IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 14, 2021, accepted March 22, 2021, date of publication March 24, 2021, date of current version April 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068899

Driving Behavior-Aware Network for 3D Object Tracking in Complex Traffic Scenes

QINGNAN LI[1,2], RUIMIN HU[2,3], (Senior Member, IEEE), ZHONGYUAN WANG[3,4], (Member, IEEE), AND ZHI DING[1]

[1]Engineering Research Center for Transportation Systems, School of Art and Design, Wuhan University of Technology, Wuhan 430070, China
[2]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China
[3]Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China
[4]Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China

Corresponding author: Ruimin Hu (hrm@whu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61502348, Grant 61671336, and Grant 91738302; in part by the Natural Science Foundation of Jiangsu Province under Grant BK20180234.

ABSTRACT Recently a large number of 3D object tracking methods have been extensively investigated and applied in a variety of applications using convolutional neural networks. Although most of them have made great progress in partial occlusion, the intricate interweaving of moving agents (e.g. pedestrians and vehicles) may lead to inferior performance of 3D object tracking in complex traffic scenes. To boost the performance of 3D object tracking in cases of severe occlusions, we present an end-to-end deep learning framework with a driving behavior-aware model that takes full advantage of spatial-temporal details in consecutive frames and learns the driving behavior from object variations in 2D center point, depth, rotation and translation in parallel. In contrast to prior work, our novelty formulates driving behavior that reasons about the possible motion trajectories of the investigated target for autonomous systems. We show in experiments that our method outperforms state-of-the-art approaches on 3D object tracking in the challenging nuScenes dataset.

INDEX TERMS 3D object tracking, driving behavior, object offsets, rotation, translation.

I. INTRODUCTION

Multi-object tracking (MOT), also called multi-target tracking (MTT), is an essential component technology in many computer vision applications such as autonomous driving [1]–[3] and robot collision prediction [4], [5]. Given a set of measurements from onboard sensors, MOT perceives road agents and surrounding environment using spatial-temporal details to identify and track objects, such as vehicles, pedestrians, etc., without any prior knowledge about object properties, shape parts, or environment variations such as lighting and weather conditions. Though a wide array of views and sensors have enabled depth information to be well exploited by many MOT techniques, onboard cameras are much cheaper and offer the promise to provide enough spatial-temporal details for detection and tracking since human observers have no difficulty in perceiving 3D world in both space and time. In this paper, we focus on 3D object tracking in video data, especially for objects that are subject to heavy occlusion in complex traffic scenes.

Impressive progress has been made over the last decade towards solving the fundamental MOT problem. The current literature on 3D object tracking can be divided into two groups, global tracking and online MOT. The first group of methods [6]–[9] assumes that all of the frames are available for processing. The idea is similar to bidirectional prediction proposed in H.264 [10] and HEVC [11], with spatial-temporal information from two directions, making the tracking process bidirectional. Though good performances have been achieved, these approaches can not afford to run real-time applications online for MOT. The second group of methods [2], [3], [12]–[16] makes use of the information upto current frame without the assumption of any prior knowledge of future frames. These approaches only rely on forward prediction but are more suitable for online tracking and real-time applications. Driven by the success of deep learning techniques, many recent approaches [2], [3], [16]–[18] generate deep features and show much better performance than hand-crafted representations [12]–[15] in these

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Ayoub Khan.

applications and furthermore exhibit a better speed-accuracy trade-off.

Data-driven MOT approaches can be further divided into two subgroups: (i) methods that are based on tracking by detection paradigm. (ii) methods that jointly perform detection and tracking. The former methods [17], [19]–[23] mainly focus on appearance feature extractor and data association. Appearance feature extractors are used to detect the locations of road agents in the form of bounding boxes from each individual frame, and then data association algorithms are proposed to associate the detected bounding boxes or additional target features across frames. These methods take advantage of the detection of individual frames. However, the detection is separated from tracking, which ignores the motion features in spatial-temporal details between consecutive frames. The latter methods [2], [3], [16] generate deep features in consecutive frames and jointly detect and track objects, which can integrate multiple cues across time such as motion features, appearance features, and interactive features that help object detection and tracking under heavy occlusion in complex traffic scenes. In this paper, we follow the paradigm of the latter methods whereby object detection and tracking are jointly processed.

Compared with 2D object tracking [17], [18], 3D object tracking [2], [3], [16] provides more spatial details for environmental perception [24], [25] in the areas of autonomous vehicles and advanced driver-assistance systems. Such methods take full use of not only the knowledge of part-whole intrinsic spatial relationships in each individual frame but also spatial-temporal details between consecutive frames, with good performance on 3D object tracking challenge in nuScenes [25] dataset. Especially, the CenterTrack [3], which assumes the objects as points and predicts the location offsets to associate objects, has achieved the competitive performance. However, the CenterTrack, based solely on supervision in the form of object 2D center offset across time, still suffers from ID switches when both appearance and motion features of the investigated target are starting to change under different occlusion levels in complex traffic scenes.

Many approaches explore knowledge-based driving behavior and teach machines to understand how the physical world is unfolding [26], [27]. Inspired by the prior works [28]–[30], we consider a natural formulation that the movements of road agents with different poses and scales are determined by human driving behavior. Based on this natural formulation, instead of encoding object center offsets on 2D plane for 3D tracking [3], we take full advantage of spatial-temporal details across consecutive frames and propose an end-to-end deep learning framework to learn the driving behavior from variations in 2D center point, depth, rotation and translation in the magnitude and direction of hidden-state vectors. By exploring such high-level driving behavior knowledge in CNN representations, our framework has a clear advantage over methods that are based on object center or bounding box offsets. Key to our approach is that the learned driving behavior aims to reason about the possible motion trajectories of the investigated target in heavy occluded or even worst-case traffic scenarios. Concretely, our framework processes the object variations in 2D center point, depth, rotation and translation in parallel, as illustrated in Fig. 1, from which we conduct driving behavior-aware transformation loss functions to formulate high-level driving behavior in consecutive frames, guiding the road agents movements in any space and time.

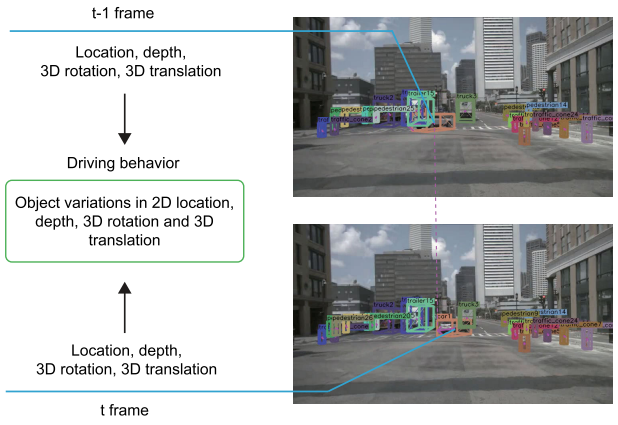

FIGURE 1. System outputs. We learn not only the object 2D center offset but also the object variations in depth, 3D rotation and translation in consecutive frames for driving behavior exploration.

