




## Article

# 3D-DIoU: 3D Distance Intersection over Union for Multi-Object Tracking in Point Cloud

Sazan Ali Kamal Mohammed <sup>1,2</sup> , Mohd Zulhakimi Ab Razak <sup>1,\*</sup>  and Abdul Hadi Abd Rahman <sup>3</sup> <sup>1</sup> Institute of Microengineering and Nanoelectronics (IMEN), Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia<sup>2</sup> Department of Automotive Technology, Erbil Technology College, Erbil Polytechnic University, Erbil 44001, Iraq<sup>3</sup> Center for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia

\* Correspondence: zul.hakimi@ukm.edu.my

**Abstract:** Multi-object tracking (MOT) is a prominent and important study in point cloud processing and computer vision. The main objective of MOT is to predict full tracklets of several objects in point cloud. Occlusion and similar objects are two common problems that reduce the algorithm's performance throughout the tracking phase. The tracking performance of current MOT techniques, which adopt the 'tracking-by-detection' paradigm, is degrading, as evidenced by increasing numbers of identification (ID) switch and tracking drifts because it is difficult to perfectly predict the location of objects in complex scenes that are unable to track. Since the occluded object may have been visible in former frames, we manipulated the speed and location position of the object in the previous frames in order to guess where the occluded object might have been. In this paper, we employed a unique intersection over union (IoU) method in three-dimension (3D) planes, namely a distance IoU non-maximum suppression (DIoU-NMS) to accurately detect objects, and consequently we use 3D-DIoU for an object association process in order to increase tracking robustness and speed. By using a hybrid 3D DIoU-NMS and 3D-DIoU method, the tracking speed improved significantly. Experimental findings on the Waymo Open Dataset and nuScenes dataset, demonstrate that our multistage data association and tracking technique has clear benefits over previously developed algorithms in terms of tracking accuracy. In comparison with other 3D MOT tracking methods, our proposed approach demonstrates significant enhancement in tracking performances.



**Citation:** Mohammed, S.A.K.; Razak, M.Z.A.; Rahman, A.H.A. 3D-DIoU: 3D Distance Intersection over Union for Multi-Object Tracking in Point Cloud. *Sensors* **2023**, *23*, 3390.

<https://doi.org/10.3390/s23073390>

Academic Editor: Cosimo Distanto

Received: 23 December 2022

Revised: 5 March 2023

Accepted: 8 March 2023

Published: 23 March 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** multi-object tracking; point cloud; 3D-DIoU; DIoU-NMS; multistage data association; tracklets; motion prediction

## 1. Introduction

An important challenge in computer vision study is multi-object tracking (MOT), which identifies and keeps track a unique identification (ID) for each object of interest in a point cloud series while predicting the locations of all objects. MOT has many important theoretical research implications and practical applications. Systems for visual security surveillance, vehicle visual navigation [1], augmented reality [2], human–computer interface, high sensitivity audio-visual (AV) [3] to name a few, all heavily rely on MOT systems with well-behaved performances. There are several difficulties that can deteriorate tracking performances in real-world applications. These difficulties include the way an object interacts, occlusion, and how close certain objects are related to one another. These difficulties lead to many unwanted detection mistakes and errors, including bounding box drift and ID changes, which cause tracking performance to degrade severely. As a result, this work proposes an improved and reliable MOT method for point cloud scenarios. Previously developed three dimension (3D) multiple object tracking (3D MOT) algorithms [4–9] adopt the tracking-by-detection pattern. Across frames, the tracklets depend directly on the 3D bounding boxes from 3D detectors.

In general, the concept of the tracking-by-detection algorithm consists of four modules: (i) input detection pre-processing module, (ii) motion module, (iii) association module, and (iv) managing tracklet life cycle. All objects of interest from the point cloud series are determined using the detector. Then, the identical objects from the detector and predicted motion model are associated using the metrics, which is established on features. A continually updating tracklet set is created by connecting the same item in many point cloud frames. In this procedure, the detector's effectiveness and the performance of the data association algorithm jointly impact the tracking accuracy and flexibility. This detection process is normally evaluated using an intersection over union (IoU) metric.

The association results can be wrong when the input detectors are inaccurate. However, refining these detectors by using the non-maximum suppression (NMS) technique can improve the association. Additionally, we found that the association metric expressed between two 3D bounding boxes should be designed properly. Neither generalized IoU or GIoU [10] nor L2 [11] work well. The inference speed of the tracking system is significantly influenced by both the detector and the data association. Therefore, the multistage association process between predictions and tracklets can express the existence of the objects. Based on these findings, using distance-IoU (DIOU) over the tracking pipeline can significantly improve the solutions. In order to tackle 3D MOT issues, we propose an improved DIOU method in this paper. Consequently, we utilize Waymo Open Dataset [12] and nuScenes [13] in order to evaluate and verify our proposed algorithm. Our method and contributions, in brief, are as follows:

- We added DIOU-NMS to 3D MOT tracking pipeline and analyze the performance;
- We proposed the use of DIOU for two-stage and multi-stage data association, which showed competitive results on both Waymo Open Dataset and nuScenes;
- We used unmatched tracklets and unmatched detection from previous stages for data association in the next stage, and the verification results on Waymo Open Dataset show better performance for cyclist objects.

By using DIOU in the tracking pipeline, we overcome premature tracklet termination where the tracking framework depends on prediction position for invisible objects. Previous work in [14] used GIoU for the tracking process, which terminates an unassociated tracklet. Instead, we used DIOU to maintain the unassociated tracklet by using its predicted position. Therefore, when a temporarily invisible object reappears, it can be associated with its original predicted position.

## 2. Related Work

Point clouds are used in 3D multi-object tracking (MOT), which works in conjunction with the detection process of 3D objects in the autonomous driving challenge. The task of connecting objects together in a complete sequence is handled by 3D MOT, which is sensible of object location in all point cloud frames. In this process, temporal consistency is vital in addressing the tracking issue. The difference between a 3D MOT system and a 2D MOT system is that 3D MOT system uses a 3D space for the detecting procedure. Recent studies have been using 3D point cloud data for MOT applications, even without the use of extra features such as RGB data.

### 2.1. Two-Dimensional MOT Methods

Based on data association, recent 2D MOT systems can be categorized into batch and online techniques. The batch approach utilizes a full sequence search for a global optimum association. Meanwhile, authors in [15] proposed a TADT system in order to learn target-aware features for better recognition of the targets under variance appearance changes. While trackers use a maximum overlap technique based on IoU values to solve this concern, there are imperfections in the IoU values that make it impossible to continuously optimize the objective function when a provided bounding box is completely contained within or without another bounding box; this makes accurate estimation of the target state extremely difficult.

Meanwhile, authors in [16] designed a tracking method based on a distance-IoU (DIOU) loss for an estimation and classification of a target. Learning to track procedures used in many fields, a method in [17] employed MOT for guiding drones and controlling them. Correspondingly, an MOT is used in unmanned aerial vehicles (UAVs) for collision prevention in [18]. Detecting lanes in driver assistance systems is challenging under bad climatic changes, hence a method [19] introduced a two-tier deep learning-based lane detection system for many images at a different number of weather situations. Texture features are extracted and an optimized deep convolution network is used for road and lane classification. On the other hand, authors in [20] proposed tracking algorithms that measure the dynamic similarity of tracklets and recover missing data due to long occlusions due to motion dynamics that provide strong cues while tracking targets with identical or very similar appearance. However, the algorithms are limited to 2D objects.

