Simultaneous Detection and Tracking with Motion Modelling for Multiple Object Tracking (Supplementary Material)

ShiJie Sun[1], Naveed Akhtar[2], XiangYu Song[3], HuanSheng Song[1], Ajmal Mian[2], and Mubarak Shah[4]

[1] Chang'an University, Xi'an, Shaanxi, China
{shijieSun,hshsong}@chd.edu.cn
[2] University of Western Australia, 35 Stirling Highway, Crawley, WA, Australia
{naveed.akhtar,ajmal.mian}@uwa.edu.au
[3] Deakin University, RWaurn Ponds, Victoria 3216, Melbourne, Australia
xiangyu.song@deakin.edu.au
[4] University of Central Florida, Orlando, FL, America
shah@crcv.ucf.edu

A Public Source Code & Dataset

Along with the submitted manuscript, we provide the source code of the proposed DMM-Net and publish the proposed Omni-MOT dataset. We also provide our implementation to generate more videos similar to the proposed dataset using the CARLA simulator [3]. Below, the links are provided for anonymous repositories for the sake of the review process. The links will be made public after the acceptance. (Click on the highlighted text to open the URL).

- DMM-Net is the source code of DMM-Net. It also contains the training and testing script for both UA-DETRAC [10] and Omni-MOT dataset, and instructions for reproducing the result of our methods.
- Omni-MOT Dataset provides the link to the dataset along with the related description.
- Omni-MOT Script is the source code for generating the Omni-MOT dataset and extending it. It includes the script for recording the MOT data and playing the recorded dataset.

B Videos for Dataset and Further Results

We provide the following videos for the review process. These and further videos will also be made public after the acceptance:

- Omni-MOT dataset videos illustrates different scenes, camera viewpoints and weather conditions used in the generated large-scale realistic dataset.
- Further results on Omni-MOT illustrates more tracking results on the proposed dataset.
- Further results on UA-DETRAC show tracking results on a representative scene from the UA-DETRAC challenge.

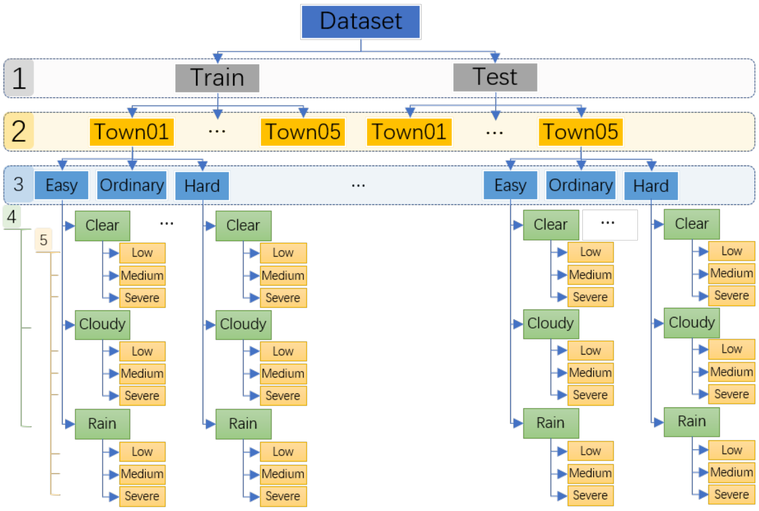


Fig. 1. Overall structure of the Omni-MOT dataset.

C Further Details of the Dataset

The structure of the proposed dataset can be best understood under five dimensions of divisions. We depict these dimensions as different levels of a block diagram in Fig. 1 for a clear overview. Along the first dimension, we split the dataset into the training and testing sets. Along the second, five towns of the CARLA simulator are employed for making the dataset diverse. For each town, we set the camera with different viewpoints for the third dimension of division. These viewpoints include three levels of difficulty, namely, Easy, Ordinary, and Hard level. Along the fourth dimension, different weather conditions split the data. These weather conditions contain Clear, Cloudy, and Rainy weather. The last variability that makes our data diverse consists of three congestion levels, namely; Low, Medium, and Severe congestion. Details are given below.

Train/Test split: The training set consists of $1,755$ videos, $8,775$K frames, 134.2K tracks, and 68.88M boxes. The testing set includes $1,755$ videos, $5,265$K frames, 122.37K tracks and 41.36M boxes.

Towns: There are five towns in our dataset, whose details are given in Table 1. Among these, Town05 is the largest city that also has three overpass roads. Town02 is the smallest city, whereas Town03 also contains a tube. Town04 is the most populous in terms of T-junctions.

