Uncertain Object Representation for Image-Based 3D Object Perception

Qitai Wang, Yuntao Chen, and Zhaoxiang Zhang, *Senior Member, IEEE*

***Abstract*—Due to the ill-posed nature of locating 3D objects based on image inputs, objects detected by camera-based detectors tend to have considerable uncertainty in their localization. Previous works in camera-based 3D detection and tracking represent each detected object as a single certain 3D bounding box, ignoring their localization uncertainty. We propose the uncertain representation of 3D objects to meet the indeterminacy of localizing objects in images. We model the localization uncertainty of objects during the detection process and represent the location of objects as a probability distribution in 3D space. For camera-based 3D detection, we propose to gather and suppress redundant predictions about an object to form its uncertain representation. For camera-based 3D multiple object tracking, we generalize the cross-frame association metric under the uncertain representation of objects for better-tracking objects with uncertain and unstable localization. As a plug-in module for camera 3D detectors, our proposed method brings a +3.5%/+3.2%/+3.7% NDS boost to BEVDet4D/BEVDet4D-Depth/DD3D on nuScenes validation set and a +4.7% NDS boost to BEVDet4D-Depth on nuScenes test set. With enhanced cross-frame association, our tracking method achieves a 48.2% AMOTA performance and reduces the remaining identity-switch cases to only 300 on nuScenes test set.**

***Index Terms*—Camera-based 3D detection, camera -based 3D multi-object tracking (3DMOT), uncertainty, object representation.**

I. Introduction

RECENTLY, remarkable progress [1], [2], [3], [4], [5] has been achieved in camera-based 3D object detection. However, due to the ill-posed nature of estimating depth from images, the stability and accuracy of current camera-based 3D detection still lag far behind LiDAR-based methods [6], [7], [8], [9]. The inaccurate object localization is detrimental to the performance of detection as well as downstream tasks that rely on camera-based detection outputs. In previous works in camera-based 3D detection and tracking, objects are represented deterministically by 3D bounding boxes, even for objects with large localization uncertainty due to occlusion or being distant from the camera. However, as far as we are concerned, it is vital to know how certain the localization of objects is, and where the objects might locate when their localization is uncertain. In this work, instead of a single deterministic 3D bounding box, we represent the location of objects as a probability distribution in 3D space to meet the uncertainty of localizing objects in images.

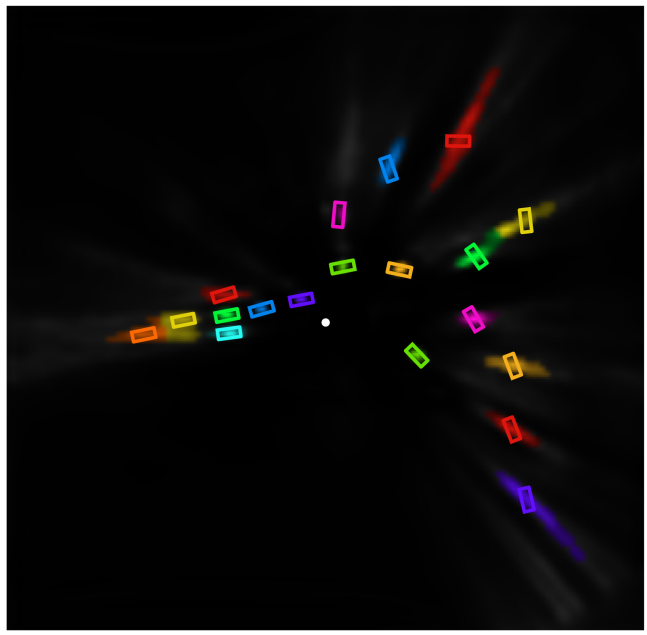


Fig. 1. *Object localization uncertainty observed under BEV view.* We draw each detected bounding box and the objectness response around them in the same color. The white dot in the center of the image indicates the ego position. Best viewed in color.

Fortunately, the uncertainty of object localization is easy to be observed in recent popular camera-based 3D detectors. Recent camera-based detectors usually elevate image features to a bird's-eye view (BEV) representation for detection. Under such a paradigm, the uncertainty of object localization can be observed in the detection objectness heat map. For example, we show the uncertainty of object localization observed in 3D detector BEVDet4D [10] in Fig. 1. We segment and show areas with high objectness responses under BEV view around detected bounding boxes, which can indicate the localization probability distribution of objects. As shown in Fig. 1, objects near the

Received 19 July 2023; revised 20 March 2025; accepted 29 April 2025. Date of publication 8 May 2025; date of current version 3 July 2025. This work was supported in part by the National Natural Science Foundation of China under Grant U21B2042 and Grant 61836014 and in part by the 2035 Innovation Program of CAS. Recommended for acceptance by V. Lempitsky. *(Corresponding author: Zhaoxiang Zhang.)*

Qitai Wang and Zhaoxiang Zhang are with the New Laboratory of Pattern Recognition (NLPR), State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: wangqitai2020@ia.ac.cn; zhaoxiang.zhang@ia.ac.cn).

Yuntao Chen is with the Hong Kong Institute of Science and Innovation, Chinese Academy of Sciences, Hong Kong SAR, China (e-mail: chenyuntao08@gmail.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2025.3568120, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2025.3568120

camera tend to have a concentrated localization probability distribution which indicates certain localization, while objects far from the camera or occluded can have high localization uncertainty, especially along the line of view. In this work, we utilize the observed object localization uncertainty for improving camera-based 3D detection and tracking.

In camera-based 3D detection, redundant predictions toward a single object along the line of view are extremely common due to the uncertain depth of objects. Those redundant predictions, most of which are false-positives, are detrimental to the performance of detectors and down-stream tasks such as multi-object tracking. Previous works neither eliminate those redundant predictions nor cluster them as a finer box representation for detected objects. In this work, we gather redundant uncertain 3D boxes covered by the uncertainty area of each object to form the uncertain box representation of objects, with its localization represented as a probability distribution in BEV view. The localization probability of an object at a BEV location is determined by the confidence score of the responding redundant box. However, we found the uncertain predictions of an object often share similar objectness scores, which leads to a nearly uniform localization probability distribution of objects. To make the localization distribution more informative, we propose a Box-Aware Filter (BAF) module to better estimate the localization confidence of uncertain 3D boxes. The BAF module utilizes a unique object-level feature sampling to better validate the rationality of uncertain boxes in perspective view. By filtering out the most likely location of object with the BAF module, we suppress the localization probability of objects in other locations to turn the localization probability distribution of objects into a near Gaussian distribution. Our proposed method significantly reduces the false positive predictions produced by camera-based 3D detectors and boosts their detection performance as a plug-in module. Furthermore, the obtained uncertain representation of 3D boxes is more informative and can greatly benefit the downstream tasks of camera-based 3D detection, such as camera-based 3D tracking.

For camera-based 3D multiple object tracking (MOT), we utilize the uncertain representation of 3D boxes to enhance the stability of the tracking process. Camera-based 3D detector often fails to provide consistent localization of the same object given poor observation conditions. To enhance cross-frame detection association, we propose to preserve and update the localization uncertainty of tracked objects. We predict the possible localization of objects in future frames based on their current uncertainty distribution and velocity, which is more robust for the cross-frame association when facing unstable localization of objects, such as before and after the object is occluded. We propose the Uncertainty-based Generalized Inter-over-Union (UGIoU) that generalizes GIoU computation between deterministic bounding boxes to between boxes under the uncertain representation, as well as its simplified version which substantially reduces the computational cost.

Our contributions in this work are as follows:

- We propose the uncertain representation of 3D bounding boxes to meet the unavoidable uncertainty in camera-based 3D detection and tracking.
- For camera-based 3D detection, we propose to gather the redundant 3D boxes of detected objects in BEV view to form the uncertain representation of objects and suppresses the false-positive redundant predictions. We propose the BAF module to locate the most possible location of objects as the center and peak of their localization probability distribution. Our proposed method boosts the performance of BEVDet4D and BEVDet4D-Depth on nuScenes validation set by 6.6% mAP and 4.9% mAP as a plug-in module. BEVDet4D-Depth equipped with our method achieves state-of-art performance on nuScenes test set. By applying the uncertain representation of objects to the famous monocular detector DD3D, we also achieve a +2.5% mAP and +1.17% mAP improvement over the monocular detector DD3D on nuScenes test split and KITTI-3D [11] test split.
- We generalize the cross-frame association metric under the uncertain representation of 3D bounding boxes. With the proposed UGIoU or its simplified version as the association metric, we reduce the remaining identity switch cases in baseline tracking results by 32% on nuScenes validation set. Our tracking algorithm outperforms all previous methods on nuScenes test set with 48.2% AMOTA and only 300 identity switches.

II. Related Works

Image-based 3D Object Detection: Image-based 3D object detection aims to predict categories and 3D bounding boxes of objects in interest from camera images. Previous works utilize various approaches to localize objects in 3D space. FCOS3D [12] generalizes the advanced 2D detector FCOS [13] to predict 3D bounding boxes with directly regressed depth. DD3D [5] introduces depth pre-training on large-scale depth datasets to enhance depth estimation. Following the model design of Centernet [7], [14], [15] directly regress bounding box with its depth while [16], [17], [18], [19] incorporate geometric clues to localize 3D objects. PGD [20] utilizes geometric relation graphs to enhance object depth estimation. [21] introduce LiDAR signals during training to help the model learn spatial cues. Recently, various works focus on predicting 3D objects from multi-camera images. Depth estimation is performed in [10], [22], [23], [24], [25], [26] to transform image features to a Bird's-Eye-View (BEV) representation to fusion information from different cameras. Another line of works [1], [2], [3], [27] continuously refines object queries located in 3D space without explicit depth estimation. All previous works represent an object in 3D with a single 3D bounding box, ignoring the localization uncertainty of objects. When facing predictions with divergence in the localization of the same object, previous works regard most uncertain predictions as redundant and utilize NMS [5], [10], [12], [19], [25], [26] or cross-predictions attention [1], [2], [3], [27] to produce the deterministic bounding boxes. However, in our opinion, the uncertain predictions convey rich localization information of objects which are hard to confidently located based on image inputs.

