOT with current state-of-the-art methods (including single modality-based and multimodality-based methods). The results show that the proposed method achieves the highest HOTA (78.52%) and AssA (82.56%) among the compared methods. It also shows superior performance in terms of the remaining metrics (MOTA, IDSW, and FPS), reaching 88.01%, 91, and 130, respectively. Specifically, MSA-MOT outperforms DeepFusionMOT, which uses the same 2D detector as our method, by significant margins, namely, 3.06% and 3.37% for HOTA and MOTA, respectively. In addition, compared to EagerMOT with the same 2D and 3D detectors, our method realizes 4.13% and 8.4% improvements in HOTA and AssA, respectively, due to the effectiveness of our multi-stage association. Notably, our method achieves significant improvements in the AssA metric, which indicates the accuracy of data association. This is mainly attributed to the proposed hierarchical module. In addition, due to the customized track management module, the IDSW value is much smaller than those of most state-of-the-art methods. Moreover, our method balances accuracy and speed.

Table 1. Comparison results on the KITTI tracking testing benchmark for the car class. A higher value in the table indicates better performance without IDSW. Bold font indicates the best performance among all the compared methods, and the values that correspond to the second-best performance are underlined. “#” indicates that the method uses the same 3D/2D detector as our method.

| Method             | Publication | Input        | HOTA (%) | AssA (%) | MOTA (%) | IDSW | FPS |
|--------------------|-------------|--------------|----------|----------|----------|------|-----|
| BeyondPixels [50]  | ICRA’18     | 2D + 3D      | 63.75    | 56.40    | 82.68    | 934  | 3.3 |
| mmMOT [15]         | ICCV’19     | 2D + 3D      | 62.05    | 54.02    | 83.23    | 733  | 50  |
| monos3DT [51]      | ICCV’19     | 2D           | 73.16    | 74.18    | 84.28    | 379  | 33.3|
| AB3DMOT [7]        | IROS’20     | 3D           | 69.99    | 69.33    | 83.61    | 113  | 212 |
| MOTS Fusion [52]   | RA-L’20     | 2D + 3D      | 68.74    | 66.16    | 84.24    | 415  | 2.3 |
| JRMOT [12]         | IROS’20     | 2D + 3D      | 69.61    | 66.89    | 76.95    | 271  | 20  |
| CenterTrack [30]   | ECCV’20     | 2D           | 73.02    | 71.20    | 88.83    | 254  | 22.2|
| Quasi-Dense [53]   | CVPR’21     | 2D           | 68.45    | 65.49    | 84.93    | 313  | 14.3|
| LGM [54]           | ICCV’21     | 2D           | 73.14    | 72.31    | 87.60    | 448  | 12.5|
| JMODT [11]         | IROS’21     | 2D + 3D      | 70.73    | 68.76    | 85.35    | 350  | 100 |
| EagerMOT [42] #    | ICRA’21     | 2D + 3D      | 74.39    | 74.16    | 87.82    | 239  | 90  |
| TripletTrack [37]  | CVPRW’22    | 2D           | 73.58    | 74.66    | 84.32    | 322  | -   |
| QD-3DT [9]         | TPAMI’22    | 2D           | 72.77    | 72.19    | 85.94    | 206  | 45  |
| PolarMOT [33] #    | ECCV’22     | 3D           | 75.16    | 76.95    | 85.0     | 462  | 170 |
| DeepFusionMOT [43] | RA-L’22     | 2D + 3D      | 75.46    | 80.06    | 84.64    | 84   | 110 |
| DetecTrack [40] #  | AAAI’22     | 2D + 3D      | 73.54    | 75.25    | 85.52    | -    | 27  |
| MSA-MOT            | Ours        | 2D + 3D      | **78.52**| **82.56**| **88.01**| 91   | 130 |