Camera viewpoints: 39 cameras are placed in each city with different viewpoints. The camera horizontal field of view is 90°. We refer to the Omni-MOT dataset videos to visually observe the viewpoints.

Weather conditions: Three kinds of weather are simulated, namely, Clear, Cloudy, and Rainy, by changing the weather parameters of the CARLA simulator. These weather parameters include cloudiness and precipitation, and their values range from 0 to 100.

Table 1. Details of five towns. Column "Size" is manually measured and its format is $Width \times Length$

Name	Size (m)	Cross	T-junction	Roundabout	Overpass	Tunnel
Town01	342×413	0	12	0	0	0
Town02	205×208	0	9	0	0	0
Town03	438×483	5	14	2	0	1
Town04	816×914	8	21	3	1	0
Town05	430×486	13	8	0	3	0

Table 2. The number of vehicles for each congestion level of different towns.

City	Low	Medium	Severe
Town01	50	75	95
Town02	50	75	95
Town03	100	170	230
Town04	100	170	230
Town05	100	170	230

The cloudiness of Clear is 15, and the cloudiness of Cloudy and Rainy are 80. Both the precipitation of Clear and Cloudy are 0, and the precipitation of Rainy is 60.

Road congestion: We include three levels of traffic congestion, i.e. low, medium, and severe congestion. Because cities have different sizes, these congestion levels are decided by different numbers of vehicles. Table 2 summarizes the number of vehicles for the chosen level of congestion in all five towns.

In the proposed dataset, five different simulated cities are considered. For each city, we use up to 39 cameras. The cameras are placed with viewpoints that have three levels of difficulty for the MOT scenarios. Namely, (a) Easy view: which results in no occlusion of the vehicles. (b) Ordinary view: that allows temporary occlusions but forbids continuous occlusions. (c) Hard view: that allows continuous occlusions in the videos. Collectively, we provide 90 scenes in the dataset that result in $3,510$ videos. There are 14.04M frames of size 1920×1080 in the OMOT dataset that are recorded in the 'XDIV' format to provide high-quality videos with acceptable memory size.

C.1 Comparison to the existing MOT datasets

To put the proposed dataset into a better perspective, we also compare it to the existing popular datasets for the MOT task. In Table 3, we provide the comparison. Our dataset comprises 3,510 videos, 14M+ frames, 250K tracks, and 110M+ bounding boxes, whose frame number is almost 1,200 times larger than MOT17. The number of provided tracks and boxes are 210 and 30 times larger than UA-DETRAC. Besides, for the proposed Omni-MOT, all the boxes and tracks are automatically generated by the enumerator that avoids any labeling error. In the table, we include nuScenes [2] and Waymo [9] for the sake of comprehensive benchmarking. Nevertheless, these datasets are related to self-driving vehicles that are captured with moving cameras.

Table 3. Comparison with other popular MOT datasets. Columns “Frames“ is the number of frames ($1k = 10^3, 1M = 10^6$), “Tracks“ is the number of tracks and “Boxes“ is the number of bounding boxes. ”-” indicates that no information is provided.

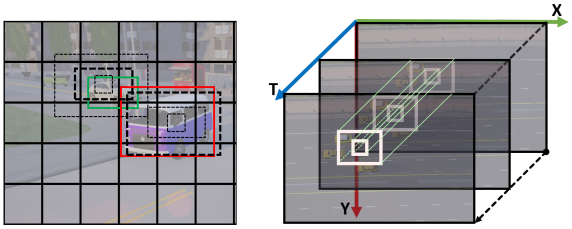
Dataset	Training			Testing		
	Frames	Tracks	Boxes	Frames	Tracks	Boxes
PETS [4]	-	-	-	1.5k	106	18.5k
KITTI [5]	8k	-	-	11k	-	-
TUD [1]	610	-	610	451	31	2.6k
MOT15 [6]	5.5k	500	39.9k	5.8k	721	61k
MOT17 [8]	5.3k	467	110k	5.9k	742	182k
UA-DETRAC [10]	84k	5.9k	578k	56k	2.3k	632k
nuScenes [2]	40k	-	1.4M	-	-	-
Waymo [9]	154k	-	8.6M	23k	-	1.3M
Omni-MOT(Ours)	**8775k**	**134.2k**	**68.88M**	**5265k**	**122.37k**	**41.36M**

Table 4. Data format of the ground truth file provided with the dataset.

