1st Place Solutions for Waymo Open Dataset Challenges - 2D and 3D Tracking

Yu Wang Sijia Chen Li Huang Runzhou Ge
Yihan Hu Zhuangzhuang Ding Jie Liao

Horizon Robotics Inc.

yu04.wang@horizon.ai

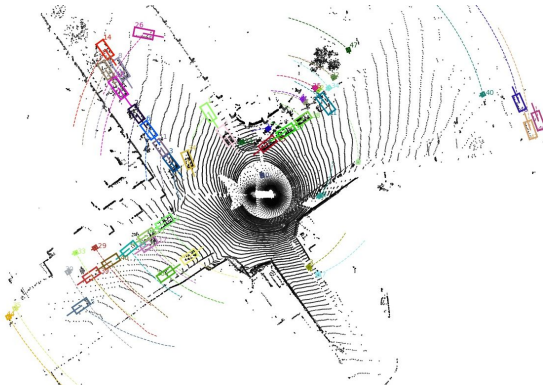
Abstract

This technical report presents the online and realtime 2D and 3D multi-object tracking (MOT) algorithms that reached the 1st places on both Waymo Open Dataset 2D tracking and 3D tracking challenges. An efficient and pragmatic online tracking-by-detection framework named HorizonMOT is proposed for camera-based 2D tracking in the image space and Lidar-based 3D tracking in the 3D world space. Within the tracking-by-detection paradigm, our trackers leverage our high-performing detectors used in the 2D/3D detection challenges and achieved 45.13% *2D MOTA/L2 and* 63.45% *3D MOTA/L2 in the 2D/3D tracking challenges.*

1. Introduction

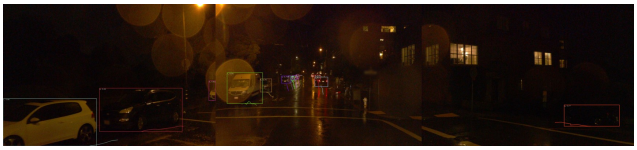
Tracking-by-detection approaches has been leading the online (no peeking into the future) multi-object tracking (MOT) benchmarks, as a result of the high-performing object detection models. There are a group of pragmatic tracking-by-detection approaches for 2D/3D multiple object tracking whose data association method is simply based on bounding box overlap or object center distance and built upon Kalman filter [12] [2] [3] [11] [4]. The majority of the participants in Waymo 2D and 3D tracking challenges is based on these methods. Despite the fact that tracking-by-detection often relies on strong object detectors, better overall performance can still be achieved by improving the data association schemes and the tracking framework.

In recent literature, there is a trend of joint detection and tracking using a single network, such as CenterTrack [14], RetinaTrack [6] and another tracker developed on top of CenterNet [13]. This paradigm was adopted by some participants in the 2D tracking challenge. The CenterTrack [14] network learns a 2D offset of the same object between two adjacent frames and associate it overtime based on center distance. The overall idea of CenterTrack is simple yet effective. One problem of CenterTrack is that it does short-term association and therefore is unable to handle long-term association or the occlusion or missing detection problem.



(a)

(b)



(c)

Figure 1: Examples of our 3D tracking (a) and 2D tracking (b)(c) results on the Waymo Open Dataset v1.2. The ego-vehicle is making a left turn in (a). We highlight in (b) two pedestrians that were tracked successfully even after occlusions in a crowed scene. Note in (c) that we do not track the object across cameras as it is not required by the challenge. Trajectories of the cars in the front camera (c) indicate large object displacement caused by pitch motion of the camera possibly due to uneven ground.

For the 2D and 3D tracking challenges, we proposed an unified and pragmatic framework named HorizonMOT that focuses on frame-to-frame prediction and association, and is applicable to both 2D camera-based tracking in the image space and Lidar-based 3D tracking in the 3D world space, as shown in Figure 1. Our trackers are online since only detections of the current frame are presented to the tracker, and the result of the current frame is decided right away without any latency. Our tracker belongs to the tracking-by-detection paradigm.