3D MOT: To further evaluate our method, we compare the performance of 3D MOT on the KITTI validation set for the car class. As shown in Table 2, MSA-MOT outperforms the previous state-of-the-art methods, achieving the highest sAMOTA (97.11%), AMOTA (50.10%), and MOTA (96.83%). Specifically, MSA-MOT outperforms the lidar-based PolarTrack, which uses the same 3D detector, by 2.79% and 2.9% in terms of sAMOTA and MOTA, respectively. In addition, compared to DeepFusionMOT, which uses the same 2D detector, the proposed method achieves remarkable improvements (approximately 5%) in the key metrics sAMOTA, AMOTA, and MOTA. Moreover, compared to EagerMOT with the same 2D and 3D detectors, sAMOTA and AMOTA are improved by 2.17% and 1.26%, respectively. Moreover, our method demonstrates significant superiority compared to multimodality methods such as mmMOT, GNN3DMOT, and DetecTrack. All performance advantages are due to the proposed multi-stage association method.
Table 2. Comparison results of 3D MOT on the KITTI tracking validation set for the car class. The bold font indicates the best performance among all the compared methods.

| Method         | Publication | Modality | mAMOTA (%) | AMOTA (%) | MOTA (%) | IDS |
|----------------|-------------|----------|-------------|-----------|----------|-----|
| FANTrack       | IV’19       | 3D + 2D  | 82.97       | 40.03     | 74.30    | 35  |
| mmMOT          | ICCV’19     | 3D + 2D  | 70.61       | 33.08     | 74.07    | 10  |
| AB3DMOT        | IROS’20     | 3D       | 93.28       | 45.43     | 86.24    | 0   |
| GNN3DMOT       | CVPR’20     | 3D + 2D  | 93.68       | 45.27     | 84.70    | 0   |
| EagerMOT#      | ICRA’20     | 3D + 2D  | 94.94       | 48.84     | 96.61    | 2   |
| PC-TCNN        | IJCAI’21    | 3D + 2D  | 95.44       | 47.64     | -        | 1   |
| PolarTrack#    | ECCV’22     | 3D       | 94.32       | -         | 93.93    | 31  |
| DetecTrack     | AAAI’22     | 3D + 2D  | 96.49       | 48.87     | 91.46    | -   |
| DeepFusionMOT# | RA-L’22     | 3D + 2D  | 91.80       | 44.62     | 91.30    | 1   |
| MSA-MOT        | Ours        | 3D + 2D  | **97.11**   | **50.10** | **96.83**| 0   |

4.4.2. Qualitative Comparison

To visually demonstrate the superiority of the proposed method, we perform a qualititative comparison on the KITTI dataset. EagerMOT is a classic method for 3D MOT and uses the same 2D and 3D detectors as our method, so we choose it as the method for comparison. Figure 5 shows the visualization results of EagerMOT and MSA-MOT in the lidar coordinate system. The left and right columns represent the visualizations corresponding to EagerMOT and our method, respectively. Specifically, in sequence 0002, the object with ID 43 in Figure 5a changes its ID in the 158th frame. However, our method achieves accurate tracking of the car. In addition, in sequence 0010, due to the severe obscuration by other objects, the car with ID 14 is incorrectly tracked many times by EagerMOT. However, our method always performs precise tracking. In sequence 0017, the vehicle with ID 11 drives far away, and its surrounding light is bright. It is difficult to track the object accurately under these conditions, but our tracker still achieves this. The results show that our method can achieve robustly track dynamically changing objects.

4.4.3. Ablation Study

To further demonstrate the effectiveness of our method, we choose three frames for the above sequences in the image dimension, as shown in Figure 6. The upper and lower parts of each sequence represent the visualization results of EagerMOT and MSA-MOT, respectively. Specifically, in sequence 0002, due to the change in motion stage and long-term occlusion, the vehicle with ID 43 obtained by EagerMOT changes its ID in the 158th frame. However, our method achieves accurate tracking of the car. In addition, in sequence 0010, due to the severe obscuration by other objects, the car with ID 14 is incorrectly tracked many times by EagerMOT. However, our method always performs precise tracking. In addition, in sequence 0017, due to the proposed hierarchical matching module reducing the loss of real objects caused by the low-scoring detections, our method accurately tracks the car with ID 11 while it is driving away.