We evaluate our method on the nuScenes dataset [25]. To ensure a fair comparison, we follow the prior work [3] and use the same model parameters released by CenterNet [31] for DLA [32] network backbone without any prior detections. We show experiments that our method outperforms the state-of-the-art CenterTrack by 0.042 and 0.026 for nuScenes validation and test set, respectively, using AMOTA metric. In summary, our end-to-end deep learning framework achieves significantly better results, especially for road agents under heavy occlusion in complex traffic scenes.

The key contributions of our work are as follows:

- An end-to-end deep learning network is proposed to learn the driving behavior from the object variations in 2D center point, depth, rotation and translation in consecutive frames.
- The learned driving behavior aims to reason about the possible motion trajectories of the investigated target in complex traffic scenes, which is contributed to improve the overall 3D tracking performance rather than a solely learned object 2D center offset.
- Our driving behavior-aware network is tested on the EvalAI nuScenes tracking online evaluation server where it outperforms the state-of-the-art approaches in terms of AMOTA.

II. PRELIMINARIES

Our method follows CenterTrack [3] and builds on the CenterNet [31] for 3D object detection, in which a single image $I \in \mathbb{R}^{H\times W\times 3}$ is taken as input and a set of detections

$\{(\hat{p}_i, \hat{s}_i^{2d}, \hat{s}_i^{3d}, \hat{d}_i, \hat{e}_i)\}_{i=0}^{N-1}$ is produced for each class $c \in \{0, \ldots, C-1\}$, where $\hat{p}_i$, $\hat{s}_i^{2d}$, $\hat{s}_i^{3d}$, $\hat{d}_i$, and $\hat{e}_i$ denote the i-th predicted object center point, the 2D bounding box size, the 3D bounding box size, the depth, and the orientation respectively. For all of the classes C, our network produces low-resolution heatmap $\hat{Y} \in [0,1]^{\frac{H}{R} \times \frac{W}{R} \times C}$, 2D bounding box heatmap $\hat{S}^{2d} \in \mathbb{R}^{\frac{H}{R} \times \frac{W}{R} \times 2}$, 3D bounding box heatmap $\hat{S}^{3d} \in \mathbb{R}^{\frac{H}{R} \times \frac{W}{R} \times 3}$, $\hat{D} \in \mathbb{R}^{\frac{H}{R} \times \frac{W}{R}}$, orientation heatmap $\hat{E} \in \mathbb{R}^{\frac{H}{R} \times \frac{W}{R} \times 8}$ for the i-th object, with a output stride $R = 4$. Each peak $\hat{p} \in \mathbb{R}^2$ in a prediction $\hat{Y}$ indicates the most likely 2D location of an object, with the corresponding confidence $\hat{w} = \hat{Y}_{\hat{p}}$.

We use the focal loss [33] to minimize the detection errors:

$$\mathcal{L}_{fl} = \frac{1}{N} \sum_{xyc} \begin{cases} (1-\hat{Y}_{xyc})^{\alpha} log(\hat{Y}_{xyc}) & if\ Y_{xyc} = 1 \\ (1-Y_{xyc})^{\beta} (\hat{Y}_{xyc})^{\alpha} & \\ log(1-\hat{Y}_{xyc}) & otherwise \end{cases} \tag{1}$$

where N is the number of predicted object center points, and Y_{xyc} denotes a ground-truth object center point rendered heatmap using a Gaussian kernel

$$Y_{xyc} = exp(-\frac{(x-\tilde{p}_x)^2 + (y-\tilde{p}_y)^2}{2\sigma_p^2}) \tag{2}$$

at 2D location (x, y) for class c. $\tilde{p}$ is a low-resolution representation $\tilde{p} = \lfloor \frac{p}{R} \rfloor$ with the downsampling stride R, where $p \in \mathbb{R}^2$ denotes each ground-truth keypoint. σ_p is an object size-adaptive standard deviation [3], [31], [34]. The prediction $\hat{Y}_{xyc} = 1$ corresponds to the object center point, while $\hat{Y}_{xyc} = 0$ is the background. The hyper-parameters of focal loss $\alpha = 2$ and $\beta = 4$ are used in our network, following the prior work CornerNet [34], CenterNet [31], and CenterTrack [3].

The 2D object size prediction is regressed by minimizing the size errors using the following function:

$$\mathcal{L}_{2ds} = \frac{1}{N} \sum_{i=1}^{N} |\hat{S}_{\tilde{p}_i}^{2d} - s_i^{2d}| \tag{3}$$

where N is the number of predicted object center points in image I, and $\hat{S}_{\tilde{p}_i}$ denotes the i-th object 2D bounding box size predicted from deep features of size output heatmap $\hat{S} \in \mathbb{R}^{\frac{H}{R} \times \frac{W}{R} \times 2}$ at the ground-truth location $\tilde{p}_i$, while s_i^{2d} indicates the ground-truth size of i-th object 2D bounding box.

A local offset $\hat{F} \in \mathcal{R}^{\frac{W}{R} \times \frac{H}{R} \times 2}$ is additionally proposed to recover the discretization error caused by the output stride R, trained with L1 loss:

$$\mathcal{L}_{off} = \frac{1}{N} \sum_{p} |\hat{F}_{\tilde{p}} - (\frac{p}{R} - \tilde{p})| \tag{4}$$

For 3D object bounding box size prediction, we add an additional channel $\hat{S}^{3d} \in \mathbb{R}^{\frac{H}{R} \times \frac{W}{R} \times 3}$ trained with L1 Loss in absolute metric:

$$\mathcal{L}_{3ds} = \frac{1}{N} \sum_{i=1}^{N} |\hat{S}_{\tilde{p}_i}^{3d} - s_i^{3d}| \tag{5}$$

where s_i^{3d} denotes the 3D bounding box size of the i-th object.