Image-based 3D Multi-Object Tracking: Image-based 3D MOT has also gained remarkable progress recently thanks to the developments in camera-based 3D detection. The early methods in image-based 3D MOT perform tracking in 2D with mature 2D MOT algorithms and lift the tracks to 3D space with estimated monocular depth [28]. CC-3DT [29] fuses multi-view object features to enhance associate objects across views. QD-3DT [30] perform 2D association first and enhance the instance association with 3D boxes depth-ordering heuristics and 3D motions. TripletTrack [31] extracts local object feature embeddings and motion descriptors with CNN or LSTM to measure the affinity between objects. MUTR3D [32] performs camera-based tracking in an end-to-end fashion by introducing 3D track queries to model spatial and appearance coherent tracklets. PF-Track [33] further refine the tracks and predict the trajectories of objects with cross-object and cross-frame attention between object queries. Due to the deterministic representation of bounding boxes, previous methods [28], [29], [30], [31] that perform tracking in a tracking-by-detection fashion have no access to the localization reliability of detections during the tracking process. Although with their detection and tracking process coupled, previous end-to-end trackers [32], [33] ignored the localization uncertainty of objects when representing each object with a single detection query. In this work, we extract the informative localization uncertainty of detections as a cost-free byproduct of the detection process for enhancing cross-frame association in camera-based 3D tracking.

Uncertainty estimation in camera-based detection: For intelligent agents e.g. autonomous driving systems, estimating the uncertainty in the output of detectors is crucial for safety considerations. Feng et al. [34] provides an excellent overview of uncertainty estimation for object detection in autonomous driving. Based on the source of different uncertainties, previous studies classify uncertainties into two categories, *epistemic* and *aleatoric* uncertainty [35]. *Epistemic* uncertainty models how certain the model fits the targeted mapping function with its parameters. *Aleatoric* uncertainty refers to the uncertainty raised from the noisy or insufficient observation. For the tasks of camera-based detection, previous works attempt to model the semantic (classification) uncertainty and spatial (2D bounding box regression) uncertainty of each detection result. Several works estimate the semantic [36], [37], [38], [39], [40] and spatial [37], [38], [39], [40] epistemic uncertainty with the Monte-Carlo Dropout [41] (MC-Dropout) or Deep Ensembles [42] approach. In short, MC-Dropout models the uncertainty in prediction by performing inference multiple times with random dropout applied to the model parameters. The Deep-Ensembles approach models the uncertainty by ensembling the outputs of multiple architecture-shared models that are independently initialized and trained. Both the sampling-based uncertainty estimation methods severely increase the computational cost of the inference process and are not equipped by most state-of-the-art object detectors [34].

More methods assume the probability distribution over the network outputs can be estimated by the model itself, and regard the uncertainties as output attributes of the model. Any detectors with cross-entropy loss as its classification loss and employing softmax over predicted classification logits provide the aleatoric uncertainty in object classification [34]. For spatial uncertainty, Gaussian distributions [19], [35], [43], [44] or Gaussian Mixture Models [45], [46] are often used to model the distribution of 2D box boundaries or 3D object depth. In most works involving uncertainty estimation, the spatial uncertainties of detected objects are modeled as a scalar variance upon the offsets of 2D box boundaries or object depth values. In this work, we model the spatial uncertainty of a detected object as a localization probability distribution in BEV space based on the predicted objectness heap map, which brings multiple advantages. Firstly, we can more accurately model the diverse localization uncertainties of each object. As shown in Figs. 1 and 6, the localization uncertainties of detected objects in BEV space are often irregular and asymmetric, unable to be described with a 2D Gaussian distribution, or even Gaussian mixture models. Secondly, our proposed uncertainty modeling pipeline does not rely on any additional model training for uncertainty estimation. Our proposed method can be applied to any 3D detector with dense prediction as a plug-in module.

III. Method

This section introduces the detailed definition of the proposed uncertain representation of 3D bounding boxes in Section III-A, as well as its application in 3D camera-based detection and tracking in Sections III-B and III-D. In Section III-C, we summarize the losses used in detector training in our experiments.

A. Uncertain Representation of 3D Bounding Boxes.

Conventionally, each object is represented by a single 3D bounding box in the 3D detection task. The properties of each 3D bounding box are represented with a vector $\mathbf{b}_{xy} = [x, y, z, \theta, l, w, h]$, including a certain 3D center location x, y, z, orientation θ, dimensions $l.w.h$. Each box has a confidence c_{xy}. Due to the depth uncertainty of locating 3D objects in images, previous 3D camera-based detectors often output redundant predictions with different depths for a single object as shown in Fig. 2. Most of those redundant predictions are false-positive and hard to be eliminated through BEV NMS since they are not overlapped in 3D. On the other hand, those redundant predictions also indicate the uncertainty and probability of object localization. Previous methods neither managed to suppress those redundant predictions nor gather them as a finer representation of detected objects.

To meet the localization uncertainty in detecting objects based on images, we generalize the representation of 3D bounding box as a set of uncertain 3D bounding boxes and their confidence $\{(\mathbf{b}_{xy}, c_{xy}) \mid (x, y) \in U_\alpha\}$, where U_α is the localization uncertain area of an object α and defined as a mask in BEV space. When performing camera-based 3D detection, we aggregate the redundant predictions converted by the uncertainty area of each object extracted from the objections heatmap. Specifically, during the detection process, we segment the objectness heat map predicted by the detection head of 3D detector to be the uncertainty areas of each object. In our experiments, a simple connected-component labeling applied on heat map areas where

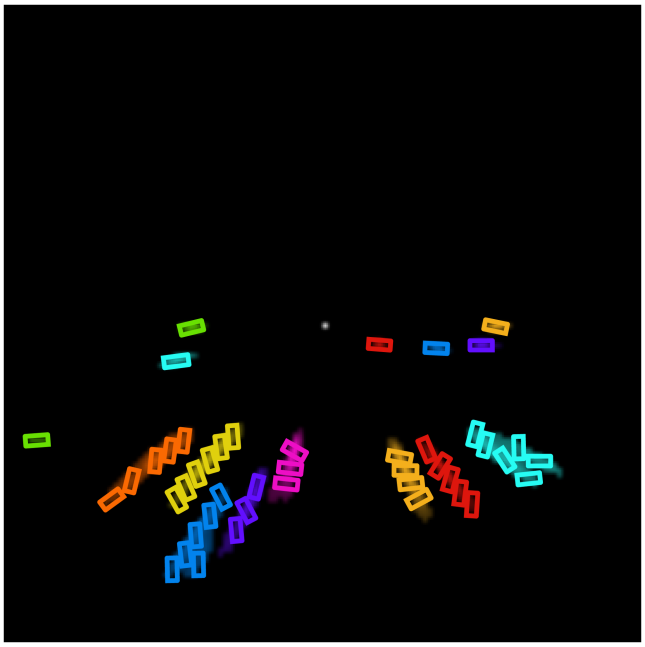


Fig. 2. *Redundant predictions observed under BEV view.* We draw redundant predictions of the same objects in the same color. The white dot in the center of the image indicates the ego position. Best viewed in color.

TABLE I
MEAN VARIANCE OF THE PROPERTIES OF REDUNDANT 3D BOXES REGARDING THE SAME VEHICLE IN CAR CATEGORY

Property	Location (longitudinal)	Location (lateral)	Dimension-L	Dimension-W
Mean	36.77 m	-	4.59 m	1.96 m
Standard Deviation	3.51 m	1.71 m	0.10 m	0.032 m

Property	Dimension-H	Yaw	Confidence	Confidence (revised)
Mean	1.72 m	-	0.31	0.33
Standard Deviation	0.055 m	0.449 rad	0.077	0.326

Results are computed based on the detection results of the BEVDet4D-R50 detector on the validation set of the nuscenes dataset.

the objectness scores are larger than T_h is enough for segmenting the uncertain areas, as shown in Fig. 1. Inside each uncertain area, we locate the predictions with maximum confidence in their neighborhood. We tell whether those predictions are referring to the same object based on their L1 object feature distance. Predictions with object feature distance larger than D_{feat} are regarded as referring to different objects. For a few objects whose uncertain areas are connected, we segment the masks by assigning grid points to their nearest object judging by distances from grid points to object centers. We record all redundant predictions inside the uncertainty areas of each object to represent the object.

In experiments, we found that redundant predictions regarding the same object tend to have very similar heights, dimensions, and orientations. As shown in Table I, the mean dimension or orientation variance of redundant predictions regarding the same object is far smaller than the variance of their locations. Therefore if we ignored the minor difference between the redundant predictions in their dimensions and orientations, we can further simplify the uncertain representation of an object α as a 3D box with fixed properties besides location, and a localization probability distribution $P_\alpha(x, y)$ in BEV space:

$$\beta = \sum_{(x,y)\in U_\alpha} c_{xy}$$

$$P_\alpha(x,y) = \begin{cases} \frac{c_{xy}}{\beta} & , \quad \text{if } (x,y) \in U_\alpha \\ 0 & , \quad \text{else} \end{cases} \tag{1}$$

B. Uncertain Representation of Boxes for 3D Detection.

Essentially, the redundant predictions regarding the same object stem from the depth uncertainty when lifting perspective features to BEV feature map. To generate the BEV feature map, BEV detectors will lift the features of 2D pixels to BEV space based on their predicted depth distribution. The same perspective feature will be lifted to different locations along the line of view in BEV space if its depth is uncertain. Therefore redundant 3D boxes regarding the same objects will be predicted in different locations in BEV space based on the similar lifted perspective features. This also answers why the redundant predictions tend to share similar 3D properties including dimensions, orientations, and even prediction confidence, as reported in Table I. We refer to box confidence predicted by the detection head **preliminary** confidence since they are simultaneously predicted with other box properties based on similar BEV features. The very similar preliminary confidence of redundant predictions regarding the same object leads to a nearly uniform localization probability distribution of objects inside their uncertainty areas, which is (removed:[counter-intuitive and]) not informative enough for indicating the true location of objects. To address this, we propose to generate **revised** confidence for each redundant prediction as a more indicative measure of the relative accuracy of redundant predictions. We propose the Box-Aware Filter (BAF) module to extract discriminative object features in perspective views based on the predicted 3D boxes and better estimate the rationality of redundant predictions. The Box-Aware Filter (BAF) module acts as a plug-in second-stage module in the 3D detection pipeline.