Figure 5. Cont.
challenging to track objects accurately when they are at a distance. However, in sequence 0010, due to the severe obscuration by other objects, the car with ID 14 is incorrectly tracked by EagerMOT. However, our method achieves accurate tracking of the car. In addition, in sequence 0017, the vehicle with ID 11 drives far away, and its surrounding light is bright. It is difficult to track the object accurately under these conditions, but our tracker still achieves this. The results show that our method can achieve superior tracking performance under challenging conditions, such as long-term occlusion, sudden changes in the motion state, and objects at a distance.

To further demonstrate the effectiveness of our method, we choose three frames for the above sequences in the image dimension, as shown in Figure 6. The upper and lower parts of each sequence represent the visualization results of EagerMOT and MSA-MOT, respectively. Specifically, in sequence 0002, due to the change in motion stage and long-term occlusion, the vehicle with ID 43 obtained by EagerMOT changes its ID in the 158th frame. However, our method achieves accurate tracking of the car. In addition, in sequence 0010, due to the severe obscuration by other objects, the car with ID 14 is incorrectly tracked many times by EagerMOT. However, our method always performs precise tracking. In addition, in sequence 0017, the vehicle with ID 11 drives far away, and its surrounding light is bright. It is difficult to track the object accurately under these conditions, but our tracker still achieves this. The results show that our method can achieve superior tracking performance under challenging conditions, such as long-term occlusion, sudden changes in the motion state, and objects at a distance.
To demonstrate the effectiveness of the proposed components, i.e., the hierarchical matching module and customized track management module, and investigate their contributions to the tracking performance, we conduct an ablation study on the KITTI validation set. For a fair analysis, we still choose EagerMOT (using the same 2D and 3D detectors) as the comparison method, as shown in Table 3. Due to the utilization of positive objects with low reliability in our hierarchical matching module, a 0.99% performance improvement in HOTA is realized, along with a 2.58% improvement in AssA. In addition, by combining all the modules, the performance is further enhanced, which indicates that the customized track management approach can improve the association accuracy and reduce the number of identity switches.

4.5. Ablation Experiments

4.5.1. Component-Wise Analysis

Table 3. Ablation study results on the validation set for the car class, using the hierarchical matching strategy (MSM) and customized track management module (CTM). Because the proposed CTM relies on the results of the matching module, it is not ablated separately. Bold font indicates the best performance.

|        | MSM | CTM  | HOTA (%) | DetA (%) | AssA (%) | MOTA (%) | IDSW |
|--------|-----|------|----------|----------|----------|----------|------|
| EagerMOT |     |      | 78.04    | 76.80    | 79.51    | 87.25    | 91   |
| Ours   | √   |      | 79.03    | 77.39    | 80.90    | 88.23    | 66   |
|        | √   | √    | 79.73    | 77.50    | 82.09    | 88.49    | 46   |

4.5.2. Hierarchical Matching Module

The affinity used to evaluate the similarity between the detections and trajectories has a crucial influence on the tracking performance. To further demonstrate the effectiveness
of the proposed hierarchical matching strategy, we perform an ablation analysis with EagerMOT under the commonly used 3D intersection over union (3D-IoU) [7] and 3D generalized intersection over union (3D GIoU) [55] metrics, as shown in Table 4. The results show that our method outperforms EagerMOT with the same detector in terms of different metrics. Moreover, the affinity metric proposed in this paper outperforms the commonly used affinity metric.