2. Detection Network

2.1. 2D Detection Network

High-performing detectors are the key to the success of tracking-by-detection approaches. We employ the one-stage anchor-free, Non-Maxima Suppression (NMS) free CenterNet framework [15] for 2D object detection. Under the CenterNet paradigm, many complicated perception tasks can be simplified in an unified framework as object center point detection and regression of object properties such as bounding box size, 3D information (*e.g.* 3D location, 3D dimension, heading), pose, or embedding.

We use Hourglass as the CenterNet backbone. As illustrated in Figure 2, two hourglass blocks are stacked and the first one only serves as providing auxiliary loss during training. We tried using both stacks for inference but it did not improve the results. As in CenterNet, the heads output the center heatmap which is of size $\frac{W}{R} \times \frac{H}{R} \times C$, where R is the output stride and set to 4, W and H are the width and height of the input image, while C is the number of classes. Width and height values are regressed for each center point and additional local offset values are regressed to recover the discretization error caused by the output stride.

We can add a pixel-wise embedding (similar to [13]) and 2nd-stage per-ROI feature extraction branches to extract Re-ID features in an end-to-end unified network but they were not used for the challenge.

2.2. 3D Detection Network

In the 3D detection track, our solution is an improvement upon our baseline detector named AFDet [5] and reached the 1st place, and we use 3D detections produced by this solution as input to our 3D tracker.

3. Tracking Framework

Tracking-by-detection consists of the following components: 1) track creation and deletion; 2) state prediction and update using the Kalman filter; 3) association between tracks and detections. We assume no ego-motion information and no future information is available.

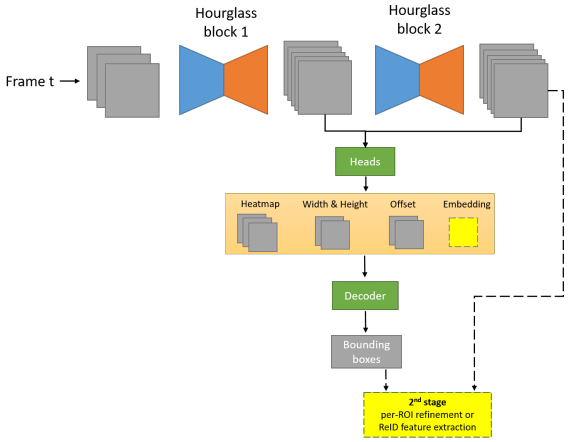


Figure 2: The main detection network is based on CenterNet with Hourglass-104 as the backbone. We did not finish the embedding head and the 2nd-stage per-ROI refinement or ReID feature extraction branches before the challenge deadline but we still include them here as future works.

3.1. Track Creation and Deletion

Similar to [12] and [11], a new track is created when a detection of the current frame is not associated with any track. As in [12] and [11], each track k has a number of frames since the last successful detection association a_k and the track will be deleted once this counter exceeds predefined maximum age A_{max}.

3.2. State Prediction and Update Using Kalman Filter

In 2D tracking, for each track we define an eight dimensional state space $(x, y, \gamma, h, \dot{x}, \dot{y}, \dot{\gamma}, \dot{h})$, which contains the center (x, y), aspect ratio γ, height h, and their respective velocities in the image space. The observation is 2D detection box and its score. We simply set the score of the track to the score of its associated detection. In 3D tracking we use 10-dimensional state space $(x, y, z, h, w, l, \theta, \dot{x}, \dot{y}, \dot{z})$ which contains the 3d location, height, width and length, heading, and the respective velocities of the 3d location values, in the 3D space. The observation is 3D detection $(x, y, z, h, w, l, \theta, s)$ where s is the detection score. At each frame, state prediction is performed first using constant velocity model, and then the track-detection association, and then the state of each track is updated if it is associated with a detection. In Kalman filter, the estimated 2D/3D box is essentially a weighted average between state space and the observation [11]. In our experiments we use the observation directly as the output instead of using the weighted average. If a track is not associated with any detections at the current frame, only the prediction step is performed and the track does not contribute to the output of the current frame.