Index	Name	Description
0	frame index	0-based frame index
1	vehicle id	Unique ID of vehicle (0-based)
2	bbox	Represents left, top, right, and bottom of the vehicle bounding box
6	3d bbox	The 8 points of the vehicle’s 3D bounding boxes in image coordinates
14	vehicle position	Vehicle’s center in the world coordinates
17	integrity	Integrity of the vehicle. Binary value in (0, 1)
18	velocity vector	Velocity vector in the world coordinates
21	acceleration vector	Acceleration vector in the world coordinates
24	wheel number	Number of vehicle wheels
25	camera view size	The width and the height of the camera view
27	camera FOV	The field of view of the camera
28	camera position	Camera position in the world coordinates
31	camera rotation	Camera rotation in the world coordinates
34	weather condition	Weather condition in the current frame.

C.2 On the ground-truth annotations

The ground-truth annotations is generated by the CARLA simulator, which allows us to capture comprehensive information on the target objects with high precision. Hence, besides being accurate, we target the ground truth data for not only multiple object tracking, but also other extended applications such as 3D estimation, velocity estimation, camera calibration, etc. To this end, we bring as much information as we can into the ground truth file. Table 4 gives details of the format of the ground truth files (available through the dataset download link provided above). The “3D bbox” at columns 6-13 contains values describing point coordinates. These points are the image projection of a minimum 3D cuboid envelope of the vehicle in the world coordinates. The column “bbox” is calculated by the minimum rectangle envelope of these points. On index 17, “integrity” encodes the visibility of vehicles. A clear description of the remaining entities is provided in the table.



(a) Anchor Boxes

(b) Anchor Tubes

Fig. 2. Illustration of the anchor tubes. The anchor boxes in (a) are a set of boxes in the 2D coordinates. The SSD model predicts the scaling and translation parameters relative to each anchor box. Extended from anchor boxes, the anchor tubes in (b) are a set of cuboids in the 3D coordinate. Each anchor tube consists of N_F boxes. Our proposed network predicts the tube shape offset parameters along the temporal dimension, the confidence for each object class, and the visibility of each box in the tube.

D Anchor Tubes

The proposed notion of anchor tube is an extension of the concept of anchor boxes in SSD [7]. In our technique, the anchor tubes are pre-defined and distributed at every position of the selected feature maps. An anchor tube is essentially a set of anchor boxes that share the same location in multiple frames along the temporal dimension, as illustrated in Fig. 2. The Fig. 2(a) depicts three pre-defined anchor boxes at each position of a 3-D feature map. The Fig. 2(b) illustrates a pre-defined anchor tube at a position of a 4-D feature map. Similar to the main goal of the anchor boxes, the anchor tubes format the network output dimensions. Consequently, our network is designed to predict the tube shape offset parameters along the temporal dimension, the confidence for each object class, and the visibility of each box in the tube.

E Further Details on Motion Model

The proposed DMM-Net outputs encoded anchor tubes directly. However, it entails predicting numerous parameters (unless a compact encoding for the tubes is used). For instance, assume that an anchor tube contains 16 boxes. In this case, the network would need to output 16x4 scalar values to describe the tube. To limit the output parameters, we introduce the motion function to describe these boxes in an encoded anchor tube. In our experiments, we use the quadratic function that only needs 3x4 motion parameters to describe an encoded anchor tube. The Eq. (1) below states the relationship between the motion parameters and the ground truth tracks.

$$\begin{cases} b^{w}_{i,t} = a^{w}_{i,t}\, exp(p_{11}t^2 + p_{12}t + p_{13} + \varDelta^{w}) \\ b^{h}_{i,t} = a^{h}_{i,t}\, exp(p_{21}t^2 + p_{22}t + p_{23} + \varDelta^{h}) \\ b^{cx}_{i,t} = p_{31}a^{w}_{i,t}t^2 + p_{32}a^{w}_{i,t}t + p_{33}a^{w}_{i,t} + a^{cx}_{i,t} + \varDelta^{cx} \\ b^{cy}_{i,t} = p_{41}a^{h}_{i,t}t^2 + p_{42}a^{h}_{i,t}t + p_{43}a^{h}_{i,t} + a^{cy}_{i,t} + \varDelta^{cy}, \end{cases} \tag{1}$$

Table 5. Further results on Omni-MOT with Medium and Hard camera views and Cloudy weather conditions: The symbol ↑ indicates higher values are better, and ↓ implies lower values are favored.