However, detections from the previous frame and the current frame are accessible to the tracker, the SORT method [21] proposed online tracking-by-detection, the tracker component runs at 260 Hz for updating its states which is useful for real-time applications. In contrast, our work uses 3D point cloud data, and the system works near real-time with 10 Hz on Waymo Open Dataset and 2Hz on nuScenes.

Several online 2D trackers [22–25] have suggested improved detection qualities and utilize the tracking-by-detection paradigm. Unfortunately, due to scale variance, the object items in RGB pictures vary in size, making association and motion models more difficult. However, 2D MOT may readily make use of rich RGB information and employ appearance models based on online learning process [24,26–28]. The design of MOT frameworks should be compatible with data taken from LiDAR or a camera.

## 2.2. Three-Dimensional MOT Methods

Previous research in 3D multi-object tracking that followed the tracking-by-detection paradigm often solved tracking problems by using a bipartite graph-matching mechanism on top of ordinary detectors. Depending on early works on 2D MOT [13–15], many strategies emphasize enhancing the relationship between detection and tracklets by simulating their movements or attendance, or a mix of the two. A state of Kalman filter in [29] is specified on the 2D plane. An AB3DMOT method [7] offers a baseline technique built on the PointRCNN detector [30] that combines the 3D Kalman filter with the Hungarian algorithm [31]. While AB3DMOT utilized 3D IoU for the association process, Chiu et al [32] employed Mahalanobis distance [10] as an alternative. On the other hand, GIoU [33] is used in Simple Track [14] for the 3D association. The 2D velocity of the detected box is predicted by learning in CenterPoint [8] followed by CenterTrack [34] and performed simple point–distance corresponding. In addition, an aspect of labeling for 3D objects in self-driving vehicles are discussed in [35].

To further prevent misperception throughout the association procedure, GNN3DMOT [36] used a Graph Neural Network to collect appearance and motion data in order to establish feature interaction between objects. Authors in [37] proposed a probabilistic multi-modal structure that covered trainable sections for 2D and 3D object feature fusion, distance space arrangement, and trajectory creation. A method in [37] joined 2D and 3D object indication gained from 2D and 3D detectors. The authors in a single graph form combined the prediction models with object identification characteristics [9]. In this paper, on the other hand, we used a simple DIOU metric for data association.

Behind early knowledge in 2D MOT [38–40], the preceding works in 3D MOT [5,7–9,24,26] frequently implemented a counting-based method for tracklet life cycle management. Different tracklets are created for each frame with detected objects that are not related to any current tracklet. The tracklets that lose their targets for a number of frames (usually fewer than 5) are terminated. Authors in [4,27] recommended that tracks are initiated and terminated based on their confidence score value, which is calculated from the confidence measurement of their related detections. Nevertheless, predictors that are not related with fresh detections are permanently terminated. In contrast, we show that by positively anticipating and conserving

the object-predicted box, predictors that have lost their targets may be appropriately preserved for future association.

### 3. Materials and Methods

In this section, a simple tracking with the PointPillars and a motion prediction technique is proposed, and the workflow of tracking procedure is shown in Figure 1. The tracking process consists of the following parts:

1. Detection: for this step, the bounding boxes are selected from the detector, as shown in Figure 2;
2. Selected Detection: by applying NMS process the number of bounding boxes decreased and the unwanted boxes are removed;
3. Tracklets, Prediction, and Motion Update: all these processes are related to each other where Kalman filter is used, as illustrated in Figures 3 and 4;
4. Multi-Stage Association: in this step the detectors in the present frame are associated with the tracklets from the previous frame. The unmatched prediction and tracklets are associated in another stage. Three-dimensional GIoU and DIoU association metrics are used in this work is coupled with Hungarian algorithm;
5. Motion update and Life Cycle Management: the creation and termination of the tracklets are updated and are determined in this step, and the final tracklets are shown in Figure 5.

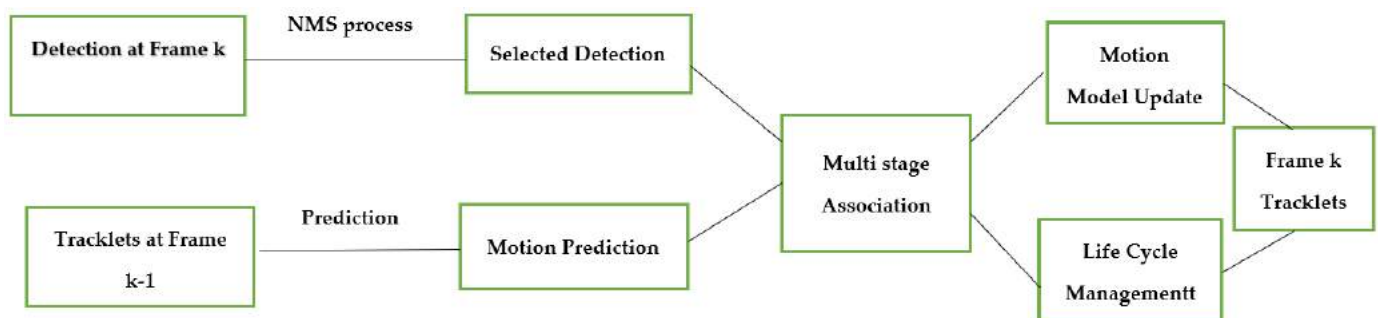


Figure 1. Three-dimensional MOT workflow steps.

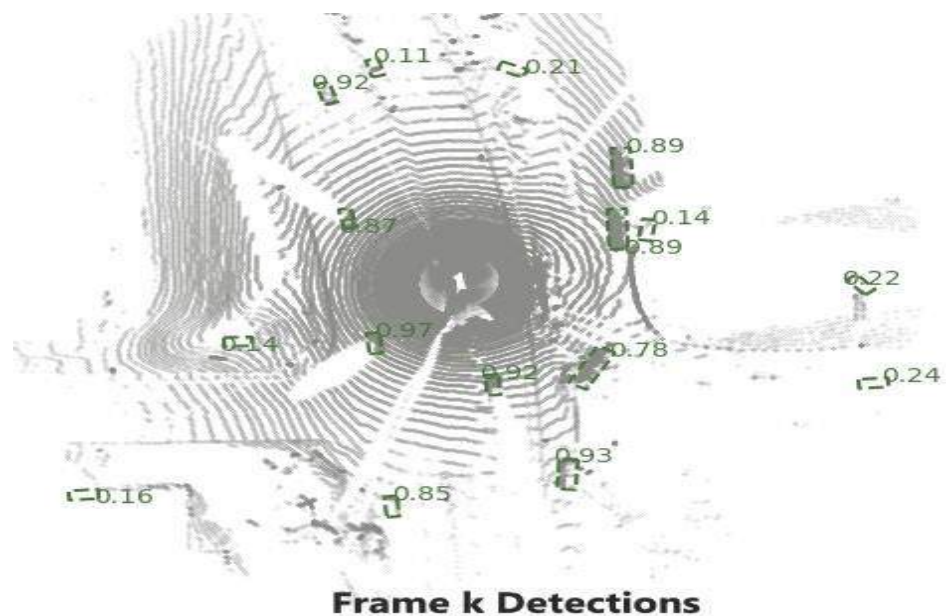


Figure 2. Vehicle bounding box detectors.