Estimating the revised confidence of boxes requires the model to be fully aware of the predicted properties of boxes. Previous two-stage 3D detectors represent 3D box in perspective views with 3D center point [1], [3], [47] or the minimum 2D bounding box of the projected 3D bounding box corners [48], [49] for feature sampling. However, projected 3D center points or 2D bounding boxes can not properly represent the 3D properties like the orientation of 3D boxes in perspective view. 3D boxes with various sizes, depths, or orientations can share similar center points or 2D bounding boxes in perspective view for feature sampling, leading to indistinguishable box features. Based on those similar object features, it is difficult for the model to judge the qualities of different 3D boxes. To address this, we propose to sample and align features with 3D local points in the 3D bounding box. We uniformly sample $8 \times 4 \times 4$ grid points within each 3D bounding box. 8 grid points are sampled along the heading direction of boxes. Given a 3D bounding box, we project grid points with fixed local positions inside the 3D bounding box to the image for feature sampling. Through this,

Fig. 3. Intuitive comparision between representations of 3D box in 2D. We only draw grid points on the visible surface of 3D bounding boxes for better visualization. Grid points are drawn in gradient color from the front to the back of 3D boxes.

we can sample features in perspective views with fixed positions inside 3D bounding boxes. Each sample point has a fixed location inside the 3D box, and its projected 2D location depends on the location, orientation, and dimensions of the 3D box. Under the multi-camera setting, we average the features sampled with a 3D grid point from different cameras if the projected grid point can be observed in more than one surrounding image.

As shown in Fig. 3, different 3D boxes may share the same 2D bounding boxes or projected center points in 2D. In such cases, features sampled with 2D ROI-Pooling or box center points may not be able to distinguish 3D boxes of differing qualities. Hence we propose to sample object 3D box-aware features with grid points fixed within 3D boxes. In Fig. 3 we draw sparser grid points within boxes for better visualization.

On the other hand, 2D-3D geometry constraints are proven to be useful for localizing objects by various previous works [16], [17], [19]. To leverage geometry clues in a simple, end-to-end manner, we encode both 2D and 3D geometry information as additional inputs to the box quality estimation module. Before sampling image features, we add standard 2D spatial sine positional encoding proposed by [50] to the image features. Additionally, for each 3D sample point, we attach its offset to the six boundaries of the 3D bounding boxes to its sampled features. This provides essential geometry information to the BAF module for better estimating the accuracy of 3D bounding boxes.

The features sampled for each box have a shape of $[8 \times 4 \times 4, 256 + 6]$. We apply a $2 \times 2 \times 2$ 3D max pooling with stride=2 to downsample the object feature as of shape $[4 \times 2 \times 2, 256 + 6]$. Then we flatten the object feature into vectors of shape $[16 \times 262]$ and then pass them through an MLP module with three fully connected layers. The output feature of the MLP module has a shape of [1024]. Finally, a linear layer is utilized to predict the GIoU_{3D} value between the 3D box and its assigned ground truth.

In the detection process, we use the BAF module to estimate the more discriminative revised confidence of each redundant prediction. To further increase the peakedness of the distribution of revised confidence of redundant predictions regarding the same object, we suppress the confidence score of redundant predictions with a process similar to Soft-NMS [51]. For each object, in each iteration, we include the uncertain prediction $\mathbf{b}_m$ with the highest revised confidence in the redundant predictions regarding an object. Here we represent 3D box $\mathbf{b}_{x_m y_m}$ as $\mathbf{b}_m$ for simplicity. We then suppress the confidence of remaining redundant predictions based on their distance from $\mathbf{b}_m$. To limit the redundant suppression in the depth direction, we only suppress redundant predictions with lateral distance within a threshold T_l. We define the lateral distance between $\mathbf{b}_m$ and a redundant box $\mathbf{b}_j$ to be the distance between them after moving $\mathbf{b}_j$ along the line of view until it has the same depth as $\mathbf{b}_m$. We multiply a suppressing factor decided by the Euclidean distance between $\mathbf{b}_m$ and $\mathbf{b}_j$ to the revised confidence c'_j of $\mathbf{b}_j$:

$$c'_j = c'_j * e^{-\gamma * dist(\mathbf{b}_m, \mathbf{b}_j)} \tag{2}$$

Where γ is the weighting factor for suppression. We illustrate the full uncertainty extraction and redundant prediction suppression pipeline in Algorithm 1.

The suppression process for redundant prediction sets of different objects proceeds in parallel.

The proposed BAF module significantly boosts the performance of our baseline model BEVDet4D [10] by providing more discriminative confidence for the redundant predictions. But could the high localization uncertainty of objects detected by BEVDet4D result from the lack of direct depth supervision during model training, instead of the uncertain nature of locating 3D objects based on images? To validate the benefits of our method on camera-based detectors with explicit depth supervision, we also train the model with lidar point supervision as described in [25] as BEVDet4D-Depth.

Moreover, we also apply the uncertain representation of objects and the proposed BAF module on the famous monocular 3D detector DD3D [5], which do not perform 3D detection under the BEV representation. For DD3D, we regard a group of boxes suppressed by the same bounding box during 2D/3D NMS as the set of uncertain boxes of an object. We apply the BAF module as a plug-in module for assessing the revised confidence of uncertain boxes and performing redundant prediction suppression. We do not perform cross-image feature sampling to keep the monocular-style pipeline of DD3D.

C. Losses

For experiments with BEVDet4D and BEVDet4D-Depth as the baseline detector, we adopt the same heat map loss, depth loss (for BEVDet4D-Depth), and 3D bounding box regression losses along with their respective loss weights as described in [10] and [25]. When training, we decode the regressed uncertain 3D boxes and calculate the GIoU_{3D} between predicted boxes and their assigned ground truth boxes as confidence targets, which are then used to supervise the predicted confidence. We use a smooth L1 loss $\mathcal{L}_{conf3D}$ to sparsely supervise the BAF module.

Algorithm 1: Full Uncertainty Extraction Pipeline.

Input: Multi-view images $\mathcal{I} \in \mathbb{R}^{6\times3\times H\times W}$.

Intermediate Variables: Set of 3D boxes $\mathfrak{B} = \{\mathbf{b}_i\}$,where $\mathbf{b}_i = [x_i, y_i, z_i, \theta_i, l_i, w_i, h_i] \in \mathbb{R}^7$; confidence of boxes $\mathfrak{C} = \{c_i\}, c_i \in \mathbb{R}$; objectness heat map $H_{\text{heat}} \in \mathbb{R}^{1\times128\times128}$; image features $\text{F}_{\text{2d}} \in \mathbb{R}^{6\times256\times\frac{H}{16}\times\frac{W}{16}}$, uncertain area $U_k \in \mathbb{R}^{128\times128}$ of an object k.

Output: Output boxes $\mathfrak{B}'$ and confidence scores $\mathfrak{C}'$ as detection results. Extracted uncertain area masks $\mathcal{U} = \{U_1, .., U_N\}$ and object localization distributions $\mathcal{P} = \{P_1(x, y), .., P_N(x, y)\}$ where $P_k(x, y) \in \mathbb{R}^{1\times128\times128}$ is the normalized localization probability distribution of a object k.

Modules/Functions: $f_{\text{backbone}} : \mathcal{I} \rightarrow \text{F}_{\text{2d}}$ represents the image backbone of the detector. $f_{\text{head}} : \text{F}_{\text{2d}} \rightarrow \{\mathfrak{B}, \mathfrak{C}, H_{\text{heat}}\}$ represents the 2D feature lifting module and the detection head in BEV space. $f_{\text{BAF}} : \{\text{F}_{\text{2d}}, \mathbf{b}_i\} \rightarrow c'_i$ represents the proposed BAF module with grid point projection, feature sampling and confidence estimation.