3.3. Association Metric

Association between detections of the current frame and tracks is based on the distance between bounding box and predicted state of track. Association metric is usually based on 2D/3D IoUs [2] [3] [11], Mahalanobis distance of 2D/3D object centers [12] [4], and cosine distance [12] between appearance/Re-ID features of 2D boxes. For 3D tracking in the 3D world space, one can also transform the 3D bounding boxes to image space and calculate association metric based on the overlap of 2D projections. In the 2D tracking challenge we adopt 2D box IoU and cosine distance. In the 3D tracking challenge we employ Euclidean distance with Gaussian kernel (with a parameter σ) between 3D centers , which is better and faster than other metrics that we tried on the validation set such as 3D box IoU, Bird's Eye View (BEV) box IoU (*i.e.* ignoring the vertical dimension) and Mahalanobis distance.

3.4. Three-stage Data Association

Typically association between tracks and detections is formulated as an assignment problem and relies on bipartite matching algorithms such as the Hungarian algorithm. In the tracking challenges we developed a three-stage data association scheme that applies to both 2D and 3D tracking and improved the MOTA scores. We first select a primary set of detections whose score is larger than $t^{(s)}$ and a secondary set in which score is within the range $[t^{(s)}/2, t^{(s)})$.

First-stage Association. We adopt the matching cascade proposed in [12] for the first stage. Association cost matrix is calculated first between tracks and the primary set of detections. We exclude unlikely associations if the cost is larger than a specified threshold $t^{(1)}$. We start from the most frequently seen track (*i.e.* with smallest track age a_k) and iterate over each track of increasing age and solve a linear assignment problem.

Second-stage Association. In the second stage, association is between un-matched tracks with age less than 3 and remaining detections in the primary detection set, we use a different association metric or relaxing the condition of the same association metric used in first stage (*e.g.* by enlarging size of a 2D bounding box to increase its overlap overtime). The association is again solved in a linear assignment problem and only admissible associations are kept by excluding unlikely associations using a specified distance threshold $t^{(2)}$.

Third-stage Association. In the third matching stage, association is between remaining un-matched tracks and detections in the secondary set. This helps to account for objects with weak detections (*e.g.* caused by partial occlusion). Admissible associations with distance lower than specified threshold $t^{(3)}$ are kept.

3.5. Appearance features for 2D camera-based tracking

Our final submission on the 2D tracking track relies on Re-ID features extracted by a small independent network. Re-ID or appearance features help handle long-term occlusion or objects with large displacement which could result in the failure of IoU based association metric. There are many scenarios which could lead to rapid displacements of object in the image plane. For example, low frame rate, vehicles in the opposite traffic direction with high relative speed, and unaccounted camera motion such as large camera pitch motion caused by bumps on the ground.

Following [12], we keep a gallery of the history associated Re-ID features of each track and the smallest cosine distance between them and the detection is used as the distance. We also introduce a maximum appearance distance $t^{(a)}$ to exclude unlikely associations.

2D Tracking Parameters	**Pedestrian**	**Vehicle**	**Cyclist**
score threshold $t^{(s)}$	0.5	0.4	0.5
max_appearance_dist $t^{(a)}$	0.15	0.06	0.15
max_iou_dist (front)	0.95	0.9	0.95
max_iou_dist (front left/right)	0.97	0.93	0.97
max_iou_dist (side)	0.99	0.95	0.99

Table 1: 2D tracking parameters. Maximum appearance distance and IoU distance are used to exclude unlikely associations during the three-stage data association.

3D Tracking Parameters	**Pedestrian**	**Vehicle**	**Cyclist**
score threshold $t^{(s)}$	0.5	0.5	0.5
max_center_dist	0.7	0.5	0.9
Gaussian kernel σ	1.5	5	3

Table 2: 3D tracking parameters. Maximum center distance is used to exclude associations with unlikely displacement in the 3D world space.

4. Experiments

4.1. Settings

Dataset. Our tracking algorithms are evaluated on the Waymo Open Dataset v1.2 [7]. We use its training set for training the 2D detection networks and 2D Re-ID networks, its validation set for verifying ideas and tuning parameters, and the test set to generate the our final submission to the leaderboard [10].