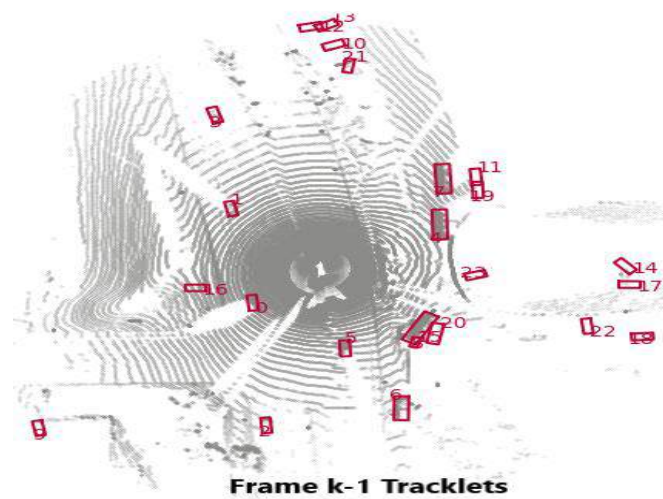


Figure 3. Vehicle tracklets bounding boxes.

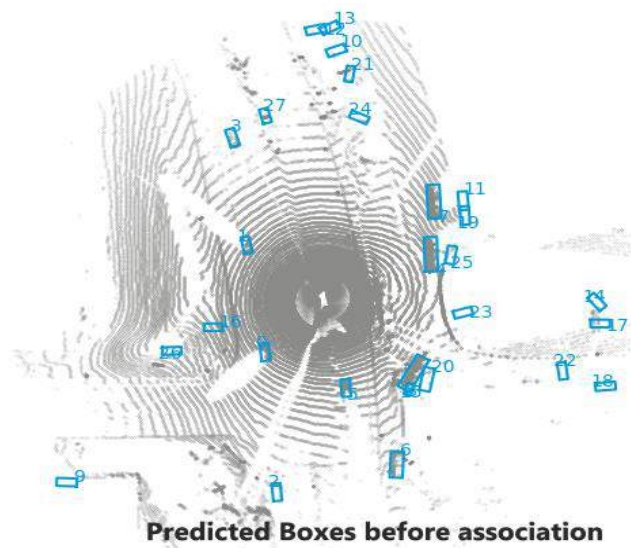


Figure 4. Motion prediction process.

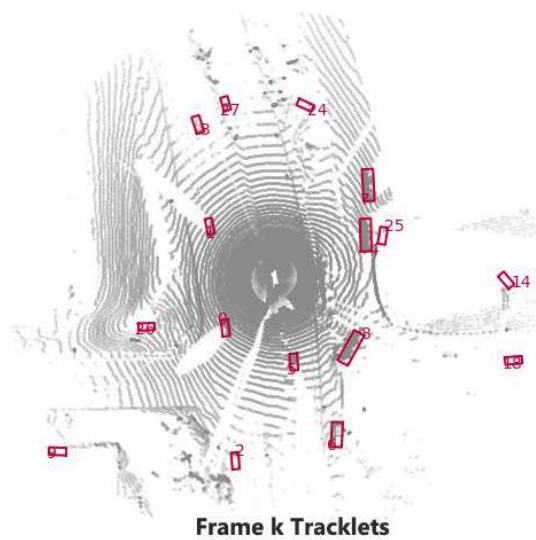


Figure 5. Tracklets at frame k.

### 3.1. Adding 3D DIoU Function

In this part, we examine and enhance the detection and multi-stage association modules by including a 3D DIoU model. In this work, we revised the NMS and its association function for bounding boxes of the conventional tracking method to enhance the tracking capability of multiscale and occluded objects.

The association speed and performance of the object tracker are directly dependent on the values of the association function. To determine the multiple object tracking (MOT) value, it is necessary to calculate the correspondence among the bounding boxes using the tracking method. In order to determine the volume of union between two bounding boxes, the intersection over union (IoU) metric [41] are used, and the consistent association function is stated as follows:

$$IoU = \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|}, \quad (1)$$

where  $B_1$  and  $B_2$  are 3D bounding boxes;  $|B_1 \cap B_2|$  indicates the volume of intersection of  $B_1$  and  $B_2$ ; and  $|B_1 \cup B_2|$  indicates the volume of the union of  $B_1$  and  $B_2$ . The IoU is equal to 0 when there is no intersection between the two 3D bounding boxes. In this case, the tracking process cannot continue.

As a solution for this gradient vanishing matter, generalized intersection over union (GIoU) equation [33] is used in the tracking technique, which is stated as follows:

$$GIoU = IoU - \frac{|D|}{|C|}, \quad (2)$$

where  $C$  is the smallest volume that covers  $B_1$  and  $B_2$ ; let  $D = C / (B_1 \cup B_2)$ , the  $C = B_1 \cup B_2 \cup D$ ;  $|C|$  and  $|D|$  stand for the volume of  $C$  and  $D$ , respectively. When  $B_2$  box contains the  $B_1$  box, then the variance between each  $B_1$  box and the  $B_2$  box are the same, GIoU, in this case, degenerates into IoU, without any tracking relationship.

IoU and GIoU only take into account the overlapping volume, and the associated functions have two drawbacks, including delayed corresponding and incorrect association. However, distance intersection over union (DIoU) uses the standardized distance among the centers of the  $B_1$  and  $B_2$  bounding boxes. The following definition relates to this association function [42]:

$$DIoU = IoU - \frac{d^2}{c^2}, \quad (3)$$

where  $d$  is the Euclidean distance length between the center points of the  $B_1$  and  $B_2$  bounding boxes;  $c$  is the diagonal length of the smallest enclosing box that encompasses the two boxes. The DIoU function causes the model to acquire quick association if the two boxes are in either the horizontal or vertical direction at the same time. Directly reducing the normalized distances between central point's using the DIoU function leads to a faster convergence rate [42] and more precise association. The IoU, GIoU, and DIoU, as expressed above, are used to describe the association between any two bounding boxes. The algorithm of 3D DIoU metric is defined as Algorithm 1.