Table 4. Ablation study results for hierarchical matching strategy on the KITTI validation set for the car class. Eager denotes EagerMOT using the same 2D and 3D detectors. Bold font indicates the best performance.

| Affinity  | HOTA (%) | AssA (%) | MOTA (%) | IDSW |
|-----------|----------|----------|----------|------|
|           | Ours     | Eager    | Ours     | Eager | Ours | Eager | Ours | Eager |
| 3D-IoU    | 78.83    | 77.16    | 80.34    | 77.76 | 88.00 | 86.67 | 206  | 234   |
| 3D-GIoU   | 79.41    | 77.82    | 81.50    | 79.08 | 88.45 | 87.09 | 95   | 126   |
| Ours      | 79.73    | 78.12    | 82.09    | 79.67 | 88.49 | 87.27 | 46   | 79    |

4.5.3. Track Management Module

In addition, we ablate the maximum number of missing frames on the KITTI validation set for the car class, as shown in Table 5. When the maximum missing frames is set to 5, the module shows suboptimal performance, which is caused by tracks being deleted prematurely. A larger value results in many redundant candidate tracklets; thus, the performance shows a decreasing trend. Notably, the best performance is achieved when the maximum missing frames is set to 11. If not specified, this value is set to 11 for all experiments in this paper.

Table 5. Ablation study for the maximum number of missing frames on the KITTI validation set for the car class. Bold font indicates the best performance.

| Frames | HOTA (%) | AssA (%) | MOTA (%) | IDSW |
|--------|----------|----------|----------|------|
| 5      | 79.06    | 80.89    | 88.40    | 56   |
| 8      | 79.24    | 81.23    | 88.48    | 50   |
| 11     | 79.73    | 82.09    | 88.49    | 46   |
| 14     | 79.65    | 82.06    | 88.48    | 46   |
| 17     | 79.64    | 82.02    | 88.45    | 46   |

4.6. Exploration Analysis

To further demonstrate the superiority of our method, we conduct an exploratory analysis. As the tracking performance of AB3DMOT differs among IoU threshold values (used to calculate the similarity between the tracking results and ground truths), we set different IoU threshold (IoU_{thres}) values, as in AB3DMOT, to verify the robustness of our method, as shown in Table 6. Compared to the two classic methods, our method almost ranks first in terms of various metrics. Moreover, although accurate tracking of pedestrians is generally challenging due to the dense bounding boxes, our method still obtains the best performance in terms of the two keys metrics (sAMOTA and AMOTA).
Table 6. Comparison of 3D MOT performance on the KITTI validation set with different IoU thresholds for the car and pedestrian classes. Bold font indicates the best performance.

| Method     | Criteria   | SAMOTA (%) | AMOTA (%) | MOTA (%) |
|------------|------------|------------|-----------|----------|
|            |            | Car        | Pedestrian| Car      | Pedestrian| Car      | Pedestrian|
| AB3DMOT    | IoU thres = 0.25 | 93.28      | 75.85     | 45.43    | 31.04     | 86.24    | 70.90     |
|            | IoU thres = 0.5  | 90.38      | 70.95     | 42.79    | 27.31     | 84.02    | 65.06     |
|            | IoU thres = 0.7  | 69.81      | -         | 27.26    | -         | 57.06    | -         |
| EagerMOT   | IoU thres = 0.25 | 94.94      | 92.95     | 48.84    | 45.96     | 96.61    | 93.14     |
|            | IoU thres = 0.5  | 95.42      | 90.57     | 48.93    | 43.79     | 94.67    | 90.66     |
|            | IoU thres = 0.7  | 85.13      | 64.49     | 39.06    | 21.91     | 84.04    | 64.67     |
| MSA-MOT    | IoU thres = 0.25 | 97.11      | 93.61     | 50.10    | 46.31     | 96.83    | 94.63     |
|            | IoU thres = 0.5  | 96.99      | 91.92     | 49.85    | 44.01     | 94.84    | 91.29     |
|            | IoU thres = 0.7  | 86.85      | 66.77     | 39.90    | 23.58     | 84.16    | 64.60     |

In addition, because the performance of the tracker is partly dependent on the detector, we compare the performance of our method under different 3D detectors. As shown in Table 7, we use three 3D detectors (Point-GNN [48], PointRCNN [24], and PV-RCNN [56]) that are widely applied in 3D MOT tasks. Experimental results show that our method still achieves superior performance with different detectors, proving that MSA-MOT has wide applicability and can be combined with various 3D detectors for robust tracking.