Input($\mathcal{I}$)
$\text{F}_{\text{2d}} = f_{\text{backbone}}(\mathcal{I})$
$\mathfrak{B}, \mathfrak{C}, H_{\text{heat}} = f_{\text{head}}(\text{F}_{\text{2d}})$
$\mathcal{U} = \{U_1, .., U_N\} = \text{Segment}(H_{\text{heat}}, \mathfrak{B})$
Have$\{\mathcal{B}_1, .., \mathcal{B}_N\}, \mathcal{B}_k = \{\mathbf{b}_i | (x_i, y_i) \in U_k\}$
for k in $(1, N)$ **do**
  $\mathcal{C}'_k = \{c'_i | \mathbf{b}_i \in \mathcal{B}_k\} = \{f_{\text{BAF}}(\text{F}_{\text{2d}}, \mathbf{b}_i) | \mathbf{b}_i \in \mathcal{B}_k\}$
  $\mathcal{B}'_k = \{\}$
  while $\mathcal{B}_k \neq \{\}$ **do** ▷ Redundant Prediction Suppression
    $m = \arg\max_i \{c'_i \mid c'_i \in \mathcal{C}'_k\}$
    $\mathcal{B}'_k \leftarrow \mathcal{B}'_k \cup \{\mathbf{b}_m\}$
    $\mathcal{B}_k \leftarrow \mathcal{B}_k - \{\mathbf{b}_m\}$
    $\mathcal{C}'_k \leftarrow \mathcal{C}'_k - \{c'_m\}$
    for $\mathbf{b}_j$ in $\mathcal{B}_k$ **do** ▷ Suppress the relevant predictions
      if $lateral\_dist(\mathbf{b}_m, \mathbf{b}_j) \leq T_l$ **then**
        $c'_j \leftarrow c'_j * e^{-\gamma * dist(\mathbf{b}_m, \mathbf{b}_j)}$
      end if
    end for
  end while
  $\beta_k = \sum_{\mathbf{b}_i \in \mathcal{B}'_k} c'_i$ ▷ Normalize
  Have $P_k(x, y)$, $P_k(x_i, y_i) = \begin{cases} \frac{c'_i}{\beta_k} & , \text{ if } (x_i, y_i) \in U_k \\ 0 & , \text{ else} \end{cases}$
end for
$\mathfrak{B}' = \mathcal{B}'_1 \cup \mathcal{B}'_2 \cup .. \cup \mathcal{B}'_N$
$\mathfrak{C}' = \mathcal{C}'_1 \cup \mathcal{C}'_2 \cup .. \cup \mathcal{C}'_N$
$\mathcal{P} = \{P_1(x, y), .., P_N(x, y)\}$
Output($\mathfrak{B}', \mathfrak{C}', \mathcal{U}, \mathcal{P}$)

We set loss weight $\lambda_c = 1.0$ for the revised box confidence loss. We train the BAF module end-to-end with the detectors.

For experiments with DD3D, we make no change to the classification, 2D bounding box regression, and 3D bounding box regression losses as described in [5] in spite of the depth regression loss. We predict depth uncertainty for bounding boxes and employ the depth loss formulated in [44]. We maintain the same weight for the depth loss as in [5]. We utilized the same confidence losses described above to supervise the attached BAF module.

D. Enhancing Cross-Frame Association in Tracking

Camera-based 3D Multi-Object Tracking (3DMOT) suffers from unstable and inaccurate object localization. The instability of camera-based 3D detection across frames mainly results from the uncertainty of object localization. Even minor changes to the observation of objects such as the motion of the ego vehicle can lead to discrete detection bounding boxes across frames, not to mention when facing occlusion.

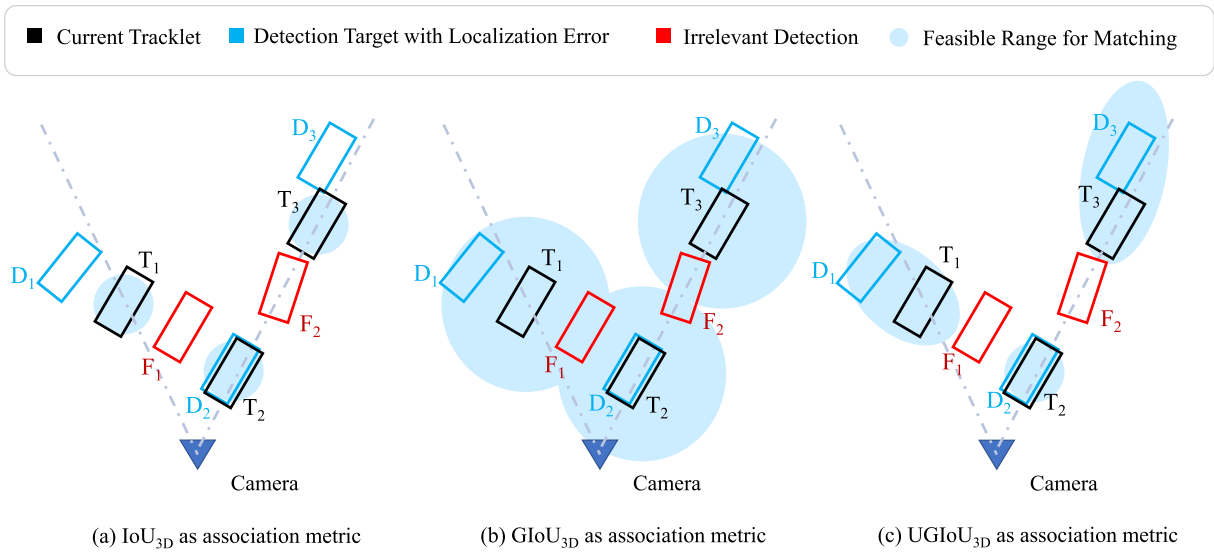
Conventional IoU-based bounding box association is not feasible if there is a large localization error between the current object detection and its past trajectory, which is common under image-based detection. To address this issue, [52] utilizes GIoU_{3D} as the matching cost, which enables association between boxes not overlapped. However, GIoU_{3D} does not distinguish between localization errors in different directions and does not distinguish boxes with high localization uncertainty or confident boxes, which both may lead to unnecessary large localization error tolerance and wrong association. To address this issue, we propose to utilize the localization uncertainty of boxes obtained in the detection process for enhancing cross-frame association in tracking.

Equipped with the uncertain representation of objects, we can predict the future localization probability of objects combined with the inferred object velocities. As discussed in Sections III-A and III-B, we can gather the uncertain 3D boxes regarding the same object as the uncertain box sets and further represent the localization of objects as a probability distribution in BEV view. In the tracking process, we preserve the latest observed object localization probability distribution $P(\mathbf{b}_{xy})$ of each object and predict the future possible location of objects by moving their localization probability distributions along their predicted future trajectory. To measure the similarity of 3D boxes under the uncertain representation, we propose the Uncertainty-based Generalized Inter-over-Union(UGIoU) that generalizes Generalized Inter-over-Union(GIoU) under the uncertain representation of boxes. Specifically, for uncertain box distribution of tracklet $P_T(x, y)$ and detection $P_D(u, v)$, we define UGIoU_{3D} as:

$$\text{UGIoU}_{\text{3D}}(T, D) = \sum_{xy}\sum_{uv} P_T(x, y)P_D(u, v) * \text{GIoU}_{\text{3D}}(\mathbf{b}_{xy}, \hat{\mathbf{b}}_{uv}) \tag{3}$$

where $\mathbf{b}_{xy}, \hat{\mathbf{b}}_{uv}$ are uncertain boxes of two objects in each possible location. When the uncertain box localization distributions degenerate to be certain one-hot distributions, the defined UGIoU_{3D} equals the GIoU_{3D} between two certain boxes.

Fig. 4 gives an intuitive comparison between IoU_{3D}, GIoU_{3D}, and UGIoU_{3D}. In Fig. 4(a), when utilizing IoU_{3D} as association metric, tracklet and detection pairs with large localization error ($\text{T}_1\&\text{D}_1, \text{T}_3\&\text{D}_3$) can not be associated. In Fig. 4(b), by utilizing GIoU_{3D} as association metric, box pairs with localization error can be associated under low association threshold. However GIoU_{3D} do not classify objects with different localization uncertainty, leading to possible wrong associations ($\text{T}_1\&\text{F}_1, \text{T}_2\&\text{F}_1, \text{T}_3\&\text{F}_2$). With the proposed UGIoU_{3D} as association metric which takes the localization distribution of objects



(a) IoU_{3D} as association metric (b) $GIoU_{3D}$ as association metric (c) $UGIoU_{3D}$ as association metric

Fig. 4. *Comparison between different association metrics for 3D camera-based multi-object tracking.* 3D bounding boxes are drawn in BEV. We draw the current detection of objects in black and their relative tracklet in blue. Irrelevant tracklets are drawn in red. The shaded areas approximately indicate the feasible range for tracklets to associate with detection.

into consideration, we can flexibly adjust the feasible association range for each bounding box.

Considering the computational cost of $UGIoU_{3D}$ is N^2 times higher than conventional $GIoU_{3D}$, which is too heavy for achieving real-time tracking, we also propose to directly measure the similarity between the localization probability distributions of boxes with Kullback-Leibler(KL) divergence:

$$D_{KL}(T, D) = \sum_{xy} P_T(x, y) * (log(P_T(x, y)) - log(P_D(x, y))) \quad (4)$$

Here we measure the localization distribution similarity of the detection and the predicted status of tracklet. We further perform a two-stage bipartite matching in the tracking process to save computational costs. We first calculate the $GIoU_{3D}$ between the most confident detection box and tracking box for matching the bounding box pairs with small localization errors. Then for boxes with large localization errors, we compute $UGIoU_{3D}$ or KL divergence with their localization probability distributions to perform a second-stage bipartite matching.

IV. Experiments

In this section, we first introduce our detailed experimental setup. Then we provide comparisons with recent works on nuScenes camera-based 3D detection benchmark, nuScenes camera-based 3D MOT benchmark, and KITTI-3D camera-based 3D detection benchmark. Ablation studies on each component in our method are conducted on nuScenes validation set.

A. Datasets and Metrics.

We evaluate our proposed method on nuScenes and KITTI datasets.

nuScenes [53] dataset is a large-scale autonomous driving benchmark containing 1000 multi-modal videos in total with six cameras. The videos are divided into 700, 150, and 100 scenes for training, validation, and testing respectively. 3D box annotations of 10 object classes are provided at 2 FPS for the videos. For nuScenes 3D detection benchmark, we report mean Average Precision (mAP), nuscenes Detection Score (NDS) as well as the true positive metrics including Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE) and mean Average Attribute Error (mAAE). For the 3D MOT task we report AMOTA, AMOTP [54], MOTA [55] and identity switches (IDS).

On **KITTI-3D** [11] Object Detection benchmark, the accuracy of 3D detection is measured with average precision defined in 3D space (3D AP) or Bird-Eye-View (BEV AP). We report our performance in the Car category on the KITTI test set.