**Algorithm 1. 3D Distance Intersection Over Union Function**

Input: the information data of  $B_1$  and  $B_2$  bounding boxes:

$$B_1 = (x_1, y_1, z_1, l_1, w_1, h_1, \theta_1), B_2 = (x_2, y_2, z_2, l_2, w_2, h_2, \theta_2)$$

Output: 3D DIoU Association Metric

1. Determining the Projections  $B'_1$  and  $B'_2$  of  $B_1$  and  $B_2$  on the bird's eye view, respectively  $B'_1 = (x^1_1, y^1_1, x^2_1, y^2_1, \theta'_1)$ ,  $B'_2 = (x^1_2, y^1_2, x^2_2, y^2_2, \theta'_2)$
2.  $A_1 \leftarrow$  the area of the 2D box  $B'_1$
3.  $A_2 \leftarrow$  the area of the 2D box  $B'_2$
4.  $I_{2D} \leftarrow$  intersection between  $B'_1$  and  $B'_2$
5.  $U_{2D} \leftarrow$  union between  $B'_1$  and  $B'_2$
6.  $I_h \leftarrow$  the height of the intersection between  $B_1$  and  $B_2$
7.  $U_h \leftarrow$  the height of the union between  $B_1$  and  $B_2$
8.  $I_w \leftarrow$  the width of the intersection between  $B_1$  and  $B_2$
9.  $U_w \leftarrow$  the width of the union between  $B_1$  and  $B_2$
10.  $I_l \leftarrow$  the length of the intersection between  $B_1$  and  $B_2$
11.  $U_l \leftarrow$  the length of the union between  $B_1$  and  $B_2$
12.  $I_v \leftarrow$  the volume of the intersection between  $B_1$  and  $B_2$
13.  $U_v \leftarrow$  the volume of the union between  $B_1$  and  $B_2$
14.  $d_x \leftarrow$  the center distance for  $(x_1 - x_2)$
15.  $d_y \leftarrow$  the center distance for  $(y_1 - y_2)$
16.  $d_z \leftarrow$  the center distance for  $(z_1 - z_2)$
17.  $d^2 \leftarrow$  the diagonal distance between  $B_1$  and  $B_2$
18.  $c^2 \leftarrow$  the diagonal distance for the smallest enclosing box that encompasses between  $B_1$  and  $B_2$
19. If  $I_{2D} \leq 0$  :  
 $I_v = 0$ ;  
 else:  
 If  $I_h \leq 0$  :  
 $I_v = 0$ ;  
 else:  
 $I_v = I_{2D} \times I_h$ ;
20.  $d^2 = d_x^2 + d_y^2 + d_z^2$ ;
21.  $c^2 = U_w^2 + U_l^2 + U_h^2$ ;
22.  $IoU_{3D} = \frac{I_v}{U_v}$ ;
23.  $DIoU_{3D} = IoU_{3D} - \frac{d^2}{c^2}$

**3.2. Non-Maximum Suppressing (NMS) Upgrade to DIoU-NMS**

To locate local maximum and to eliminate non-maximum bounding boxes, the NMS approach is used. Most object-tracking systems use NMS as a pre-processing stage, which is often used to choose the bounding boxes before starting the tracking operation. Depending on the score for classification confidence, which is the foundation of the original NMS, the bounding box that has the highest confidence score can be maintained. Since IoU and classification confidence scores are typically not strongly correlated, it is difficult to pinpoint many classification labels with a high number of confidence scores. When using the tracking method with the original NMS technique, analysis is only performed over overlapping regions, increasing the likelihood of missing and false detection, specifically in scenes with extremely overlapping objects.

We use DIoU-NMS to increase the detection efficiency for the occluded object. DIoU-NMS uses DIoU as a tool for suppressing the redundant bounding boxes, in contrast to the original NMS, which uses IoU as the criterion. DIoU-NMS takes into account the distance length between the center points of the two bounding boxes in addition to the overlapping area. The DIoU-NMS is stated as

$$s_i = \begin{cases} s_i, & IoU - R_{DIoU}(M, B_i) < \varepsilon \\ 0, & IoU - R_{DIoU}(M, B_i) \geq \varepsilon' \end{cases} \quad (4)$$

where  $s_i$  stands for the score of classification confidence;  $IoU$  is stated in an Equation (1);  $\varepsilon$  denotes the value of NMS threshold;  $M$  represents the highest-scoring bounding boxes; and  $B_i$  is the pending bounding box. When conducting DIoU-NMS, the distance between the centers of two bounding boxes is taken into account concurrently with IoU. The distance is indicated by  $R_{DIoU}$  and the equivalent equation is as follows:

$$R_{DIoU} = \frac{p^2(b_1, b_2)}{c^2} \quad (5)$$

where  $p^2$  denotes the length of the central distance measured between the bounding box  $b$  center point and the bounding box  $b_2$  ones;  $c^2$  is the smallest box's diagonal, which contains both boxes.

## 4. Results

### 4.1. Datasets

There have been several MOT datasets recommended and used during the last few years. Waymo Open Dataset [12] and nuScenes [13] are the most commonly used and most considerable benchmark for MOT. Waymo Open Dataset (WOD) includes a perception dataset and a motion dataset. The total number of scenes in the dataset is 1150, divided into 150, 202, and 798 scenes for testing, validation, and training, respectively. While the motion dataset comprises 103,354 sequences, the perception dataset has 1950 lidar sequences that have been annotated. Each sequence is recorded for 20 s at a sample rate of 10 Hz. For each frame, point cloud data and 3D ground truth boxes for vehicles, pedestrians, and cyclists are provided. By using the evaluation metrics stated in [12], we recorded multiple object tracking accuracy (MOTA), multiple object tracking precision (MOTP) [43], Miss, Mismatch, and false Positive (FP) for objects with the L2 difficulty level.

NuScenes [13] provides ground truth 3D box annotations at 20 frames per second and LiDAR scans at 2 frames per second (fps) for a total of 1000 driving sessions. We report identity switches (IDS), AMOTA [7], and MOTA for nuScenes. AMOTA, the average value of MOTA, serves as the main indicator for assessing 3DMOT on nuScenes, is created by merging MOTA over several recalls. Meanwhile, AMOTP, the average value of MOTP, indicates an error value for the association process. Hence, the value for MOTP and AMOTP should be kept as small as possible.

### 4.2. DIoU-NMS Results

Our approach aims to increase the precision without considerably reducing the recall. We apply a strict DIoU-NMS to the input detections, and it is found that the ID switch only recorded 479 switches, in comparison with the IoU method, which is 519, as shown in Table 1.

**Table 1.** NMS for IoU and DIoU in the detection process with GIoU in association two stage.

NMS Metric	AMOTA	AMOTP	RECALL	MOTA	ID Switch
IoU	0.687	0.573	0.725	0.592	519
DIoU	0.688	0.573	0.722	0.592	479

In addition, when DIoU-NMS is applied to the Waymo Open Dataset, the MOTA is higher than that resulting from IoU-NMS. Similarly, the mismatch value improved and reached 0.077% for vehicle class, as in Table 2. Meanwhile, MOTA results reached 51% and the mismatch value is equal to 0.4% for pedestrian objects, as shown in Table 3.



**Table 2.** Comparison of the tracking results for vehicle objects using different NMS metrics on the validation set of Waymo Open Dataset.

NMS Metric	MOTA	MOTP	Miss	Mismatch (%)	FP
IoU	0.544	0.168	0.355	0.08	0.099
DIoU	0.547	0.1678	0.354	0.077	0.097

**Table 3.** Comparison of the tracking results for pedestrian objects using different NMS metrics on the validation set of Waymo Open Dataset.