Table 7. Ablation results on the 3D detector for the car and pedestrian classes. Bold font indicates the best performance.

| 3D Detector | SAMOTA (%) | MOTA (%) | IDs | SAMOTA (%) | MOTA (%) | IDs |
|-------------|------------|----------|-----|------------|----------|-----|
| Point-GNN   | 97.21      | 96.68    | 0   | 94.11      | 94.47    | 17  |
| PointRCNN   | 97.18      | 95.44    | 0   | 81.37      | 81.95    | 2   |
| PV-RCNN     | 94.56      | 95.54    | 0   | 94.63      | 95.18    | 6   |

5. Conclusions

In this paper, we propose a novel multi-stage association framework for 3D multimodality multi-object tracking. In this framework, a hierarchical matching module is proposed to improve the utilization of true objects by matching detections and tracks in sequence, which achieves accurate association between detections and tracks. Then, based on the track reliability output from the matching module, the proposed customized track management sets larger maximum missing frames for reliable trajectories than for unreliable tracks, which further improves the association accuracy. Comprehensive experiments are conducted on the challenging KITTI benchmark. Our method achieves state-of-the-art performance among the competitive approaches. Abundant ablation experiments further demonstrate the effectiveness of the proposed modules.

MSA-MOT follows the commonly used constant-velocity motion model to predict the state of the object, where it is difficult to deal with the complex change of state. Therefore, a more accurate state estimation method is worth exploring in the future work.

Author Contributions: Conceptualization, Z.Z.; methodology, Z.Z.; software, Z.Z.; validation, J.N. and H.W.; formal analysis, Z.Z., J.N. and H.W.; investigation, J.N.; data curation, Z.Z.; writing—original draft preparation, Z.Z.; writing—review and editing, Z.Z., J.N. and H.W.; visualization, Z.Z.; supervision, H.W.; project administration, Z.H.; funding acquisition, Z.H. and M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Hangzhou Major Science and Technology Innovation Project of China under Grant 2022AIZD0022, and the Zhejiang Provincial Key Lab of Equipment Electronics.
Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The publicly available dataset used in this paper can be found here: http://www.cvlibs.net/datasets/kitti/eval_tracking.php (accessed on 10 September 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nie, J.; He, Z.; Yang, Y.; Gao, M.; Dong, Z. Learning Localization-aware Target Confidence for Siamese Visual Tracking. IEEE Trans. Multimed. 2022, early access.

2. Nie, J.; Wu, H.; He, Z.; Gao, M.; Dong, Z. Spreading Fine-grained Prior Knowledge for Accurate Tracking. IEEE Trans. Circuits Syst. Video Technol. 2022, 32, 6186–6199. [CrossRef]

3. Wojke, N.; Bewley, A.; Paulus, D. Simple Online and Realtime Tracking with a Deep Association Metric. In Proceedings of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017; pp. 3645–3649.

4. Bae, S.-H.; Yoon, K.-J. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 40, 595–610. [CrossRef] [PubMed]

5. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple Online and Realtime Tracking. In Proceedings of the IEEE International Conference on Image Processing, Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468.

6. Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; Bennamoun, M. Deep learning for 3D point clouds: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 2020, 43, 4338–4364. [CrossRef]

7. Weng, X.; Wang, J.; Held, D.; Kitani, K. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, NV, USA, 25–29 October 2020; pp. 10359–10366.

8. Pischmann, J.; Pfeifer, T.; Protzel, P. Factor Graph Based 3D Multi-Object Tracking in Point Clouds. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, NV, USA, 25–29 October 2020; pp. 10343–10350.

9. Hu, H.-N.; Yang, Y.-H.; Fischer, T.; Darrell, T.; Yu, F.; Sun, M. Monocular quasi-dense 3D object tracking. IEEE Trans. Pattern Anal. Mach. Intell. 2022; early access.