The depth output channel $D \in \mathbb{R}^{\frac{W}{R} \times \frac{H}{R}}$ consists of two convolutional layers separated by a ReLU using the inverse sigmoidal transformation at the output layer. We use the output transformation proposed by Eigen *et al.* [35] $d = \frac{1}{\sigma(\hat{d}_i)} - 1$ to minimize the depth errors using the following function:

$$\mathcal{L}_{dep} = \frac{1}{N} \sum_{i=i}^{N} |\frac{1}{\sigma(\hat{d}_i)} - 1 - d_i| \tag{6}$$

where N is the number of predicted object center points, and d_i denotes the ground-truth absolute depth.

Following the prior works [2], [3], [31], [36], the orientation θ prediction is to solve a fundamental softmax classification problem. An 8-scalar encoding scheme is proposed to transform the orientation θ into 8 scalars for classification with L1 loss:

$$\mathcal{L}_{orie} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{2} (softmax(\hat{b}_k, c_k) + c_k |\hat{a}_k - a_k|) \tag{7}$$

where N is the number of predicted center points in image I, and k indicates to one of the angular bins $B = \{B_1, B_2\}$, in which $B_1 = [-\frac{7\pi}{6}, \frac{\pi}{6}]$ and $B_2 = [-\frac{\pi}{6}, \frac{7\pi}{6}]$. Two scalers $b_k \in \mathbb{R}^2$ in each angular bin are used for softmax classification, while the rest scalers $a_k = (sin(\theta - m_k), cos(\theta - m_k))$ are serves as in-bin offset to the bin center $m_k = \mathbb{I}(\theta \in B_k)$. At inference time, the decoding scheme is proposed to recover the predicted orientation θ transformed from such 8-scalers using the following equation:

$$\hat{\theta} = arctan2(\hat{a}_{j1}, \hat{a}_{j2}) + m_j \tag{8}$$

where j is the index of the highest confidence in softmax classification.

Thus, we have detailed the objective loss functions $\mathcal{L}_{fl}$, $\mathcal{L}_{2ds}$, $\mathcal{L}_{3ds}$, $\mathcal{L}_{off}$, $\mathcal{L}_{dep}$ and $\mathcal{L}_{orie}$ for object detections, including object localization, 2D/3D bounding box size regression, orientation classification, etc.

III. METHOD

Object motions, such as rotation, acceleration or deceleration, driven by human behavior in complex traffic scenes, play an important role in 3D object tracking. In this section, we propose an end-to-end deep learning network with driving behavior-aware architecture and corresponding loss functions for driving behavior exploration. We first introduce the overview of our network architecture, and then detail the driving behavior-aware architecture and conduct spatial-temporal relative transformation loss functions. Our framework aims to learn the high-level driving behavior knowledge from the motions of road agents in consecutive frames.

A. ARCHITECTURE OVERVIEW

The overview of our network architecture is shown in Fig. 2, which takes current image $I^{(t)} \in \mathbb{R}^{H \times W \times 3}$, previous image $I^{(t-1)} \in \mathbb{R}^{H \times W \times 3}$, and a heatmap rendered

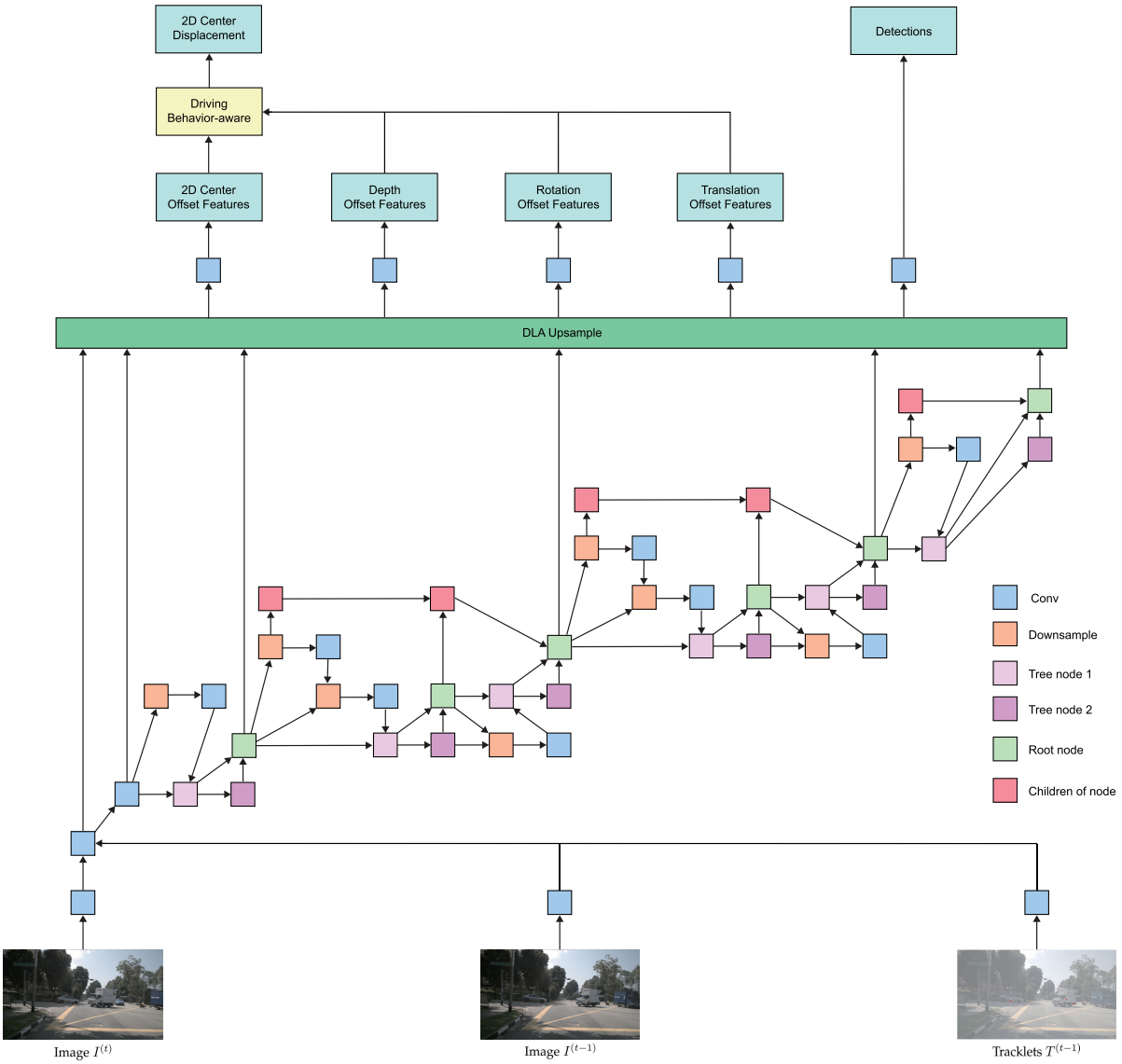

FIGURE 2. The overview of our driving behavior-aware network. The current frame, the previous frame, and a heatmap rendered from tracked object centers are sent into the deep layer aggregation structure with hierarchical and iterative skip connections. The architecture consists of the standard convolutional layers, downsampling layers, tree nodes, root nodes and children of nodes. The iterative deep aggregation is used in the upsampling layers for dense prediction. The driving behavior is learned from object variations in 2D center point, depth, rotation and translation in parallel.