B. Implementation Details

Experiments on top of multi-camera 3D detectors: We follow most of the settings of the official implementation[1] of BEVDet4D [10] with modifications including enlarging the positive assign range of ground truth objects in BEV view. We use ResNet-50 or SwinTransforme-Base [56] as the bachbone of BEVDet4D, named BEVDet4D-R50 and BEVDet4D-Base. We also train BEVDet4D with lidar points supervision as BEVDet4D-Depth.

Models are trained with AdamW [57] optimizer, learning rate 2e-4, and a total batch size of 24. Input images are clipped to 704 × 256 / 1600 × 640 for model with ResNet-50 / Swin-B as its backbone following [10]. We do not train our model

[1]https://github.com/HuangJunJie2017/BEVDet

TABLE II
COMPARISONS WITH RECENT WORKS ON NUSCENES TEST SET

Method	Temporal	BackBone	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
DD3D§ [5]	×	V2-99*	0.418	0.477	0.572	0.249	0.368	1.014	0.124
DETR3D§# [62]	×	V2-99*	0.412	0.479	0.641	0.255	0.394	0.845	0.133
BEVDet [26]	×	Swin-B	0.424	0.488	0.524	0.242	0.373	0.950	0.148
BEVFormer-S [1]	×	V2-99*	0.435	0.495	0.589	0.254	0.402	0.842	0.131
PETR [2]	×	V2-99*	0.441	0.504	0.593	0.249	0.383	0.808	0.132
BEVDepth§# [25]	✓	V2-99*	0.503	0.600	0.445	0.245	0.378	0.320	0.126
BEVFormer [1]	✓	V2-99*	0.481	0.569	0.582	0.256	0.375	0.378	0.126
BEVDet4D-Base§# [10]	✓	Swin-B	0.451	0.569	0.511	0.241	0.386	0.301	0.121
UVTR [63]	✓	V2-99*	0.472	0.551	0.577	0.253	0.391	0.508	0.123
PolarFormer [27]	✓	V2-99*	0.490	0.572	0.556	0.256	0.364	0.440	0.127
FrustumFormer [64]	✓	V2-99*	0.516	0.589	0.555	0.249	0.372	0.389	0.126
BEVFormer v2 [65]	✓	InternImage-B	0.540	0.620	0.488	0.251	0.335	0.302	0.122
DD3D+ours§	×	V2-99*	0.443	0.485	0.544	0.258	0.443	1.123	0.117
BEVDet4D-Depth-Base+ours†	✓	Swin-B	0.545	0.616	0.437	0.264	0.429	0.283	0.150

†: with lidar supervision. §: with test-time augmentation. #: trained with CBGS. *: using V2-99 backbone that was pre-trained with external data.

with CBGS [58] nor perform test-time augmentation as [10]. On nuScenes validation set, our models are trained for 20 epochs. On nuScenes test set, we slightly extend the training schedule to 24 epochs.

We set the objectness threshold $T_h = 0.1$ and $D_{\text{feat}} = 2.5$ for extracting uncertain areas. We set the suppressing factor $\mu = 0.25$ and lateral distance threshold for soft suppression $T_t = 1.0$ m.

Experiments on top of monocular 3D detector DD3D: We use DLA-34 [59]/V2-99 [60] as the backbone network for the experiments on nuScenes and KITTI. Our model is trained end-to-end using an SGD optimizer with a learning rate of 2.5e-5 and a batch size of 32. For experiments on nuScenes validation set, our model is trained for 120 k iterations. We use CBGS [58] and train the model for 240 k iterations for nuScenes test set. As for KITTI, the model is trained for 25 k iterations with a batch size of 48. All the other hyper-parameters are the same as in [5]. For all experiments, we use random seeds and do not report cherry-picked results.

3D MOT on nuScenes dataset: We adopt ImmortalTracker [61] as a simple, training-free baseline. We replace the association metric used in ImmortalTracker with our proposed UGIoU_{3D} or D_{KL}. We take detection results predicted by BEVDet4D+BAF as inputs to the tracking process.

C. Main Results

Detection results on nuScenes: Table II shows comparisons of detection results between DD3D/BEVDet4D/BEVDet4D-Depth with our proposed methods applied and other state-of-the-art methods on the nuScenes test set. We report the performance of detectors applied with our proposed methods including gathering uncertain boxes as the uncertain representation of objects, locating the center and the peak of object localization distribution with BAF module, and suppressing redundant predictions under the uncertain representation of objects. Our proposed methods significantly boost the performance of BEVDet4D-Depth-Base and achieve leading performance without CBGS or test-time augmentation applied. Our model outperforms all pervious methods besides BEVFormer v2, which utilizes a much stronger InternImage-B model as its image backbone. Under the monocular setting, compared to DD3D, our model also achieves a +2.5% mAP increase and +0.8% NDS increase. Our method achieves the best mAP performance among all state-of-art methods without temporal aggregation.

We report our experiment results on nuScenes validation set in Table III.

Detection results on KITTI-3D: Table IV presents the performance of DD3D applied with our proposed methods on KITTI-3D test set. Our method obtains a 2.08/0.82/1.03 BEV AP improvement and a 0.30/0.23/0.44 3D AP improvement under easy/moderate/hard setting over DD3D reported performance. We cannot reproduce the performance of DD3D reported on KITTI-3D test set. We trained the baseline DD3D model without any modification to the open-sourced official implementation and report its performance on KITTI-3D test set as DD3D†. Compared to the performance of DD3D†, the model coupled with BAF achieves a 3.16/1.94/1.90 BEV AP increase and a 1.77/1.17/1.34 3D AP increase.

Tracking results on nuScnens: Table V reports the performance of our tracking method on the nuScenes test set. Our method outperforms all previous methods on AMOTA, AMOTP, and MOTA metrics with a large gap on nuScenes test set. We reduce the identity switch cases in our tracking results to only 300, significantly fewer than most of the previous methods. PF-Track-F shares similar IDS performance with our method, but it relies on a customized and 3D detector to perform detection with tracklet queries for enhancing cross-frame association. The 3D detector needs to be trained with customized label assigning and tracking losses, which are not proven to be beneficial for detection performance, if not detrimental. This leads to the limited applicability of PF-Track to be implemented upon modern, scaled-up 3D detectors. On the other hand, our proposed uncertain representation of objects serves as a plug-in module

TABLE III
COMPARISONS WITH RECENT WORKS ON NUSCENES VALIDATION SET

Method	BackBone	Size	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
DETR3D# [62]	Res-101	1600×900	0.349	0.434	0.716	0.268	0.379	0.842	0.200
PETR# [2]	Res-101	1600×900	0.370	0.442	0.711	0.267	**0.383**	0.865	0.201
BEVDet§# [26]	Swin-B	1600×640	0.397	0.477	0.595	0.257	0.355	0.818	0.188
BEVFormer [1]	Res-101	1600×900	0.416	0.517	0.673	0.274	0.372	0.395	0.198
BEVDet4D-Tiny# [10]	Swin-T	704×256	0.338	0.476	0.672	0.274	0.460	0.337	0.185
BEVDet4D-Base#	Swin-B	1600×640	0.421	0.545	0.579	0.258	0.329	0.301	0.191
BEVDepth†# [25]	Res-101	512×1408	0.412	0.535	0.565	0.266	0.358	0.331	0.190
PolarFormer [27]	Res-101	1600×900	0.432	0.528	0.648	0.270	0.348	0.409	0.201
FrustumFormer [64]	Res-101	1600×900	0.457	0.546	0.624	0.265	0.362	0.380	0.191
DD3D-DLA34+ours§	DLA-34	1600×900	0.394	0.439	0.660	0.272	0.458	1.080	0.190
BEVDet4D-R50+ours	Res-50	704×256	0.390	0.482	0.658	0.284	0.648	0.333	0.206
BEVDet4D-Depth-R50+ours†	Res-50	704×256	0.399	0.489	0.635	0.283	0.639	0.337	0.206
BEVDet4D-Depth-Base+ours†	Swin-B	1600×640	0.489	0.570	0.542	0.271	0.441	0.274	0.215

†: with lidar supervision.§: with test-time augmentation. #: trained with CBGS.

TABLE IV
STATE-OF-THE-ART COMPARISONS FOR 3D DETECTION FOR CAR CATEGORY ON KITTI-3D TEST SET

Methods	BEV AP			3D AP		
	Easy	Mod	Hard	Easy	Mod	Hard
M3D-RPN [66]	21.02	13.67	10.23	14.76	9.71	7.42
D4LCN [67]	22.51	16.02	12.55	16.65	11.72	9.51
DCD [17]	32.55	21.50	18.25	23.81	15.90	13.21
DD3D[5]	30.98	22.56	20.03	23.22	16.34	14.20
monoDDE[44]	33.58	23.46	20.37	24.93	17.14	15.10
DD3D†	29.90	21.44	19.19	21.75	15.40	13.38
DD3D†+ours	33.06	23.38	21.09	23.52	16.57	14.64

DD3D†: we report the performance of our baseline DD3D model based on the open-sourced official implementation of DD3D without any modification.

TABLE V
STATE-OF-THE-ART COMPARISONS FOR CAMERA-BASED 3D MOT ON NUSCENES TEST SET

Method	AMOTA↑	AMOTP↓	MOTA↑	IDS↓
Test Split				
QD3DT [30]	0.217	1.550	0.198	6856
TripletTrack [31]	0.268	1.504	0.245	1044
MUTR3D [32]	0.270	1.494	0.245	6018
PolarDETR [31]	0.273	1.185	0.238	2170
SRCN3D[49]	0.398	1.317	0.359	4090
PF-Track-F[33]	0.434	1.252	0.378	**249**
Ours	**0.482**	**1.065**	**0.407**	300

for any camera-based 3D detector, and our tracking algorithm follows a tracking-by-detection fashion. This makes our method easy to be built upon and benefits from strong 3D detectors, hence achieving a much higher AMOTA performance.

D. Ablation Studies

We conduct all the ablation studies on nuScenes validation split.