10. Luo, C.; Yang, X.; Yuille, A. Exploring Simple 3d Multi-Object Tracking for Autonomous Driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10488–10497.

11. Huang, K.; Hao, Q. Joint Multi-Object Detection and Tracking with Camera-LiDAR Fusion for Autonomous Driving. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Prague, Czech Republic, 28–30 September 2021; pp. 6983–6989.

12. Shenoi, A.; Patel, M.; Gwak, J.; Goebel, P.; Sadeghian, A.; Rezatofighi, H.; Martín-Martín, R.; Savarese, S. Jrmot: A Real-Time 3D Multi-Object Tracker and a New Large-Scale Dataset. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, NV, USA, 25–29 October 2020; pp. 10335–10342.

13. Chiu, H.-k.; Li, J.; Ambroš, R.; Bohg, J. Probabilistic 3D Multi-Modal, Multi-Object Tracking for Autonomous Driving. In Proceedings of the IEEE International Conference on Robotics and Automation, Xi’an, China, 30 May–5 June 2021; pp. 14227–14233.

14. Gautam, S.; Meyer, G.P.; Vallespi-Gonzalez, C.; Becker, B.C. SDVTracker: Real-Time Multi-Sensor Association and Tracking for Self-Driving Vehicles. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 3012–3021.

15. Zhang, W.; Zhou, H.; Sun, S.; Wang, Z.; Shi, J.; Loy, C.C. Robust Multi-Modality Multi-Object Tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 2365–2374.

16. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. Int. J. Rob. Res. 2013, 32, 1231–1237. [CrossRef]

17. Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. Fairmot: On the Fairness of Detection and Re-Identification in Multiple Object tracking. Int. J. Comput. Vision 2021, 129, 3069–3087. [CrossRef]

18. Wu, H.; Nie, J.; He, Z.; Zhu, Z.; Gao, M. One-Shot Multiple Object Tracking in UAV Videos Using Task-Specific Fine-Grained Features. Remote Sens. 2022, 14, 3853. [CrossRef]

19. Wu, H.; Nie, J.; Zhu, Z.; He, Z.; Gao, M. Leveraging temporal-aware fine-grained features for robust multiple object tracking. J. Supercomput. 2022; early access.

20. Kalman, R.E. A New Approach to Linear Filtering and Prediction Problems. J. Fluids Eng. 1960, 82, 35–45. [CrossRef]

21. Wu, J.; Cao, J.; Song, L.; Wang, Y.; Yang, M.; Yuan, J. Track to Detect and Segment: An Online Multi-Object Tracker. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12352–12361.

22. Pang, B.; Li, Y.; Zhang, Y.; Li, M.; Lu, C. Tubetk: Adopting Tubes to Track Multi-Object in a One-Step Training Model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6308–6318.

23. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. Pointpillars: Fast Encoders for Object Detection from Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12697–12705.
51. Hu, H.-N.; Cai, Q.-Z.; Wang, D.; Lin, J.; Sun, M.; Krahenbuhl, P.; Darrell, T.; Yu, F. Joint Monocular 3D Vehicle Detection and Tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 5390–5399.

52. Luiten, J.; Fischer, T.; Leibe, B. Track to reconstruct and reconstruct to track. IEEE Robot. Autom. Lett. 2020, 5, 1803–1810. [CrossRef]

53. Pang, J.; Qiu, L.; Li, X.; Chen, H.; Li, Q.; Darrell, T.; Yu, F. Quasi-Dense Similarity Learning for Multiple Object Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 164–173.

54. Wang, G.; Gu, R.; Liu, Z.; Hu, W.; Song, M.; Hwang, J.-N. Track without Appearance: Learn Box and Tracklet Embedding with Local and Global Motion Patterns for Vehicle Tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 9876–9886.

55. Pang, Z.; Li, Z.; Wang, N. SimpleTrack: Understanding and Rethinking 3D Multi-object Tracking. arXiv 2021, arXiv:2111.09621.

56. Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; Li, H. Pv-rcnn: Point-Voxel Feature Set Abstraction for 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10529–10538.