NMS Metric	MOTA	MOTP	Miss	Mismatch (%)	FP
IoU	0.505	0.311	0.397	0.45	0.092
DIoU	0.510	0.311	0.388	0.40	0.097

#### 4.3. Association Results

We used 3D box detection from the CenterPoint method as the input data. To select boxes with scores higher than 0.7 on the Waymo Open Dataset, 3D IoU-NMS was set to 0.7. To associate between detection and prediction boxes, we used two-stage data association, namely 3D GIoU and DIoU. In this case, we associated the detection and prediction boxes by using DIoU in the first stage, then we re-associated again any un-associated boxes with DIoU for the detection and tracklets in the second stage. A similar second stage approach was applied to the third and higher order stage data association. The results for two-stage data associations are shown in Table 4 for vehicle class, and Table 5 for pedestrian class using the Waymo Open Dataset. The first row in both Tables 4 and 5 represents two-stage data association results, where 3D GIoU metric is coupled with the Hungarian algorithm to match between detections and tracklets. On the other hand, in the second row, we used 3D DIoU instead of GIoU metrics.

**Table 4.** Comparisons for 3D MOT two-stage association on vehicle class, Waymo Open Dataset validation set.

Two-Stage	MOTA	MOTP	Miss	Mismatch (%)	FP
GIoU	0.5612	0.1681	0.3344	0.078	0.1035
DIoU	0.5892	0.1736	0.3165	0.1559	0.092

**Table 5.** Comparisons for 3D MOT two-stage association on pedestrian class, Waymo Open Dataset validation set.

Two-Stage	MOTA	MOTP	Miss	Mismatch (%)	FP
GIoU	0.5776	0.3125	0.3090	0.425	0.1091
DIoU	0.5972	0.3518	0.3360	0.96	0.0570

On the other hand, Table 6 represents a three-stage data association, where only the cyclist class has a low false positive value (FP), the unmatched detections and unmatched tracklets are associated using 3D DIoU coupled with the Hungarian algorithm.

**Table 6.** Three-dimensional MOT three-stage association on cyclist class, Waymo Open Dataset validation set.

Three-Stage	MOTA	MOTP	Miss	Mismatch (%)	FP
DIoU	0.6018	0.2855	0.3033	0.6613	0.0881

#### 4.4. Comparison with Previous Techniques

In this part, we incorporate the aforementioned methods into the combined DIoU-NMS and DIoU data association in order to demonstrate how the performance can be enhanced. Table 7 below shows our proposed 3D MOT trackers perform better than the baselines. In the case of the Waymo Open Dataset, although the dimensions of vehicles and pedestrians are much different, DIoU-NMS and two-stage DIoU data association are adequate and appropriate for both vehicle and pedestrian objects due to high tracking performance values, as the results are illustrated in Tables 7 and 8. The comparison for vehicle class in the Waymo Open Dataset test set is tabulated in Table 7, which CenterPoint [8] recognitions are utilized. For comparison, the results from AB3DMOT [7] and Chiu et al. [5] are also presented. On the same note, Table 8 highlights the results for pedestrian class in the Waymo Open Dataset test set. Meanwhile, the three-stage technique is only applicable to cyclist objects due to the limitation of MOTA value computation for vehicles and pedestrians in this multi-stage evaluation. DIoU-NMS shows effective results on the nuScenes dataset, as shown in Table 9. In this case, CenterPoint [8] detection is utilized and compared. In all tests, 2 Hz frame rate is used for the detection.

**Table 7.** Comparison on Waymo Open Dataset test set, vehicle class.

Method	MOTA	MOTP	Mismatch (%)
AB3DMOT [7]	0.5773	0.1614	0.26
Chiu et al. [5]	0.4932	0.1689	0.62
CenterPoint [8]	0.5938	0.1637	0.32
This Work	0.6061	0.1738	0.094

**Table 8.** Comparison on Waymo Open Dataset test set, pedestrian class.

Method	MOTA	MOTP	Mismatch (%)
AB3DMOT [7]	0.5380	0.3163	0.73
Chiu et al. [5]	0.4438	0.3227	1.83
CenterPoint [8]	0.5664	0.3116	1.07
This Work	0.615	0.329	0.521

**Table 9.** Comparison on the nuScenes test set.

Method	AMOTA	AMOTP	MOTA	ID Switch
AB3DMOT [7]	0.151	1.501	0.154	8987
Chiu et al. [5]	0.550	0.798	0.459	736
CenterPoint [8]	0.638	0.555	0.537	730
CBMOT [4]	0.649	0.592	0.545	517
OGR3MOT [9]	0.656	0.620	0.554	248
This Work	0.658	0.568	0.557	569

#### 4.5. Comparison between GIoU and DIoU

As a comparison between the association metric between GIoU and DIoU, the score threshold for selecting the boxes is set equal to 0.7. The GIoU association threshold is equal to 1.5 and in the case of DIoU, it is equal to 1. Meanwhile, the NMS-IoU threshold is equal to 0. The figures below show the associations between the detection results (green boxes) and the predicted results (blue boxes). In this case, it can be seen that for both DIoU and GIoU in the first stage, and for DIoU, the predicted box is preserved until the object is detected again. Meanwhile, in the GIoU case, the predicted box is terminated when the object is temporarily not observed, causing an identity switch, as illustrated in Figure 6. At frame 9, which the figures shown in the first row, the detected box for vehicle ID number 2 (green box) is associated with its predicted box (blue box). The second row showed frame 11, which contains the tracklet for vehicle ID 10. We also use DIoU with a predicted box

for vehicle ID 2 for association process. However, when we apply GIoU on the association process, we obtained the tracklet only for vehicle ID number 10 and the predicted box for vehicle ID 2 is terminated, as shown by the first column in the second row of Figure 6. At frame 28, the tracklet for vehicle 3 and 11 and predicted position for vehicles 0, 1, 2, and 6 are shown in final row, second column when we apply DIoU on the association process. On the other hand, when we apply GIoU on association process, we obtain the tracklet for vehicle 3 and 11 only, and the predicted boxes are terminated for vehicles 0, 1, 2, and 6, which lead to an increase in ID switch and lowering of the multiple object tracking accuracy (MOTA) value.