Type	View	**Camera**	**IDF1↑**	**IDP↑**	**IDR↑**	**Rcll↑**	**Prcn↑**	**GT↑**	**MT↑**	**PT↑**	**ML↓**	**FP↓**	**FN↓**	**IDs↓**	**FM↓**	**MOTA↑**	**MOTP↑**
Test	Hard	Camera_1	35.8%	41.8%	31.3%	55.8%	74.6%	116	9	34	73	3359	7801	78	184	36.3%	73.2%
	Hard	Camera_12	32.0%	38.6%	27.3%	48.0%	67.7%	136	1	38	97	2948	6704	113	188	24.2%	70.5%
	Ordinary	Camera_9	47.7%	51.1%	44.7%	68.5%	78.3%	61	16	32	13	2398	3986	42	150	49.2%	73.5%
Train	Hard	Camera_0	44.5%	48.2%	41.4%	63.7%	74.1%	137	17	36	84	2979	4865	66	155	40.9%	75.8%
	Hard	Camera_1	30.7%	36.8%	26.3%	49.7%	69.7%	148	5	38	105	3236	7537	117	202	27.4%	73.2%
	Ordinary	Camera_5	68.9%	70.9%	66.9%	81.5%	86.3%	94	32	40	22	3443	4957	62	222	68.3%	81.1%
Average	-		47.0%	52.0%	42.8%	63.5%	77.3%	692	80	218	394	18363	35850	478	1101	44.4%	76.1%

where $(a^{cx}_{i,t}, a^{cy}_{i,t}, a^{w}_{i,t}, a^{h}_{i,t})$ is the pre-defined box (center x, center y, width, height) of i^{th} anchor tube at t^{th} frame, $(b^{cx}_{i,t}, b^{cy}_{i,t}, b^{w}_{i,t}, b^{h}_{i,t})$ is the box of i^{th} ground truth track at t^{th} frame, $\{p_{11}, \cdots, p_{43}\}$ represents the motion parameters of the i^{th} encoded anchor tube, and $(\mathit{\Delta}^{cx}, \mathit{\Delta}^{cy}, \mathit{\Delta}^{w}, \mathit{\Delta}^{h})$ is the localization error of our network output. From Eq. (1), we can see that the box center $(b^{cx}_{i,t}, b^{cy}_{i,t})$ of the ground truth track is a quadratic function of time, while $(b^{w}_{i,t}, b^{h}_{i,t})$ is more a complicated function. The motion parameters are able to successfully model object motion for short time slots to generate effective tracklets.

F Further Quantitative Results

To better evaluate our technique and putting the values reported in the paper into a better perspective, we provide further results of DMM-Net on two additional viewpoints Omni-MOT. These results are reported in Table 5. The experiments are conducted for cloudy weather conditions. The selected scenes are from Town 05 (with 230 vehicles) that are indexed 1, 9, and 12 in the dataset, where scene-1 and scene-12 are with hard camera view, and scene-9 is with ordinary camera view. Similar to the experiments in the paper, we train the network for 22 epochs and use the same evaluation matrices as used in the paper. The results show good performance of DMM-Net on the realistic dataset with accurate ground-truth. For these experiments, we observed that our network was often able to track vehicles that are fully occluded. This is a direct benefit of using motion modeling for tracking. On the flip side, we also observed a slight drift of the bounding boxes for stationary objects due to the amplification of motion caused by noisy detection. Nevertheless, this problem was never observed to cause critical problems. These observations can be verified in the videos provided on the URL links above.

References

1. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2008), https://academic.microsoft.com/paper/2138302688
2. 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 (2019)
3. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An Open Urban Driving Simulator. In: Proceedings of the 1st Annual Conference on Robot Learning. pp. 1–16 (2017), http://arxiv.org/abs/1711.03938
4. Ferryman, J., Shahrokni, A.: PETS2009: Dataset and challenge. In: Proceedings of the 12th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS-Winter 2009 (2009). https://doi.org/10.1109/PETS-WINTER.2009.5399556
5. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3354–3361 (2012). https://doi.org/10.1109/CVPR.2012.6248074
6. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. arXiv:1504.01942 [cs] pp. 1–15 (2015), http://arxiv.org/abs/1504.01942
7. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single shot multibox detector. ECCV **9905 LNCS**, 21–37 (2016). https://doi.org/10.1007/978-3-319-46448-0_2
8. Milan, A., Leal-Taixe, L., Reid, I., Roth, S., Schindler, K.: MOT16: A Benchmark for Multi-Object Tracking. CoRR **abs/1603.0** (2016), http://arxiv.org/abs/1603.00831
9. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset (2019)
10. Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M.C., Qi, H., Lim, J., Yang, M.H., Lyu, S.: UA-DETRAC: A New Benchmark and Protocol for Multi-Object Detection and Tracking (2015), http://arxiv.org/abs/1511.04136