Impact of the uncertain representation of objects applied to multi-camera and monocular detectors: Here we show the detailed impact of our methods applied to multi-camera detectors BEVDet4D and BEVDet4D-Depth, as well as the monocular detector DD3D. As shown in Table VI, our proposed methods brings +6.6% mAP boost and +3.5% NDS boost to BEVDet4D on nuScenes validation set. We also boost the performance of BEVDet4D-Depth by +4.9% mAP and +3.2% NDS. For monocular detector DD3D, BAF module can also bring a significant +4.8%mAP and +3.7%NDS boost.

Suppressing redundant predictions: As introduced in Section III-B, we suppress redundant predictions for each object with a customized soft suppression strategy. This design avoids the elimination of redundant predictions (hard suppression) to increase the maximum recall rate of objects. As shown in Table VII, performing a hard suppression on the detection results leads to a severe decrease in the mAP performance due to the decreased recall rate.

Effect of the 3D box-aware sampling: Table VIII demonstrates the impact of the 3D box-aware sampling in the BAF module to each detector. We replace the 3D box-aware sampling with a 2D ROI-pooling on image features based on the 2D bounding box of projected 3D boxes or a point feature sampling in BEV feature map. Experiment results in Table VIII demonstrate that 3D box-aware feature sampling can lead to more precise 3D box quality estimation. Scene-level feature extraction with 3D-to-2D projection is widely used for BEV view transformation under the multi-camera setting. However, as far as we know, we are the first to utilize 3D structural information for instance-level feature extraction in image-based 3D detection. Beyond the monocular setting, the proposed novel BAF module is also beneficial for multi-camera detectors, with or without temporal aggregation. Considering 3D-to-2D projection is already utilized for scene-level feature extraction in BEVDet4D,

TABLE VI
ABLATION STUDY OF BAF APPLIED TO MONOCULAR DETECTOR DD3D, MULTI-CAMERA DETECTOR BEVDEPTH AND MULTI-CAMERA DETECTOR WITH TEMPORAL AGGREGATION BEVDET4D

Method	+ours	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
DD3D-DLA34	×	0.346	0.402	0.699	0.281	0.542	1.121	0.185
	✓	0.394	0.439	0.660	0.272	0.458	1.080	0.190
BEVDet4D-R50	×	0.324	0.447	0.700	0.283	0.601	0.361	0.212
	✓	0.390	0.482	0.658	0.284	0.648	0.333	0.206
BEVDet4D-R50-Depth	×	0.350	0.457	0.695	0.287	0.605	0.373	0.223
	✓	0.399	0.489	0.635	0.283	0.639	0.337	0.206

TABLE VII
ABLATION ON THE SUPPRESSION STRATEGY TOWARDS REDUNDANT PREDICTIONS

Suppression	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
-	0.381	0.477	0.669	0.283	0.650	0.325	0.205
Hard	0.309	0.445	0.636	0.284	0.639	0.328	0.208
Soft	0.390	0.482	0.658	0.284	0.648	0.333	0.206

TABLE VIII
ABLATION STUDY ON THE BAF MODULE

Detector	feature sampling	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
DD3D	BEV	0.371	0.421	0.694	0.296	0.481	1.055	0.172
	2D ROIAlign	0.378	0.426	0.682	0.283	0.477	1.041	0.184
	3D Box-Aware	0.394	0.439	0.660	0.272	0.458	1.080	0.190
BEVDet4D	BEV	0.361	0.464	0.663	0.279	0.650	0.308	0.203
	2D ROIAlign	0.370	0.469	0.687	0.292	0.664	0.326	0.194
	3D Box-Aware	0.390	0.482	0.658	0.284	0.648	0.333	0.206
BEVDet4D-Depth	BEV	0.378	0.481	0.647	0.275	0.644	0.304	0.206
	2D ROIAlign	0.387	0.482	0.647	0.283	0.657	0.314	0.204
	3D Box-Aware	0.399	0.489	0.635	0.283	0.639	0.337	0.206

We report the performance of detectors without the BAF module, with the BAF module but replace the proposed 3D bounding box-aware feature sampling with 2D ROIAlign, and with the proposed BAF module.

we can conclude that the benefits we can gain from scene-level or instance-level 3D structure-aware feature sampling are orthogonal. This point was not well-recognized in previous arts.

Ablation on the sample grid: In the proposed BAV module, 3D grid points are utilized for feature sampling. In Table IX we show the ablation experiments on the distribution of 3D grid points. When we sample grid points inside 3D bounding boxes with fixed grid size as described in Section III-B, a grid size as large as $4 \times 4 \times 4$ is enough to provide enough sample points. We also noticed that assigning more grid points along the heading direction of boxes (X-axis) performs better than a Uniformly distributed grid. A grid size of $8 \times 4 \times 4$ with a total of 128 points performs better than a grid size of $6 \times 6 \times 6$ with a total of 216 points. We think this is because most objects in driving scenes have a longer length than their width or height, e.g. cars or buses. So that more sample points along their heading direction lead to more uniformly distributed 3D sample points.

TABLE IX
ABLATION ON GRID SIZE

Grid Type	Grid size	mAP↑	NDS↑
Fixed Grid size	2×2×2	0.377	0.468
	4×4×4	0.388	0.478
	8×4×4	0.390	0.482
	6×6×6	0.388	0.479
	8×8×8	0.391	0.482
Grid Type	Ceil size/m	mAP↑	NDS↑
Fixed Ceil size	1.0×1.0×1.0	0.375	0.469
	0.5×0.5×0.5	0.378	0.474

Sampling grid points under fixed ceil size instead of fixed point num performs worse than the latter. We speculate that this is because a fixed ceil size cannot ensure sampling features in

TABLE X
ABLATION ON THE FORWARD MODULE IN BAF

Module	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
MLP	0.390	0.482	0.658	0.284	0.648	0.333	0.206
3D conv	0.391	0.481	0.645	0.285	0.674	0.341	0.205
Self-attn	0.386	0.479	0.650	0.281	0.652	0.345	0.209

TABLE XI
ABLATION ON THE HEATMAP DEFINITION FOR MODELING DETECTION UNCERTAINTY

Uncertainty Source	Detection Confidence	mAP↑	NDS↑	AMOTA↑	IDS↓
Objectness Map	Preliminary	0.361	0.464	0.458	310
Revised Confidence Map	Revised	0.381	0.478	0.469	337
Objectness Map	Revised	0.381	0.478	0.469	299

TABLE XII
ABLATION ON THE WEIGHT OF REVISED BOX CONFIDENCE LOSS

λ_c	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
0.25	0.387	0.476	0.655	0.288	0.680	0.344	0.202
0.5	0.390	0.482	0.659	0.288	0.662	0.333	0.206
1.0	0.390	0.482	0.658	0.284	0.648	0.333	0.206
2.0	0.396	0.473	0.676	0.286	0.657	0.349	0.212

the corners of 3D boxes and hence fails to convey the precise boundary of 3D boxes.

Forward module in BAF: In Table X we replace the forward MLP in BAF with 3D convolutions or attention layers. These three structures perform nearly equivalent as the forward module in BAF.

Heatmap for modeling detection uncertainty: In Section II-I-A, we extract the uncertainty of detections from the objectness heat map. In Section III-B, we estimate the revised confidence of uncertain boxes with the proposed BAF module and hence can generate a revised confidence heat map around each final prediction. In Table XI, we tested extracting uncertainty from the revised confidence heat map instead of the objectness map. We found although the revised confidence of uncertain bounding boxes can locate the more precise boxes and increase the detection performance as well as the AMOTA metric for 3D MOT, uncertainty extracted from the revised confidence map performs worse than uncertainty extracted from objectness map. We speculate that this is because the BAF module suppressed the revised confidence of inaccurate uncertain boxes, which in turn makes the distribution of revised confidence less indicative for predicting the possible location of boxes in future frames.

Loss weight for the revised box confidence loss: In Table XII, the loss weight λ_c of the revised box confidence loss is optimal when set to 0.5 or 1.0. A higher λ_c will force the model to pay imbalanced attention to the precision (mAP) of 3D boxes and harm the model's NDS performance.

Uncertainty-guided box association metric for camera-based 3D MOT: Table XIII shows the ablation experiment on association metrics for tracking. Replacing IoU_{3D} with GIoU_{3D} as the association metric allows boxes without overlap to be associated. This strengthens tracking objects with unstable detections across frames and reduces identity switches.

TABLE XIII
ABLATION ON ASSOCIATION METRIC IN 3D MOT

Metric	AMOTA↑	MOTA↑	IDS↓	FPS↑
IoU_{3D}	0.439	0.386	1329	30.6
GIoU_{3D}	0.460	0.414	431	15.4
UGIoU_{3D}	0.471	0.417	378	0.5
GIoU_{3D}&UGIoU_{3D}	0.474	0.419	295	4.9
GIoU_{3D}&D_{KL}	0.472	0.417	299	13.0

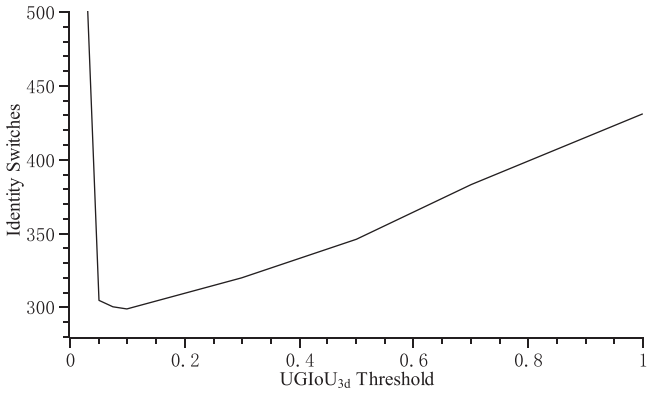


Fig. 5. Ablation on the association threshold for 3D MOT.