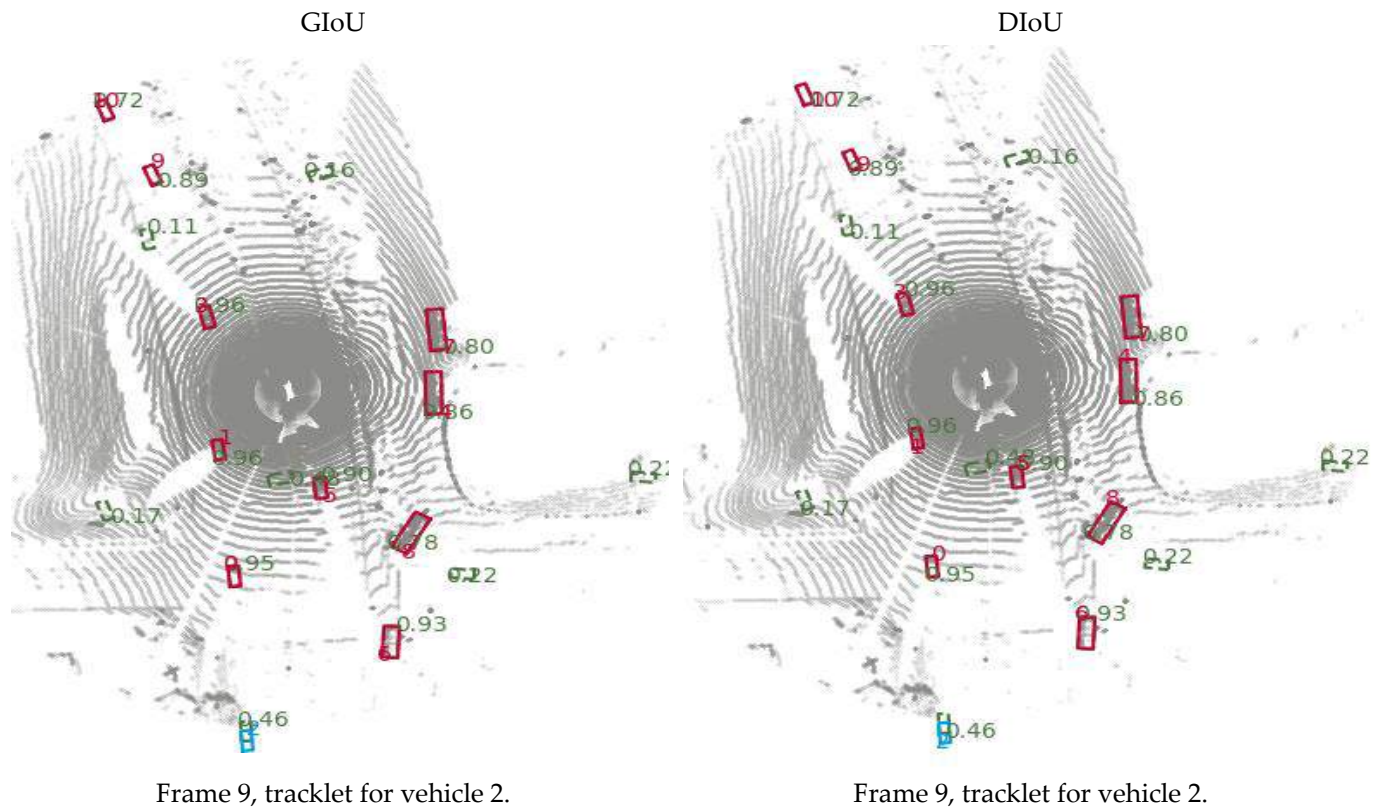
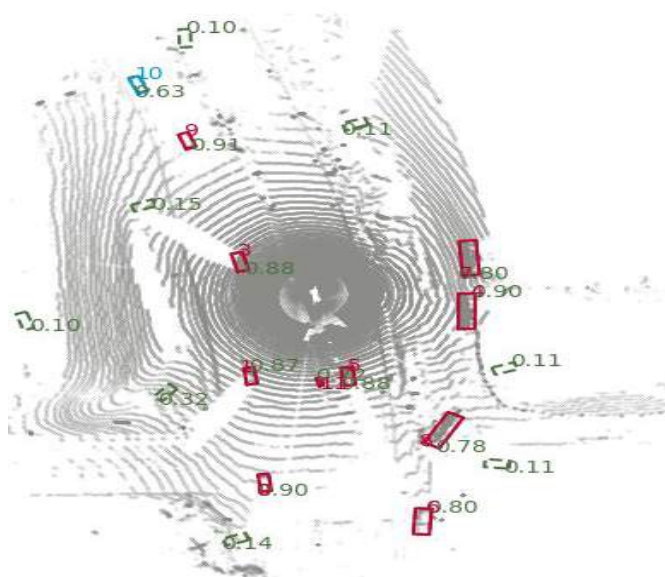
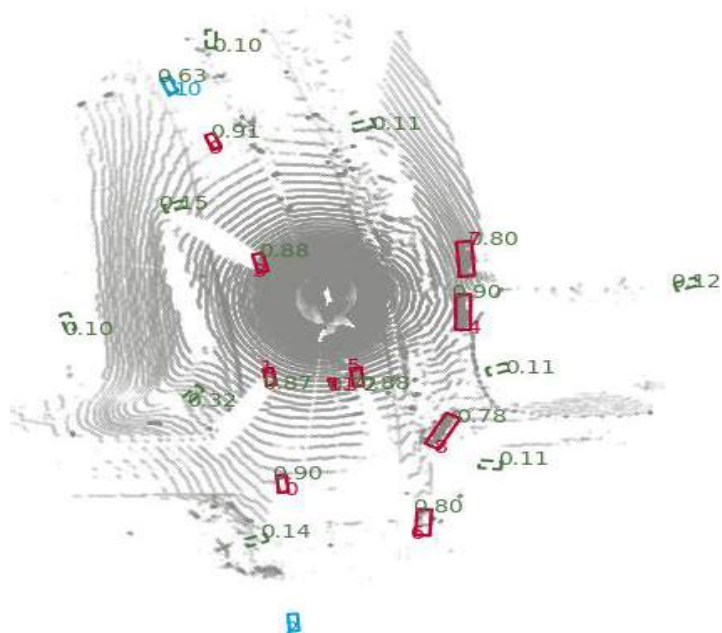


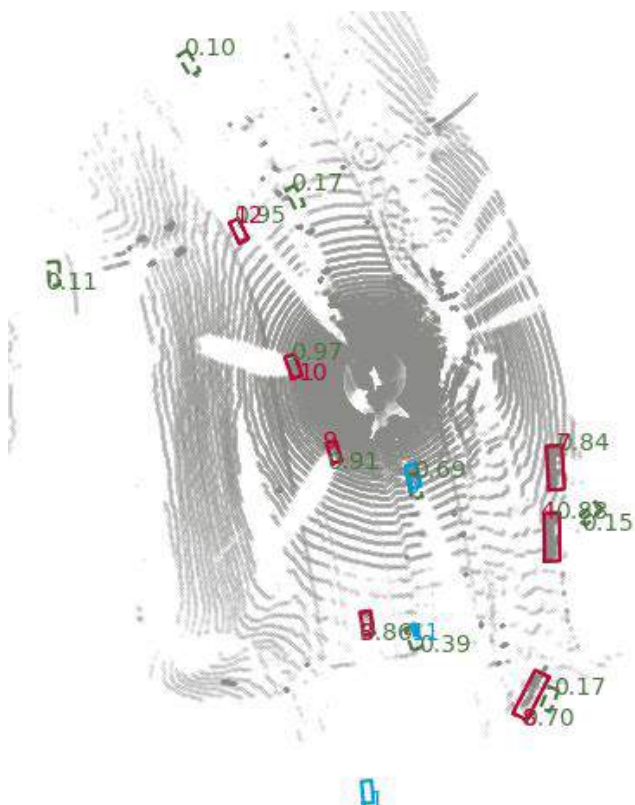
Figure 6. Cont.