from tracked objects in the previous image $T^{(t-1)} = \{b_0^{(t-1)}, b_1^{(t-1)}, b_2^{(t-1)}, \ldots, b_N^{(t-1)}\}$ as inputs, where $b_i^{(t-1)} = (p, s^{2d}, s^{3d}, d, e, w, id)$ indicates the i-th tracked object described by its 2D center location $p \in \mathbb{R}^2$, 2D bounding box size $s^{2d} \in \mathbb{R}^2$, 3D bounding box size $s^{3d} \in \mathbb{R}^3$, depth $d \in \mathbb{R}$, orientation $e \in \mathbb{R}^8$, detection confidence w, and the unique identity $id \in \mathbb{I}$. We use Deep Layer Aggregation (DLA) [32] as network backbone, and create convolutional heads of object center point, 2D and 3D bounding box size, depth, and orientation for 3D object detection. The DLA structure can fuse information across layers with hierarchical and iterative skip connections to make networks with better accuracy and fewer parameters. We use DLA-34 for a good trade-off between time complexity and tracking performance. We use the driving behavior-aware hierarchical architecture to learn the object variations in 2D center point, depth, rotation, and translation for driving behavior exploration in 3D object tracking challenge.

B. DRIVING BEHAVIOR-AWARE HIERARCHICAL ARCHITECTURE

The motion property of each object in complex traffic scenes is an important cue for tracking targets that are occluded or lost. One key challenge is to handle the intricate interweaving of target and its neighboring interference objects under occlusion, where the motion of the target may be non-linear, especially if we reason on several motion components. The motion components of a road agent can be analyzed on the

basis of valuable instance-aware semantic information such as object 2D location, depth, pose, velocity and their corresponding relative variations, such as 2D displacement, depth offset, rotation offset and translation offset in consecutive frames. By exploring such semantic information and their relative variations, we discover the latent geometric consistency from two views of the same object. Inspired by this natural formulation, our proposed driving behavior-aware hierarchical architecture is able to learn this non-linearities from consecutive frames, and build the relationships between motion components and corresponding relative variations, formulating driving behavior that contributes 3D object tracking in complex traffic scenes.

As for existing works, they associate objects through time by producing an object 2D center offset heatmap $\hat{O}^{cp} \in \mathbb{R}^{\frac{H}{R}\times\frac{W}{R}\times 2}$. With a 2D displacement offset prediction, they simply associate objects across time. However, our motivation is to learn the high-level driving behavior knowledge from object motions in consecutive frames. In order to formulate object motions in CNN representations, we conduct our driving behavior-aware architecture hierarchically merge the feature hierarchy from object depth offset heatmap $\hat{O}^{dep} \in \mathbb{R}^{\frac{H}{R}\times\frac{W}{R}}$, 3D rotation offset heatmap $\hat{O}^{rot} \in \mathbb{R}^{\frac{H}{R}\times\frac{W}{R}\times 4}$, 3D translation offset heatmap $\hat{O}^{tra} \in \mathbb{R}^{\frac{H}{R}\times\frac{W}{R}\times 3}$ and object 2D center offset deep features. The driving behavior-aware deep features can be defined as:

$$G_{beh} = CAT(\hat{O}^{cp}, \hat{O}^{dep}, \hat{O}^{rot}, \hat{O}^{tra}) \tag{9}$$

where $\hat{O}^{cp}$, $\hat{O}^{dep}$, $\hat{O}^{rot}$, $\hat{O}^{tra}$ denote the deep features that represent object variations in 2D center point, depth, 3D rotation and translation respectively. Compared with the state-of-the-art CenterTrack framework that is based solely on object 2D displacement supervised feature representations, our driving behavior-aware hierarchical architecture encodes object motion components and object variations in consecutive frames, producing a sufficiently better high-level knowledge-based 2D displacement offset for 3D object tracking in complex traffic scenes.

C. BEHAVIOR-AWARE RELATIVE TRANSFORMATION LOSS

In this section, we detail the key techniques of behavior-aware relative transformation functions across consecutive frames, which stand in contrast to previous networks under the supervision of a simple object 2D displacement.

The current state-of-the-art 3D tracking framework [3], designed to focus on minimizing the residual error between the ground-truth and the predicted object 2D displacement in the absence of any other motion components, suffers from severe degradation of performance or even failure in the presence of heavy occluded scenes. Instead, our framework learns not only the object 2D center offset but also the object variations in the depth, rotation and translation driven by human behavior and formulates high-level driving behavior knowledge that contributes to 3D object tracking for autonomous driving systems. Concretely, we focus on minimizing the residual errors of object variations in 2D center point, depth, rotation, and translation. For each object at ground-truth location $p^{(t)}$, the offsets $\hat{O}^{cp}_{p^{(t)}}$, $\hat{O}^{dep}_{p^{(t)}}$, $\hat{O}^{rot}_{p^{(t)}}$, $\hat{O}^{tra}_{p^{(t)}}$ capture the differences of 2D center point, depth, rotation and translation in the current frame and the previous frame respectively, from which the high-level driving behavior knowledge is learned in our framework.

We learn object 2D displacement using the same regression objective as size or location refinement:

$$L_{ocp} = \frac{1}{N}\sum_{i=1}^{N} |\hat{O}^{cp}_{p_i^{(t)}} - (p_i^{(t-1)} - p_i^{(t)})| \tag{10}$$

where $\hat{O}^{cp}_{p_i^{(t)}}$ denotes the predicted 2D displacement of i-th object at time t, while $p_i^{(t-1)}$ and $p_i^{(t)}$ are the ground-truth 2D center location of i-th object. Likewise, the training loss for depth offset is defined as follows:

$$L_{odep} = \frac{1}{N}\sum_{i=1}^{N} |\hat{O}^{dep}_{p_i^{(t)}} - (d_i^{(t-1)} - d_i^{(t)})| \tag{11}$$

where the $\hat{O}^{dep}_{p_i^{(t)}}$ denotes the predicted depth offset of i-th object at time t, while $d_i^{(t-1)}$ and $d_i^{(t)}$ are the ground-truth depth of i-th object at time $(t-1)$ and t respectively.