	MOTA/L1 ↑	MOTP/L1 ↑	MOTA/L2 ↑	MOTP/L2 ↑	FP/L2 ↓	Mis-match/L2 ↓	Miss/L2 ↓
HorizonMOT	51.01	14.18	45.13	14.27	7.13	2.25	45.49
Quasi-Dense R101	51.18	15.12	45.09	15.20	7.20	1.31	46.41
CascadeRCNN-SORTv2	50.22	14.85	44.15	14.85	6.94	2.44	46.46
Online V-IOU	46.07	13.28	40.09	13.40	5.73	3.42	50.76
DSTNet	43.83	15.76	38.01	15.84	6.28	1.34	54.38

Table 3: Top-5 2D tracking results on the test set of Waymo Open Dataset, with MOTA/L2 as the primary metric.

	MOTA/L1 ↑	MOTP/L1 ↑	MOTA/L2 ↑	MOTP/L2 ↑	FP/L2 ↓	Mis-match/L2 ↓	Miss/L2 ↓
HorizonMOT3D	65.13	23.96	63.45	23.96	7.28	0.29	28.99
PV-RCNN-KF	57.14	24.95	55.53	24.97	8.66	0.63	35.18
Probabilistic	49.16	24.80	47.65	24.82	8.99	1.01	42.35
3DMOT FS-H	46.52	24.69	45.07	24.74	9.04	1.90	44.00
3DMOT-MD-TPF	44.03	26.30	42.56	26.31	10.07	0.44	46.93

Table 4: Top-5 3D tracking results on the test set of Waymo Open Dataset, with MOTA/L2 as the primary metric.

Evaluation Metric. Waymo open dataset uses multiple object tracking metrics from [1]. MOTA is the main metric that takes into account the number of misses, false positives and mismatches. It is calculated for two difficulty levels. L1 metrics are calculated only for level 1 ground truth, while L2 metrics are computed by considering both level 1 and level 2 ground truth.

4.2. Implementation Details

2D and 3D Detections. In contrast to the original CenterNet, we use Gaussian kernels as in [16] which takes into account the aspect ratio of the bounding box to produce training samples for both center localization and size regression. During training, we use 768×1152 as the input size and a learning rate of $1.25e$-4. Due to lack of computational resource and sheer size of the dataset, we first trained a main network with weights pretrained on COCO on a $1/10$ subset of the training images for all 3 object categories (i.e. car, pedestrian, cyclist) for 25 epochs. A daytime expert model and a nighttime expert model were fine-tuned from this main network using only daytime or nighttime training images in the subset. To handle the imbalanced training data problem (*i.e*. pedestrian and especially cyclist have significantly less training samples than vehicle class), we also fine-tuned an expert model using only images with pedestrian and cyclist samples. We then fine-tuned 4 more models on the entire validation set, the entire training set, and images in the entire training set with pedestrian and cyclist samples, and nighttime images in the entire training set, respectively for 8-10 epochs. In inference we use flip and multi-scale (0.5, 0.75, 1, 1.25, 1.5) augmentation. To serve as tracker input, outputs of the 8 models were merged by naive NMS with IoU overlap threshold set to 0.5. Note that in the 2D detection challenge we use weighted boxes fusion instead to merge the results.

As input to our 3D tracker we rely on the 3D detections produced by our solution in the 3d detection challenge. Details of this solution can be found in our technical report for that challenge.

Re-ID Network. We use an independent Re-ID network with 11 3×3 and 3 1×1 convolutional layers and a max-pooling and average pooling layer and a downsampling factor of 16. Input image is normalized to 128×64 for pedestrian, and 128×128 for car/cyclist. The network was trained from scratch as classification network by adding a fully-connected layer and we prepared a total of 2844, 20041, and 906 unique objects for the pedestrian, car, and cyclist respectively from a subset of the Waymo 2D training images. The classification layer is removed during inference and the 512-dimension feature embedding servers as Re-ID feature.