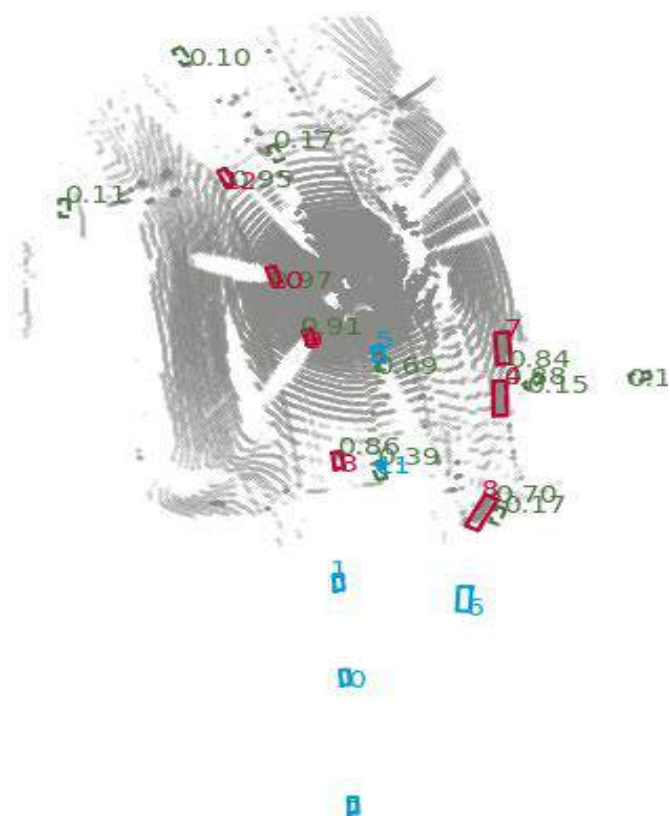
Frame 11, tracklet for vehicle 10.



Frame 11, tracklet for vehicle 10 and predicted position for vehicle 2.



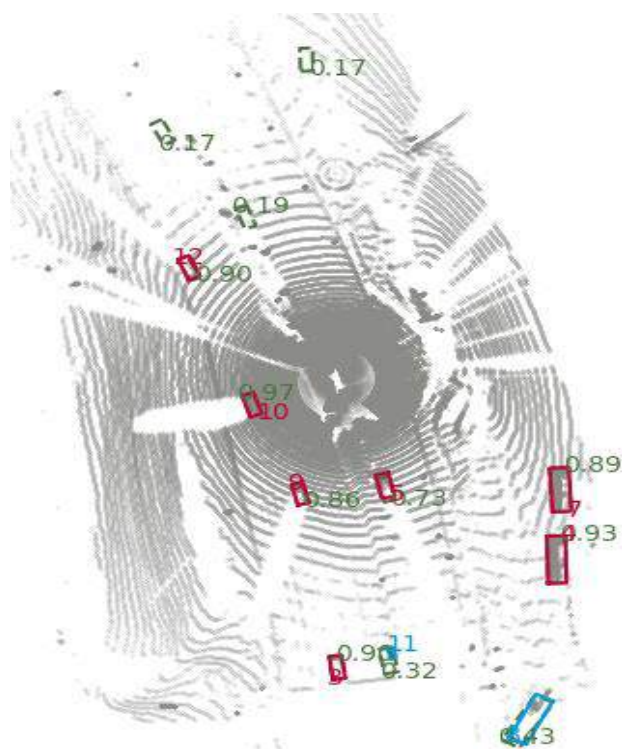
Frame 26, tracklet for vehicle 5, 11 and predicted position for vehicle 1.



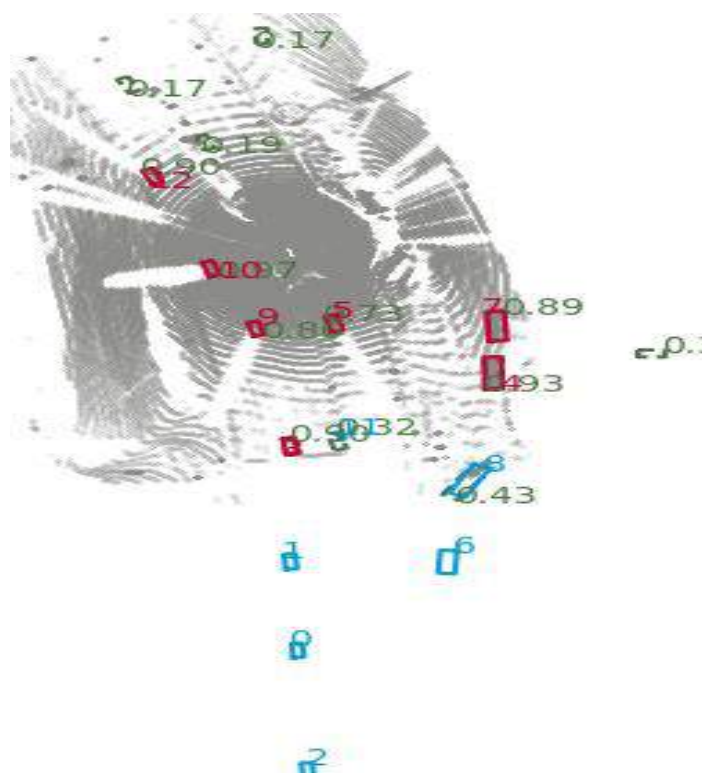
Frame 26, tracklet for vehicle 5, 11 and predicted position for vehicle 0, 1, 2, 6.

Figure 6. Cont.





Frame 28, tracklet for vehicle 3 and 11.



Frame 28, tracklet for vehicle 3 and 11 and predicted position for vehicle 0, 1, 2, 6.

**Figure 6.** Comparison between GIoU and DIoU for Association process.

## 5. Conclusions

It was discovered that tracklet termination leads to identity switches in 3D MOT, which are common and unresolved issues in recent 3D MOT studies. Therefore, in this paper, we proposed a hybrid method of using DIoU-NMS and DIoU in order to improve the association between tracklet and prediction boxes for objects. We found that by using the combination of DIoU-NMS and DIoU, the identity switch cases can be reduced.

Additionally, we used DIoU for multi-stage association, which lead to an increase in MOTA values for small objects on the Waymo Open Dataset. Experiment results show that DIoU-NMS can significantly reduce the identity switches when it is used in selecting the detectors for tracking. Our approach achieved 479 ID switches for the vehicle objects on nuSense compared with 519 for GIoU only. While the mismatches were improved slightly for vehicles and pedestrian objects, the MOTA results also recorded better performance in tracking on the Waymo Open Dataset. Meanwhile, two-stage data association results demonstrated significant improvements in MOTA values with 58.9% and 59.7% for vehicle and pedestrian objects, respectively. The FP values also significantly improved, which are 11.1% and 47.7% for vehicle and pedestrian objects, respectively. In addition, using DIoU for three-stage association reduced the false positive detection as well as improved MOTA values.

In comparison to previous work, our method recorded significant improvement in ID mismatches, which achieved at least 63.8% and 28.6% reductions for vehicle and pedestrian objects, respectively. Similarly, test results on Waymo Open Dataset show MOTA values for both vehicles and pedestrian objects reach over 60%, overtaking all previously issued LiDAR-based methods. The results show great potential for future 3D MOT analysis and can pave the path for many real-time 3D tracking-by-detection applications.

**Author Contributions:** Conceptualization and investigation, S.A.K.M.; Supervision, writing—review and editing, M.Z.A.R.; Validation, Rahman, A.H.A.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by Ministry of Higher Education (MOHE) Malaysia under research grant FRGS/1/2020/STG07/UKM/02/3 and the APC was partially funded by Universiti Kebangsaan Malaysia (UKM).

**Data Availability Statement:** Data supporting the conclusions of this manuscript are provided within the article and will be available from the corresponding author upon request.