Since synchronized keyframes are sampled at a fixed frame rate in nuScenes dataset [25], we can transform the motion components in consecutive frames, from vector-based representation of quaternion offset and velocity offset to relative rotation and translation offsets. Thus, the relative rotation loss L_{orot} is defined as follows:

$$L_{orot} = \frac{1}{N}\sum_{i=1}^{N} 2arcsin\left(\frac{1}{2\sqrt{2}}||\hat{O}^{rot}_{p_i^{(t)}} - \mathcal{R}_i^{(t-1,t)}||_F\right) \tag{12}$$

where the offset $\hat{O}^{rot}_{p_i^{(t)}}$ is the i-th object relative rotation matrix of the predicted vector-based quaternion representation at the ground-truth local location $p_i^{(t)}$ at time t, while the residual $\mathcal{R}_i^{(t-1,t)}$ is defined as:

$$\mathcal{R}_i^{(t-1,t)} = R_i^{(t)}(R_i^{(t-1)})^{-1} \tag{13}$$

where $(R_i^{(t-1)})^{-1}$ is the inverse of $R_i^{(t-1)}$, which is the i-th ground-truth object rotation matrix of vector-based quaternion at time $t-1$, and likewise for the ground-truth $R_i^{(t)}$. On the other hand, the relative translation loss L_{otra} is defined as follows:

$$L_{otra} = \frac{1}{N}\sum_{i=1}^{N} |\hat{O}^{tra}_{p_i^{(t)}} - (\gamma_i^{(t-1)} - \gamma_i^{(t)})| \tag{14}$$

where $\hat{O}^{ts}_{p_i^{(t)}}$ denotes the predicted translation offset, while $\gamma_i^{(t-1)}$ and $\gamma_i^{(t)}$ are the ground truth translation of i-th object at time $t-1$ and t respectively.

Having defined the above relative transformation losses $L_{ocp}, L_{odep}, L_{orot}$ and L_{otra}, the overall loss for behavior-aware relative transformation can be written as:

$$L_{beh} = L_{ocp} + L_{odep} + L_{orot} + L_{otra}. \tag{15}$$

By exploring the object variations in motion components that consist of 2D center offset, depth offset, rotation and translation offset in consecutive frames, our framework in contrast to prior work [3] that aims to formulate driving behavior for efficient 3D object tracking with a finer 2D displacement. We then use a simple greedy matching algorithm to associate objects across time. For i-th object at position $\hat{p}_i^{(t)}$ at time t, we greedily associate it with the closest unmatched object at position $\hat{p}_i^{(t)} - \hat{D}_{\hat{p}_i^{(t)}}$, in descending order of confidence w. A new tracklet will be assigned if there is not any matched prior detection within a threshold τ, which is defined as the geometric mean of width and height of the predicted bounding boxes.

IV. EXPERIMENTS

To demonstrate our end-to-end deep network robust to heavy occlusion in complex traffic scenes, we evaluate our method on the challenging nuScenes [25] dataset presented in Sec. IV-A. The corresponding results are reported in Sec. IV-D, where the two main metrics AMOTA, AMOTP and the secondary metrics MT, ML, IDS, FP and FN, etc., are used for evaluation, detailed in Sec. IV-B. We also present our implementation details in Sec. IV-C and the analysis on our driving behavior-aware representations in Sec. IV-E.

A. DATASETS

The nuScenes dataset is a public large-scale dataset for autonomous driving. It consists of 1000 scenes of 20s duration each, and keyframes are sampled at 2Hz in each scene with 6 slightly overlapping images in a panoramic 360° view, resulting in 168k training, 36k validation, and 36k test images. All of the 23 object classes are annotated in the form of cuboids modeled as x, y, z, width, length, height, yaw angle and other properties such as visibility, activity, and pose. We follow the baseline [37] and current state-of-the-art CenterTrack [3], and use the annotated keyframes for training and validation. We also evaluated our proposed driving behavior-aware network on the nuScenes [25] test set by submitting tracking results to the EvalAI tracking online evaluation server.

B. METRICS

AMOTA [3], [25], [37], average multi object tracking accuracy, compared with the common multi-object tracking accuracy [39], [40], is a weighted average of MOTA across different output thresholds, defined as follows:

$$AMOTA = \frac{1}{n-1} \sum_{r \in \{\frac{1}{n-1}, \frac{1}{n-2}, \ldots, 1\}} MOTAR$$
$$MOTAR = max(0, 1 - \alpha \frac{IDS_r + FP_r + FN_r - (1-r) \times P}{r \times P}) \tag{16}$$

where the n-point interpolation $n = 40$. The parameters $\alpha = 0.2$ (AMOTA@0.2) and $\alpha = 0.1$ (AMOTA@0.1) are set by the nuScenes [25] benchmark. The IDS_r, FP_r, and FN_r denote the total number of identity switches, false positives, and false negatives respectively, all of which only consider top confident samples that achieve the recall threshold r. P refers to the total number of ground-truth positives among all frames.

AMOTP [3], [25], [37], average multi object tracking precision, is defined as follows:

$$AMOTP = \frac{1}{N} \sum_{r \in \{\frac{1}{n-1}, \frac{1}{n-2}, \ldots, 1\}} \frac{\sum_{i,t} d_{i,t}}{\sum_t TP_t} \tag{17}$$

where $d_{i,t}$ indicates the position error of track i at time t, and TP_t is the number of matches at time t.

C. IMPLEMENTATION DETAILS

Our driving behavior-aware network consists of DLA [32] backbone, CenterTrack heads [3], and our proposed behavior-aware architecture, implemented using Pytorch and optimized with Adam using learning rate $4e-5$ and batchsize 10. Data augmentations include random horizontal flip, random scale, cropping, and color jittering, while rendering pipeline [41], tensor completion [42] or image inpainting [43] can be further leveraged by 3D object tracking framework to handle heavy occlusions in future work. We train our network on a machine with an Intel E5-2680v4 and 1 TitanXp GPU. The network is trained for 320 epochs with a learning rate drop at 300 epochs by a factor 10.

Our network follows CenterTrack [3] that uses nuScenes input resolution 800×448 from all the 6 cameras and fuses network outputs without handling duplicate detections at the intersection of views [38]. The hyperparameters are set at $\lambda_{fp} = 0.1$, $\lambda_{fn} = 0.4$, with the output threshold $\theta = 0.1$ and the heatmap rendering threshold $\tau = 0.1$. The loss weights for variations in 2D center point, depth, rotation and translation are set to 1 while the rest loss weights are set the same way as CenterTrack.