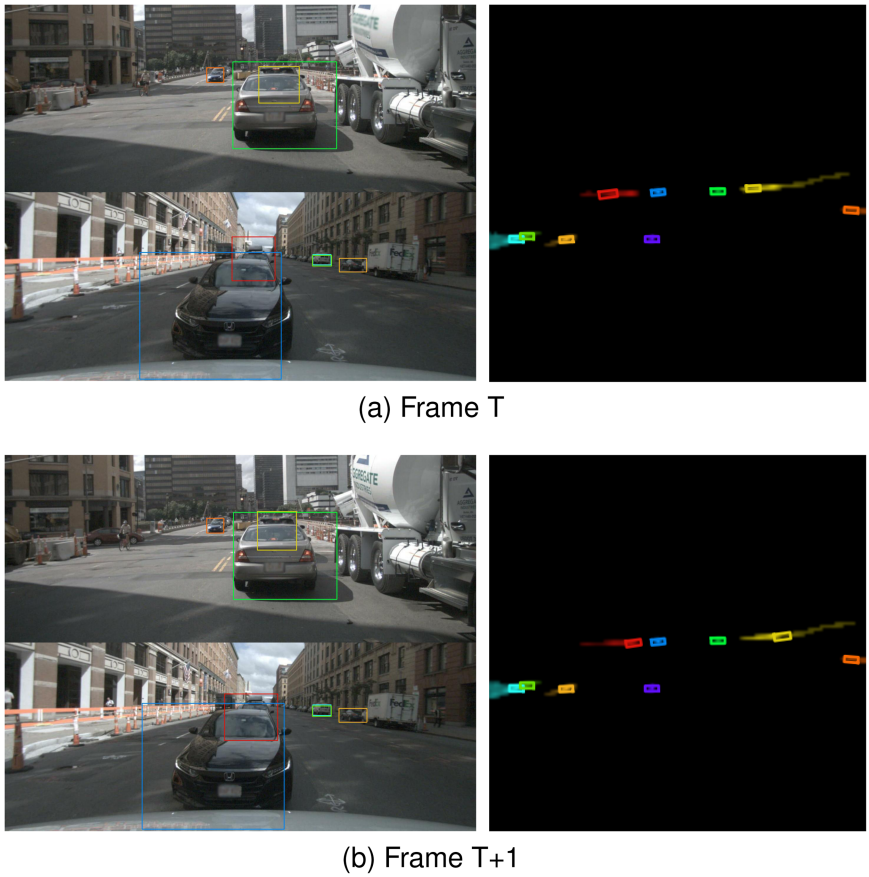


Fig. 6. *Qualitative result for 3D MOT.* In the given scene, the second vehicles in front of and behind the ego vehicle are severely occluded. Their localization in adjacent frames is unstable but can be associated with the visualized uncertain area.

TABLE XIV
ABLATION ON THE MAX SIZE OF UNCERTAINTY AREA

Uncertainty Area Size / m	AMOTA↑	IDS↑
2×2	0.471	359
4×4	0.471	315
8×8	0.472	299
16×16	0.470	300

TABLE XV
LATENCY IMPACT ABLATION

Backbone	BAF module	Uncertainty Extraction	Latency
Res-50	×	×	0.242s
	✓	×	0.261s
	✓	✓	0.269s
Swin-B	×	×	3.752s
	✓	×	3.774s
	✓	✓	3.783s

By representing the location of 3D bounding boxes with uncertain distributions and using UGIoU_{3D} as association metric, we further improve the AMOTA performance of our tracker by 1.1% and further reduce identity switches. However, the high computational cost of UGIoU_{3D} also leads to the significantly decreased efficiency of the tracker (15.4 FPS to 0.5 FPS). To address this, we propose to perform a two-stage bipartite matching in the tracking process, represented by GIoU_{3D}&UGIoU_{3D} in Table XIII. We first match 3D boxes with small localization errors with GIoU_{3D} to save computational cost, and then solve the remaining hard cases with UGIoU_{3D}. The two-stage association strategy not only reduces the computational cost of utilizing UGIoU_{3D} as the association metric, but also further reduces identity switches by taking the benefits of two metrics.

Furthermore, we also propose the simplified substitute of UGIoU_{3D}, the D_{KL}, which is nearly equivilant to UGIoU_{3D} as the second-stage assocation metric. Our two-stage tracker equipped with D_{KL} can run at 13 FPS, close to the one-stage tracker with GIoU_{3D} as the association metric.

UGIoU_{3D} tracking association threshold: As shown in Fig. 5, setting the UGIoU_{3D} tracking association threshold to around 0.1 0.2 is optimal. A lower association threshold will introduce more wrong associations while a higher threshold will depress the possible uncertainty-guided correct association.

The max size of uncertainty area: When modeling the localization uncertainty distribution of detections from the objectness map, we need to set a max range of uncertainty area for each detection. In Table XIV, we perform ablation experiments on

TABLE XVI
PER-CATEGORY TRACKING PERFORMANCE ANALYSIS

Metrics	Association	Car	Pedestrian	Truck	Bus	Bicycle	Motorcycle	Trailer
AMOTA	$GIoU_{3D}$	0.633	0.303	0.531	0.604	0.376	0.449	0.326
	$GIoU_{3D}$&$UGIoU_{3D}$	0.645	0.326	0.533	0.603	0.389	0.461	0.328
AMOTP	$GIoU_{3D}$	0.887	1.524	1.065	1.088	1.273	1.219	1.471
	$GIoU_{3D}$&$UGIoU_{3D}$	0.892	1.530	1.065	1.088	1.282	1.233	1.474
IDS	$GIoU_{3D}$	175	239	6	0	2	5	4
	$GIoU_{3D}$&$UGIoU_{3D}$	129	153	6	0	3	4	4

TABLE XVII
PER-CATEGORY DETECTION PERFORMANCE ANALYSIS

Metrics	+ours	Car	Pedestrian	Truck	Bus	Bicycle	Motorcycle	Trailer	Construction Vehicle	Traffic Cone	Barrier
AP	×	0.548	0.374	0.278	0.324	0.260	0.274	0.109	0.091	0.524	0.528
	✓	0.554	0.381	0.332	0.420	0.349	0.410	0.158	0.114	0.527	0.551
ATE	×	0.480	0.719	0.684	0.695	0.598	0.707	1.074	0.860	0.492	0.502
	✓	0.494	0.706	0.694	0.690	0.535	0.614	1.046	0.939	0.462	0.516
ASE	×	0.159	0.297	0.224	0.200	0.283	0.273	0.250	0.512	0.355	0.264
	✓	0.172	0.295	0.232	0.215	0.256	0.267	0.234	0.547	0.350	0.269
AOE	×	0.098	0.803	0.109	0.108	1.457	0.953	0.429	1.476	-	0.166
	✓	0.223	0.777	0.229	0.205	1.190	1.051	0.467	1.472	-	0.162
AVE	×	0.333	0.463	0.317	0.704	0.193	0.481	0.295	0.125	-	-
	✓	0.367	0.467	0.283	0.639	0.181	0.423	0.189	0.119	-	-
AAE	×	0.208	0.266	0.219	0.257	0.008	0.235	0.131	0.335	-	-
	✓	0.217	0.258	0.217	0.309	0.010	0.236	0.073	0.320	-	-

the max size of the uncertainty area. Experiments show that an uncertain area max size over 8 m× 8 m is optimal.

Qualitative results for 3D MOT: In the given scenes shown in Fig. 6, the second vehicles in front of and behind the ego vehicle are severely occluded, resulting in unstable localization among adjacent frames. As shown in BEV view, the uncertain localization distribution of two vehicles can be extracted from the objectness heat map and guide cross-frame detection association.

Latency analysis: The latency of the detector is primarily determined by the latency of its backbone. As shown in Table XV, the computational cost introduced by the BAF module and uncertainty extraction is negligible compared to the computational cost of large-scale backbones like SwinTransforme-Base.

Per-category performance analysis: We show the per-category performance of our method on detection and tracking tasks in Tables XVII and XVI. Experiment results show that our proposed method can boost the detection and tracking performance regarding all object categories on the nuScenes dataset.

V. CONCLUSION

In this work, we delve into the uncertain object representation in the field of camera-based 3D detection and 3D multi-object tracking. We propose to represent the location of objects as a probability distribution in 3D space to meet the uncertainty of localizing objects in images. With the uncertainty area of object localization extracted as a vice-product in the detection process, we gather the redundant predictions of objects to generate their uncertain localization representation in BEV view. Redundant prediction suppression with the proposed BAF module significantly boosted the performance of our baseline multi-camera 3D detector BEVDet4D and monocular 3D detector DD3D on nuScenes and KITTI-3D datasets. We further extend the application of the uncertain representation of objects to camera-based 3D multi-object tracking. We propose $UGIoU_{3D}$ and D_{KL} for enhancing cross-frame association by measuring the similarity between uncertain object localization distributions. Our proposed tracking method outperforms all previous methods on nuScenes tracking benchmark, which illustrates the benefits of the uncertain representation of objects to the downstream tasks of 3D camera-based detection.

REFERENCES

[1] Z. Li et al., "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 1–18.
[2] Y. Liu, T. Wang, X. Zhang, and J. Sun, "PETR: Position embedding transformation for multi-view 3D object detection," 2022, *arXiv:2203.05625*.
[3] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. Conf. Robot Learn.*, 2022, pp. 180–191.

[4] H. Chen, P. Wang, F. Wang, W. Tian, L. Xiong, and H. Li, "EPro-PnP: Generalized end-to-end probabilistic perspective-N-points for monocular object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2781–2790.

[5] D. Park, R. Ambrus, V. Guizilini, J. Li, and A. Gaidon, "Is pseudo-LiDAR needed for monocular 3D object detection?," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3142–3152.

[6] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12697–12705.

[7] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11784–11793.