2D Tracking. Cosine distance between Re-ID features is used in the first-stage matching, 2D IoU distance is used in the second and third-stage matching. We double or triple the size of the bounding boxes in the second and third-stage respectively when calculating the IoU overlap to account for objects with large displacement. Table 1 summarizes all the parameters used in my 2D tracking experiments. Note that we use different IoU matching thresholds for front, front left and right, and side cameras. We allow larger IoU distance (*i.e*. smaller overlap) in admissible associations for front left/right and side cameras since the displacement of some objects (especially pedestrians) tend to be very large. We assign the score of associated detection to the track as its score.

3D Tracking. Euclidean distance with Gaussian kernel be-

tween 3D centers is used throughout the three-stage associations. We use different σ values for each class as shown in Table 2. We also assign the score of associated detection to the track as its score.

Module Name	**MOTA/L1**	**MOTA/L2**
Baseline	41.73	36.03
+ Third-stage association	46.10	39.68
+ ReID models	48.79	42.11

Table 5: Ablation study on the 2D tracking validation set

4.3. Results

As shown in Table 3 and Table 4, our tracking algorithm reached the 1st pace on the official Waymo Open Dataset 2D and 3D tracking leaderboards [8] [9] and achieved the highest MOTA/L2 scores. In particular, our trackers return the lowest miss rate. Some qualitative results are shown in Figure 3 and Figure 4.

4.4. Ablation Study

On the 2D tracking validation set with 202 sequences, we study the effect of introducing the 3rd-stage association and using the Re-ID models. Our baseline performance is produced without these two components. As shown in Table 5, the 3rd-stage association results in a 3.65% improvement in terms of MOTA/L2 and the Re-ID models can further improve the performance by 2.43%.

5. Conclusion

An accurate, online and unified 2D and 3D tracking framework is proposed and achieved the 1st places on the Waymo Open Dataset 2D and 3D tracking challenges. In the future we will continue our ongoing work with the above-mentioned joint detection and tracking framework.

References

[1] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008, 01 2008.

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. *2016 IEEE International Conference on Image Processing (ICIP)*, Sep 2016.

[3] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *International Workshop on Traffic and Street Surveillance for Safety and Security at IEEE AVSS 2017*, Lecce, Italy, Aug. 2017.

[4] Hsu-kuang Chiu, Antonio Prioletti, Jie Li, and Jeannette Bohg. Probabilistic 3d multi-object tracking for autonomous driving. *arXiv preprint arXiv:2001.05673*, 2020.

[5] Runzhou Ge, Zhuangzhuang Ding, Yihan Hu, Yu Wang, Sijia Chen, Li Huang, and Yuan Li. Afdet: Anchor free one stage 3d object detection. In *CVPR Workshops*, 2020.

[6] Zhichao Lu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Retinatrack: Online single stage joint detection and tracking. *arXiv preprint arXiv:2003.13870*, 2020.

[7] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. *arXiv preprint arXiv:1912.04838*, 2019.

[8] Waymo. Waymo 2d tracking leaderboard. https://waymo.com/open/challenges/2d-tracking/, 2020.

[9] Waymo. Waymo 3d tracking leaderboard. https://waymo.com/open/challenges/3d-tracking/, 2020.

[10] Waymo. Waymo challenges. https://waymo.com/open/challenges/, 2020.

[11] Xinshuo Weng and Kris Kitani. A baseline for 3d multi-object tracking. *arXiv preprint arXiv:1907.03961*, 2019.

[12] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[13] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. A simple baseline for multi-object tracking. *arXiv preprint arXiv:2004.01888*, 2020.

[14] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *arXiv preprint arXiv:2004.01177*, 2020.

[15] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[16] Guodong Xu Zheng Yang Haifeng Liu Deng Cai Zili Liu, Tu Zheng. Training-time-friendly network for real-time object detection. *arXiv preprint arXiv:1909.00700*, 2019.

Figure 3: Qualitative results of 2D tracking. Each row shows 3 frames from the same sequence. Each object is showed in an unique color and assigned an unique ID. Note that they are not consecutive frames, the time interval is larger than one frame.