D. EVALUATION ON nuScenes DATASET

We compare our approach with the official monocular 3D tracking baseline Mapillary [38] + AB3D [37], and the current state-of-the-art method CenterTrack [3] in nuScenes [25] validation and test set. The AMOTA, AMOTP, MOTAR, MT, ML, IDS, FP and FN are reported to evaluate the performances on the nuScenes [25] validation and test set, which are listed in Table 1 and Table 2. Our driving behavior-aware framework outperforms the state-of-the-art CenterTrack in both validation and test set. More detailed results in terms of MOTA are listed in Table 3.

Qualitative results of 3D object detection and tracking are predicted from four video clips. The first video clip, from nuScenes [25] dataset, is adopted by CenterTrack [3] for visualization. The second and third video clips are from

TABLE 1. Quantitative evaluation of 3D object tracking on nuScenes validation set.

	AMOTA ↑	AMOTP ↓	MOTAR ↑	MT ↑	ML ↓	IDS ↓	FP ↓	FN ↓
CenterTrack [3]	0.068	1.543	**0.332**	538	4231	2962	15733	73573
Ours	**0.110**	**1.485**	0.325	592	**4120**	**2214**	**15582**	**71614**

TABLE 2. Quantitative evaluation of 3D object tracking on nuScenes test set.

	AMOTA ↑	AMOTP ↓	MOTAR ↑	MT ↑	ML ↓	IDS ↓	FP ↓	FN ↓
Mapillary [38] +AB3D [37]	0.018	1.790	0.091	499	4700	10420	113596	83202
GDLG	0.045	1.819	0.242	448	**4094**	12040	40742	78327
CenterTrack [3]	0.046	1.543	0.231	573	5235	**3807**	**17574**	89214
Ours	**0.072**	**1.489**	**0.349**	**855**	4718	3831	19874	**81843**

TABLE 3. Quantitative results on the nuScenes test set in terms of MOTA.

	Bicycle	Bus	Car	Motorcycle	Pedestrian	Trailer	Truck
CenterTrack [3]	-	**0.072**	0.202	0.011	0.030	-	**0.004**
Ours	**0.024**	0.030	**0.266**	**0.083**	**0.087**	**0.016**	0.002

nuScenes [25] dataset either, considering adverse weather and extreme lighting conditions. The fourth video clip is captured from an in-car camera in a tree-lined road scene, where various random light spots are formed by the light that passes through the trees, resulting light intensity variations in a short time. Note that the position of in-car camera has been changed, which is different from roof mounted cameras used in nuScenes dataset.

Qualitative results predicted from the first video clip are shown on the upper side of the Fig. 3. The 3D object detections of the black car and the silver car in the top row are inferior to that in the bottom row, which shows that our driving behavior-aware framework, compared with the current state-of-the-art CenterTrack [3], has a significant improvement on 3D object detection task. Furthermore, the colors (vehicle IDs) of 3D bounding box of the black car in the top row have changed 4 times while that in the bottom row have changed only once, which shows that our driving behavior-aware framework outperforms the CenterTrack considerably as far as both 3D object detection and tracking in complex traffic scenes.

Qualitative results predicted from the second video clip are shown on the middle side of the Fig. 3. CenterTrack [3] fails to detect the truck in the second column and our approach has always been tracking this white truck across time, from the start to the end, with a unique ID 1002, while the CenterTrack lost this truck and assigned a new ID 1086 from ID 1083, which shows that our method outperforms the CenterTrack on 3D object tracking task in heavily occluded scenes under inclement weather conditions.

Qualitative results predicted from the third video clip are shown on the bottom side of the Fig. 3. In the night-time scene, most of color and texture information of target objects are lost, which presents a challenge to the network because many objects (e.g. cars and bicycles) are symmetric across at least one axis. From different viewpoints, the objects may appear visually identical especially in the night-time scene, resulting in ambiguous poses with respect to an azimuth rotation of π. Furthermore, the truncation level of the target is increasing gradually, which is a key technical challenge in performing 3D object detection and tracking in complex traffic scenes. For 3D object detection challenge, our method first detects the white car in the fourth column while the CenterTrack [3] does not detect this white car, which shows that our approach is more robust to night-time illumination. For 3D object tracking challenge, CenterTrack [3] fails to track the highly truncated target in the fifth column while our approach is able to track it, which shows that our approach is more robust to object truncation under heavy truncated scenes.

Qualitative results predicted from the fourth video clip are shown Fig. 4. The generalization capabilities of different algorithms are compared from the fourth video clip. Both the white and blue car have an increasing truncation level in a tree-lined road scene where the light intensity is increasing during this period. Compared with the CenterTrack [3] trained on nuScenes [25] dataset, our approach tracks both the blue and white car one more frame visualized in the third and fifth columns of Fig. 4, which shows that our method, also trained on nuScenes dataset, has an improvement on generalization and robustness to nonlinear illumination variations across consecutive frames.

E. ANALYSIS

In this section, we study the effects of our proposed driving behavior-aware architecture discussed in Section III-B. Our driving behavior-aware network explores object variations in 2D center point, depth, rotation and translation in consecutive frames, and formulates the driving behavior contributed to 3D object tracking in complex traffic scenes. In detail, we evaluate our 3D object tracking performances by computing MOTAR-Recall curves, MOTA-Recall curves and MOTP-Recall curves, as shown in Fig. 5.

FIGURE 3. Qualitative results on nuScenes dataset. The arrows in each image are not the vehicle orientations, however, such arrows indicate the current object offsets predicted from the corresponding objects in the last frame.

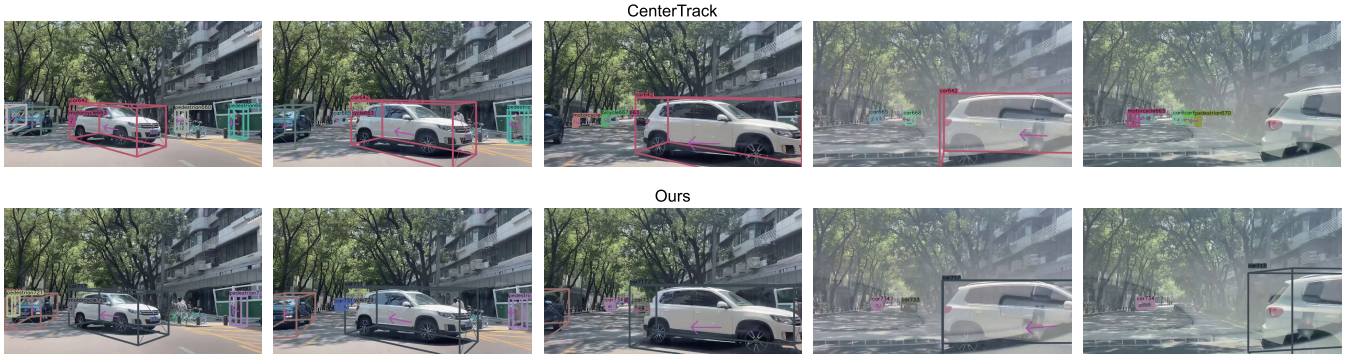

FIGURE 4. Qualitative results in a tree-lined road scene. Our method is robust to lighting variations, and outperforms the CenterTrack [3] under heavy truncation.