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

## References

- Pang, Z.; Li, Z.; Wang, N. Model-Free Vehicle Tracking and State Estimation in Point Cloud Sequences. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 8075–8082.
- Qi, C.R.; Zhou, Y.; Najibi, M.; Sun, P.; Vo, K.; Deng, B.; Anguelov, D. Offboard 3d Object Detection from Point Cloud Sequences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6134–6144.
- Liu, Y.; Wang, W.; Chambers, J.; Kilic, V.; Hilton, A. Particle Flow SMC-PHD Filter for Audio-Visual Multi-Speaker Tracking. In Proceedings of the Latent Variable Analysis and Signal Separation: 13th International Conference, LVA/ICA 2017, Grenoble, France, 21–23 February 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 344–353.
- Benbarka, N.; Schröder, J.; Zell, A. Score Refinement for Confidence-Based 3D Multi-Object Tracking. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 8083–8090.
- Kuang Chiu, H.; Prioletti, A.; Li, J.; Bohg, J. Probabilistic 3d Multi-Object Tracking for Autonomous Driving. *arXiv* **2020**, arXiv:2001.05673.
- Pöschmann, J.; Pfeifer, T.; Protzel, P. Factor Graph Based 3d Multi-Object Tracking in Point Clouds. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October–24 January 2020; pp. 10343–10350.
- Weng, X.; Wang, J.; Held, D.; Kitani, K. 3d Multi-Object Tracking: A Baseline and New Evaluation Metrics. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October–24 January 2020; pp. 10359–10366.
- Yin, T.; Zhou, X.; Krahenbuhl, P. Center-Based 3d Object Detection and Tracking. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11784–11793.
- Zaech, J.-N.; Liniger, A.; Dai, D.; Danelljan, M.; Van Gool, L. Learnable Online Graph Representations for 3d Multi-Object Tracking. *IEEE Robot. Autom. Lett.* **2022**, *7*, 5103–5110. [\[CrossRef\]](#)
- Mahalanobis, P.C. On the Generalised Distance in Statistics. *Proc. Natl. Inst. Sci. India* **1936**, *12*, 49–55.
- Zhou, Y.; Tuzel, O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4490–4499.
- Sun, P.; Kretschmar, H.; Dotiwala, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2446–2454.
- Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. Nuscenes: A Multimodal Dataset for Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11621–11631.
- Pang, Z.; Li, Z.; Wang, N. Simpletrack: Understanding and Rethinking 3d Multi-Object Tracking. In *Proceedings of the Computer Vision—ECCV 2022 Workshops: Tel Aviv, Israel, 23–27 October 2022*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 680–696.
- Li, X.; Ma, C.; Wu, B.; He, Z.; Yang, M.-H. Target-Aware Deep Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1369–1378.
- Yuan, D.; Shu, X.; Fan, N.; Chang, X.; Liu, Q.; He, Z. Accurate Bounding-Box Regression with Distance-IoU Loss for Visual Tracking. *J. Vis. Commun. Image Represent.* **2022**, *83*, 103428. [\[CrossRef\]](#)
- Zhihao, C.A.I.; Longhong, W.; Jiang, Z.; Kun, W.U.; Yingxun, W. Virtual Target Guidance-Based Distributed Model Predictive Control for Formation Control of Multiple UAVs. *Chin. J. Aeronaut.* **2020**, *33*, 1037–1056.
- Huang, Y.; Liu, W.; Li, B.; Yang, Y.; Xiao, B. Finite-Time Formation Tracking Control with Collision Avoidance for Quadrotor UAVs. *J. Franklin Inst.* **2020**, *357*, 4034–4058. [\[CrossRef\]](#)
- Dewangan, D.K.; Sahu, S.P. Lane Detection in Intelligent Vehicle System Using Optimal 2-Tier Deep Convolutional Neural Network. *Multimed. Tools Appl.* **2023**, *82*, 7293–7317. [\[CrossRef\]](#)
- Dicle, C.; Camps, O.I.; Sznaiar, M. The Way They Move: Tracking Multiple Targets with Similar Appearance. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2304–2311.
- Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple Online and Realtime Tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468.



22. Bergmann, P.; Meinhardt, T.; Leal-Taixe, L. Tracking without Bells and Whistles. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 941–951.
23. Lu, Z.; Rathod, V.; Votel, R.; Huang, J. Retinatrack: Online Single Stage Joint Detection and Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14668–14678.
24. Sadeghian, A.; Alahi, A.; Savarese, S. Tracking the Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 300–311.
25. 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. Vis.* **2021**, *129*, 3069–3087. [[CrossRef](#)]
26. Leal-Taixé, L.; Canton-Ferrer, C.; Schindler, K. Learning by Tracking: Siamese CNN for Robust Target Association. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 27–30 July 2016; pp. 33–40.
27. Li, J.; Gao, X.; Jiang, T. Graph Networks for Multiple Object Tracking. In Proceedings of the IEEE/CVF winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020; pp. 719–728.
28. Wojke, N.; Bewley, A.; Paulus, D. Simple Online and Realtime Tracking with a Deep Association Metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649.
29. Patil, A.; Malla, S.; Gang, H.; Chen, Y.-T. The H3d Dataset for Full-Surround 3d Multi-Object Detection and Tracking in Crowded Urban Scenes. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 9552–9557.
30. Shi, S.; Wang, X.; Li, H. PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 770–779.
31. Kuhn, H.W. The Hungarian Method for the Assignment Problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [[CrossRef](#)]
32. Chiu, H.; Li, J.; Ambrus, R.; Bohg, J. Probabilistic 3d Multi-Modal, Multi-Object Tracking for Autonomous Driving. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–June 5 2021; pp. 14227–14233.
33. Rezaatofghi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
34. Zhou, X.; Koltun, V.; Krähenbühl, P. Tracking Objects as Points. In *Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 474–490.
35. Yang, B.; Bai, M.; Liang, M.; Zeng, W.; Urtasun, R. Auto4d: Learning to Label 4d Objects from Sequential Point Clouds. *arXiv* **2021**, arXiv:2101.06586.
36. Weng, X.; Wang, Y.; Man, Y.; Kitani, K.M. Gnn3dmot: Graph Neural Network for 3d Multi-Object Tracking with 2d-3d Multi-Feature Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6499–6508.
37. Kim, A.; Ošep, A.; Leal-Taixé, L. Eagermot: 3d Multi-Object Tracking via Sensor Fusion. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 11315–11321.
38. He, J.; Huang, Z.; Wang, N.; Zhang, Z. Learnable Graph Matching: Incorporating Graph Partitioning with Deep Feature Learning for Multiple Object Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5299–5309.
39. Liu, Y.; Wang, W.; Kilic, V. Intensity Particle Flow Smc-Phd Filter for Audio Speaker Tracking. *arXiv* **2018**, arXiv:1812.01570.
40. Liu, Y.; Hu, Q.; Zou, Y.; Wang, W. Labelled Non-Zero Particle Flow for Smc-Phd Filtering. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5197–5201.
41. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. Unitbox: An Advanced Object Detection Network. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 516–520.
42. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
43. Bernardin, K.; Stiefelwagen, R. Evaluating Multiple Object Tracking Performance: The Clear Mot Metrics. *EURASIP J. Image Video Process.* **2008**, *2008*, 1–10. [[CrossRef](#)]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.