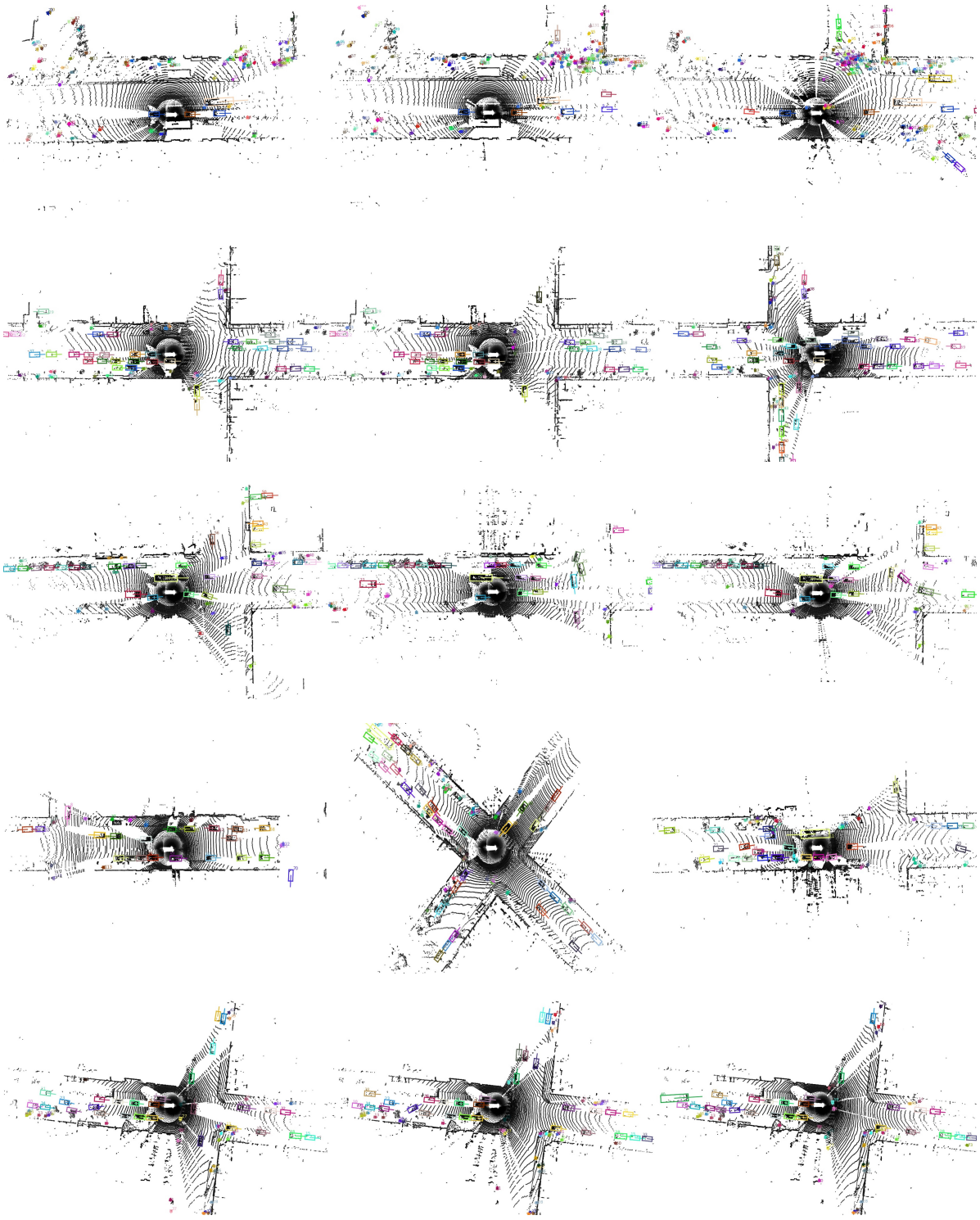


Figure 4: Qualitative results of 3D tracking. Each row shows 3 frames from the same sequence. Each object is showed in an unique color and assigned an unique ID. Note that they are not consecutive frames, the time interval is larger than one frame. For better visualization purpose point cloud data is sub-sampled.