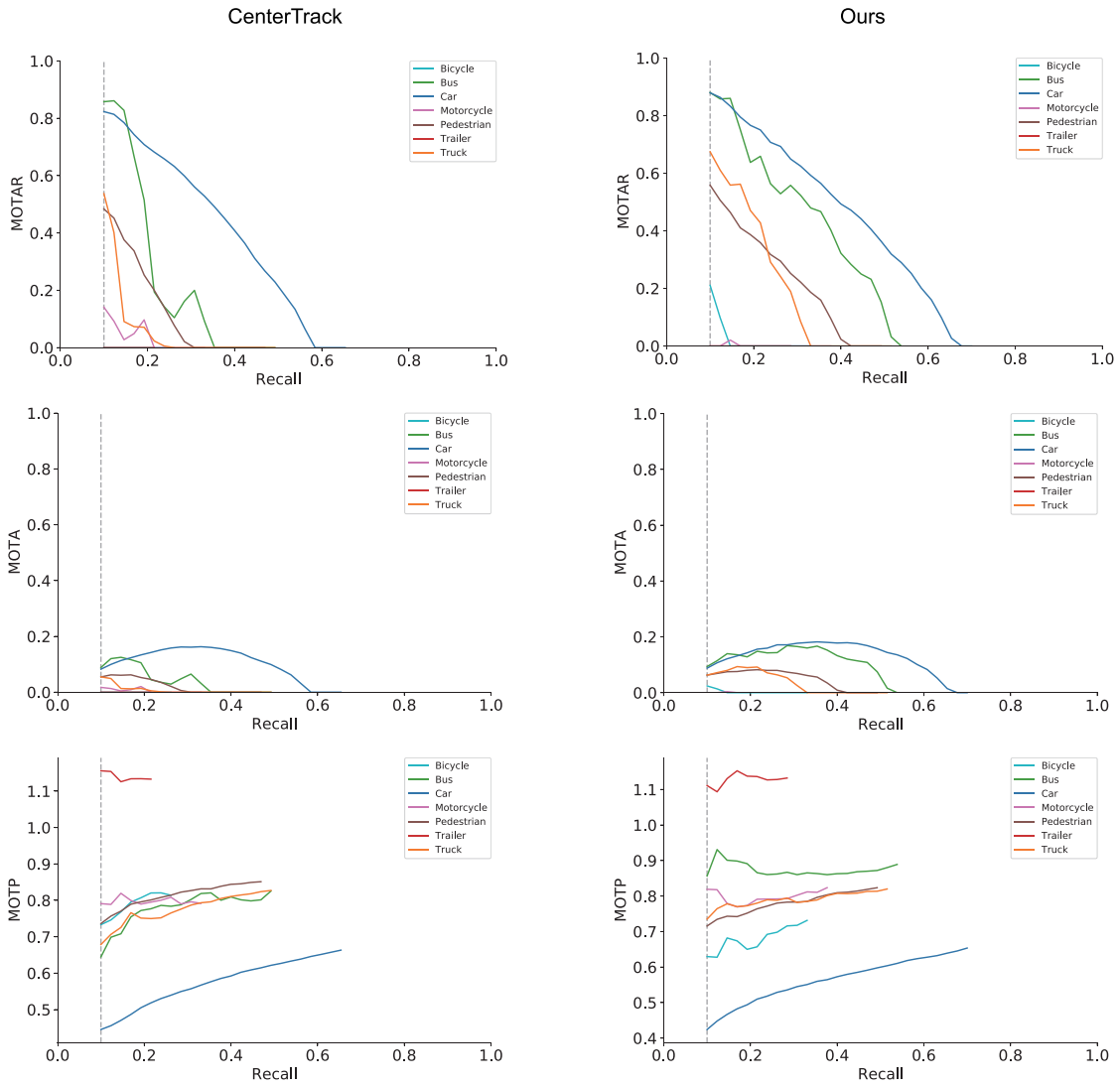

FIGURE 5. MOTAR-Recall curves, MOTA-Recall curves and MOTP-Recall curves on the nuScenes [25] dataset.

A comparison of the MOTAR-Recall curves provided by the nuScenes [25] validation set shows that our driving behavior-aware model has a distinct advantage to better avoid false positives, false negatives and ID switches related to MOTAR, especially for the car, bus and truck classes.

MOTA, multi object tracking accuracy, is the main metric considered in many autonomous driving datasets, such as KITTI [24] dataset for 2D tracking, nuScenes [25] dataset for 3D tracking, etc. We compute MOTA-Recall curves across all 7 tracking categories provided by nuScenes [25] dataset for 3D tracking. For objects that are symmetric across at least one axis, e.g., the left side of a bus looks like the right side flipped, the MOTA-Recall curves shows that the learned driving behavior exercises a strong influence on 3D symmetric objects tracking in complex traffic scenes. Our driving behavior-aware network, using ground-truth object variations in 2D center point, depth, rotation and translation in consecutive frames as supervision, significantly outperforms the solely direct 2D displacement supervised CenterTrack [3] in the challenge of 3D object tracking.

MOTP, multi object tracking precision, is another main metric adopted by nuScenes [25] benchmark for all 7 tracking classes. The MOTP-Recall curves are computed to evaluate the misalignment between the annotated and the predicted bounding boxes. Compared with the current state-of-the-art CenterTrack [3], the remarkable small position errors returned by our proposed framework are in the range suitable for 3D object tracking in complex traffic scenes, especially for the car and bicycle classes.

V. CONCLUSION

In this paper, we introduced an end-to-end deep convolutional neural network with a driving behavior-aware model for 3D object tracking. We designed our network architecture

and objective functions carefully and demonstrated that the driving behavior, formulated from object variations in 2D center point, depth, rotation and translation, served as a significant guidance for object association under heavy occlusion. Experimentally, our method outperforms state-of-the-art methods on nuScenes benchmark. We hope these results motivate future research on 3D object tracking in complex traffic scenes.

REFERENCES

[1] S. Scheidegger, J. Benjaminsson, E. Rosenberg, A. Krishnan, and K. Granström, "Mono-camera 3D multi-object tracking using deep learning detections and PMBM filtering," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 433–440.

[2] H.-N. Hu, Q.-Z. Cai, D. Wang, J. Lin, M. Sun, P. Kraehenbuehl, T. Darrell, and F. Yu, "Joint monocular 3D vehicle detection and tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5390–5399.

[3] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 474–490.

[4] S. Kayukawa, K. Higuchi, J. Guerreiro, S. Morishima, Y. Sato, K. Kitani, and C. Asakawa, "BBeep: A sonic collision avoidance system for blind travellers and nearby pedestrians," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2019, pp. 1–12.

[5] A. Manglik, X. Weng, E. Ohn-Bar, and K. M. Kitani, "Forecasting time-to-collision from monocular video: Feasibility, dataset, and challenges," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Macao, China, 2019, pp. 8081–8088.

[6] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[7] A. Dehghan, S. M. Assari, and M. Shah, "GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4091–4099.

[8] V. Chari, S. Lacoste-Julien, I. Laptev, and J. Sivic, "On pairwise costs for network flow multi-object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5537–5545.

[9] K. Yoon, D. Kim, Y.-C. Yoon, and M. Jeon, "Data association for multi-object tracking via deep neural networks," *Sensors*, vol. 19, no. 3, p. 559, Jan. 2019.

[10] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[11] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[12] A. Ess, B. Leibe, K. Schindler, and L. van Gool, "Robust multiperson tracking from a mobile platform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1831–1846, Oct. 2009.

[13] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1201–1208.

[14] S. Wang and C. C. Fowlkes, "Learning optimal parameters for multi-target tracking with contextual interactions," *Int. J. Comput. Vis.*, vol. 122, no. 3, pp. 484–501, May 2017.

[15] A. Osep, W. Mehner, M. Mathias, and B. Leibe, "Combined image- and world-space tracking in traffic scenes," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 1988–1995.

[16] S. Song and M. Chandraker, "Joint SFM and detection cues for monocular 3D localization in road scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3734–3742.

[17] L. Leal-Taixé, C. Canton-Ferrer, and K. Schindler, "Learning by tracking: Siamese CNN for robust target association," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 33–40.

[18] X. Kong, B. Xin, Y. Wang, and G. Hua, "Collaborative deep reinforcement learning for joint object search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1695–1704.

[19] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3464–3468.

[20] K. Fang, Y. Xiang, X. Li, and S. Savarese, "Recurrent autoregressive networks for online multi-object tracking," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 466–475.

[21] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3539–3548.

[22] S. Sharma, J. A. Ansari, J. K. Murthy, and K. M. Krishna, "Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 3508–3515.

[23] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649.

[24] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[25] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.

[26] M. S. Shirazi and B. T. Morris, "Looking at intersections: A survey of intersection monitoring, behavior and safety analysis of recent studies," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 4–24, Jan. 2017.

[27] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, "Deep learning-based vehicle behavior prediction for autonomous driving applications: A review," *IEEE Trans. Intell. Transp. Syst.*, early access, Aug. 4, 2020, doi: 10.1109/TITS.2020.3012034.

[28] S. Suwajanakorn, N. Snavely, J. J. Tompson, and M. Norouzi, "Discovery of latent 3D keypoints via end-to-end geometric reasoning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 2059–2070.

[29] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "DenseFusion: 6D object pose estimation by iterative dense fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3343–3352.

[30] P. Yi, Z. Wang, K. Jiang, Z. Shao, and J. Ma, "Multi-temporal ultra dense memory network for video super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2503–2516, Aug. 2020.

[31] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*. [Online]. Available: http://arxiv.org/abs/1904.07850

[32] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[34] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.

[35] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2366–2374.

[36] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3D bounding box estimation using deep learning and geometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7074–7082.

[37] X. Weng, J. Wang, D. Held, and K. Kitani, "3D multi-object tracking: A baseline and new evaluation metrics," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 10359–10366.

[38] A. Simonelli, S. R. Bulo, L. Porzi, M. López-Antequera, and P. Kontschieder, "Disentangling monocular 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1991–1999.

[39] L. Leal-Taixé, A. Milan, K. Schindler, D. Cremers, I. Reid, and S. Roth, "Tracking the trackers: An analysis of the state of the art in multiple object tracking," 2017, *arXiv:1704.02781*. [Online]. Available: http://arxiv.org/abs/1704.02781

[40] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The CLEAR 2006 evaluation," in *Proc. Int. Eval. Workshop Classification Events, Activities Relationships*. Berlin, Germany: Springer, 2006, pp. 1–44.

[41] H. Su, C. R. Qi, Y. Li, and L. J. Guibas, "Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2686–2694.

[42] L. Zhang, L. Song, B. Du, and Y. Zhang, "Nonlocal low-rank tensor completion for visual data," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 673–685, Feb. 2021.

[43] N. Wang, Y. Zhang, and L. Zhang, "Dynamic selection network for image inpainting," *IEEE Trans. Image Process.*, vol. 30, pp. 1784–1798, 2021.

QINGNAN LI received the B.S. degree from Wuhan University, Wuhan, China, in 2008, and the M.S. degree from the Wuhan University of Technology, Wuhan, in 2011, and the Ph.D. degree from Wuhan University, in 2020.

He is currently an Associate Professor with the Engineering Research Center for Transportation Systems, Wuhan University of Technology. His research interests include video coding and decoding and object detection and tracking.

RUIMIN HU (Senior Member, IEEE) received the B.S. and M.S. degrees from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 1984 and 1990, respectively, and the Ph.D. degree in communication and electronic system from the Huazhong University of Science and Technology, Wuhan, China, in 1994.

He is currently the Director of the National Engineering Research Center for Multimedia Software, Wuhan University, and the Key Laboratory of Multimedia Network Communication Engineering, Hubei. He has published two books and more than 100 scientific articles. His research interests include multimedia understanding and image/video processing.

ZHONGYUAN WANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in communication and information system from Wuhan University, Wuhan, China, in 1995, 2001, and 2008, respectively.

He is currently a Professor with the National Engineering Research Center for Multimedia Software (NERCMS), School of Computer, Wuhan University. His research interests include image/video processing and multimedia communications.

ZHI DING received the B.S. and M.S. degrees from the Wuhan University of Technology, Wuhan, China, in 2006 and 2010, respectively.

He is currently an Associate Professor with the Engineering Research Center for Transportation Systems, Wuhan University of Technology. His research interests include computer graphics and image processing.