[8] L. Fan et al., "Embracing single stride 3D object detector with sparse transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8458–8468.

[9] L. Fan, F. Wang, N. Wang, and Z. -X. Zhang, "Fully sparse 3D object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 351–363.

[10] J. Huang and G. Huang, "BEVDet4D: Exploit temporal cues in multi-camera 3D object detection," 2022, *arXiv:2203.17054*.

[11] 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.*, 2012, pp. 3354–3361.

[12] T. Wang, X. Zhu, J. Pang, and D. Lin, "FCOS3D: Fully convolutional one-stage monocular 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 913–922.

[13] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.

[14] X. Liu, N. Xue, and T. Wu, "Learning auxiliary monocular contexts helps monocular 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1810–1818.

[15] X. Ma et al., "Delving into localization errors for monocular 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4721–4730.

[16] Y. Chen, L. Tai, K. Sun, and M. Li, "MonoPair: Monocular 3D object detection using pairwise spatial relationships," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12093–12102.

[17] Y. Li, Y. Chen, J. He, and Z. Zhang, "Densely constrained depth estimator for monocular 3D object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 718–734.

[18] Z. Liu, D. Zhou, F. Lu, J. Fang, and L. Zhang, "AutoShape: Real-time shape-aware monocular 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15641–15650.

[19] Y. Zhang, J. Lu, and J. Zhou, "Objects are different: Flexible monocular 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3289–3298.

[20] T. Wang, Z. Xinge, J. Pang, and D. Lin, "Probabilistic and geometric depth: Detecting objects in perspective," in *Proc. Conf. Robot Learn.*, 2022, pp. 1475–1485.

[21] Z. Chong et al., "MonoDistill: Learning spatial features for monocular 3D object detection," 2022, *arXiv:2201.10830*.

[22] Y. Wang, W. -L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8445–8453.

[23] Y. You et al., "Pseudo-LiDAR++: Accurate depth for 3D object detection in autonomous driving," 2019, *arXiv:1906.06310*.

[24] R. Qian et al., "End-to-end pseudo-LiDAR for image-based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5881–5890.

[25] Y. Li et al., "BEVDepth: Acquisition of reliable depth for multi-view 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 2, 2023, pp. 1477–1485.

[26] J. Huang, G. Huang, Z. Zhu, and D. Du, "BEVDet: High-performance multi-camera 3D object detection in bird-eye-view," 2021, *arXiv:2112.11790*.

[27] Y. Jiang et al., "PolarFormer: Multi-camera 3D object detection with polar transformers," in *Proc. AAAI Conf. Artif. Intell.* vol. 37, no. 1, 2023, pp. 1042–1050.

[28] P. Tokmakov, J. Li, W. Burgard, and A. Gaidon, "Learning to track with object permanence," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 10840–10849.

[29] T. Fischer, Y. Yang, S. Kumar, M. Sun, and F. Yu, "CC-3DT: Panoramic 3D object tracking via cross-camera fusion," in *Proc. Mach. Learn. Res.*, 2022, pp. 2294–2305.

[30] H. -N. Hu, Y. -H. Yang, T. Fischer, T. Darrell, F. Yu, and M. Sun, "Monocular quasi-dense 3D object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1992–2008, Feb. 2023.

[31] N. Marinello, M. Proesmans, and L. Van Gool, "TripletTrack: 3D object tracking using triplet embeddings and LSTM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 4499–4509.

[32] T. Zhang, X. Chen, Y. Wang, Y. Wang, and H. Zhao, "MUTR3D: A multi-camera tracking framework via 3D-to-2D queries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4537–4546.

[33] Z. Pang, J. Li, P. Tokmakov, D. Chen, S. Zagoruyko, and Y. Wang, "Standing between past and future: Spatio-temporal modeling for multi-camera 3D multi-object tracking," 2023, *arXiv:2302.03802*.

[34] D. Feng, A. Harakeh, S. L. Waslander, and K. Dietmayer, "A review and comparative study on probabilistic object detection in autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 9961–9980, Aug. 2022.

[35] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," in *Proc. Adv. Neural Inf. Process. Syst. 30: Annu. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 5574–5584. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html

[36] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf, "Dropout sampling for robust object detection in open-set conditions," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 3243–3249.

[37] F. Kraus and K. Dietmayer, "Uncertainty estimation in one-stage object detection," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2019, pp. 53–60.

[38] D. Miller, F. Dayoub, M. Milford, and N. Sünderhauf, "Evaluating merging strategies for sampling-based uncertainty techniques in object detection," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 2348–2354.

[39] D. Miller, N. Sünderhauf, H. Zhang, D. Hall, and F. Dayoub, "Benchmarking sampling-based probabilistic object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 42–45. [Online]. Available: http://openaccess.thecvf.com/content%5C_CVPRW%5C_2019/html/Uncertainty%5C_and%5C_Robust%0Aness%5C_in%5C_Deep%5C_Visual%5C_Learning/Miller%5C_Benchmarking%5C_Samplingbased%5C_%0AProbabilistic%5C_Object%5C_Detectors%5C_CVPRW%5C_2019%5C_paper.htm

[40] A. Harakeh, M. Smart, and S. L. Waslander, "BayesOD: A Bayesian approach for uncertainty estimation in deep object detectors," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 87–93.

[41] Y. Gal et al., "Uncertainty in deep learning," Ph.D. dissertation, Univ. Cambridge, 2016.

[42] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6405–6416.

[43] J. Gast and S. Roth, "Lightweight probabilistic deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3369–3378.

[44] Z. Li, Z. Qu, Y. Zhou, J. Liu, H. Wang, and L. Jiang, "Diversity matters: Fully exploiting depth clues for reliable monocular 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2791–2800.

[45] G. P. Meyer, A. Laddha, E. Kee, C. Vallespi-Gonzalez, and C. K. Wellington, "LaserNet: An efficient probabilistic 3D object detector for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12677–12686.

[46] S. Choi, K. Lee, S. Lim, and S. Oh, "Uncertainty-aware learning from demonstration using mixture density networks with sampling-free variance modeling," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 6915–6922.

[47] Y. Zhang, W. Zheng, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "A simple baseline for multi-camera 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 3, 2023, pp. 3507–3515.

[48] X. Shi, Q. Ye, X. Chen, C. Chen, Z. Chen, and T. -K. Kim, "Geometry-based distance decomposition for monocular 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15172–15181.

[49] Y. Shi et al., "SRCN3D: Sparse R-CNN 3D surround-view camera object detection and tracking for autonomous driving," 2022, *arXiv:2206.14451*.

[50] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[51] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS–improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5561–5569.

[52] Z. Pang, Z. Li, and N. Wang, "SimpleTrack: Understanding and rethinking 3D multi-object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 680–696.
[53] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11618–11628.
[54] X. Weng and K. Kitani, "A baseline for 3D multi-object tracking," 2019, *arXiv:1907.03961*.
[55] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, 2008.
[56] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Compu. Vis.*, 2021, pp. 10012–10022.
[57] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[58] B. Zhu, Z. Jiang, X. Zhou, Z. Li, and G. Yu, "Class-balanced grouping and sampling for point cloud 3D object detection," 2019, *arXiv:1908.09492*.
[59] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2403–2412.
[60] Y. Lee and J. Park, "CenterMask: Real-time anchor-free instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13906–13915.
[61] Q. Wang, Y. Chen, Z. Pang, N. Wang, and Z. Zhang, "Immortal tracker: Tracklet never dies," 2021, *arXiv:2111.13672*.
[62] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. 5th Annu. Conf. Robot Learn.*, London, U.K., vol. 164, Nov. 2021, pp. 180–191. [Online]. Available: https://proceedings.mlr.press/v164/wang22b.html
[63] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3D object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 18442–18455.
[64] Y. Wang, Y. Chen, and Z. Zhang, "FrustumFormer: Adaptive instance-aware resampling for multi-view 3D detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5096–5105.
[65] C. Yang et al., "BEVFormer V2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17830–17839.
[66] G. Brazil and X. Liu, "M3D-RPN: Monocular 3D region proposal network for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9287–9296.
[67] M. Ding et al., "Learning depth-guided convolutions for monocular 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 1000–1001.
[68] S. Chen, X. Wang, T. Cheng, Q. Zhang, C. Huang, and W. Liu, "Polar parametrization for vision-based surround-view 3D detection," 2022, *arXiv:2206.10965*.

Qitai Wang received the bachelor's degree from Tsinghua University, Beijing, in 2020. He is currently working toward the PhD degree with the Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing. His research interests include 3D perception, generative models, and driving simulation.

Yuntao Chen received the bachelor's degree from the University of Science and Technology, Bejing, in 2016, and the PhD degree from the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2021. His research interests include object recognition, 3D scene understanding, and generative models.

Zhaoxiang Zhang (Senior Member, IEEE) received the BS degree in circuits and systems from the University of Science and Technology of China in 2004, and the PhD degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences in 2009. From 2009 to 2011, he was with the School of Computer Science and Engineering, Beihang University, as an assistant professor and an associate professor from 2012 to 2015. In 2015, he returned to the Institute of Automation, Chinese Academy of Sciences as a full professor with the Center for Research on Intelligent Perception and Computing and the National Laboratory of Pattern Recognition. He has authored or coauthored more than 200 papers in international journals and conferences, including reputable international journals such as *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IJCV, and JMLR, and top-level international conferences such as CVPR, ICCV, ECCV, ICLR, NeurIPS, AAAI, and IJCAI. His research interests include computer vision, pattern recognition, and machine learning. He specifically focuses on biologically inspired intelligent computing and its applications in human analysis and scene understanding. He is the associate editor for IJCV, *IEEE Transactions on Circuits and Systems for Video Technology*, *Pattern Recognition*, and *Frontiers of Computer Science*. He was the area chair of reputable international conferences such as CVPR, ICCV, AAAI, IJCAI, ACM MM, ICPR, and ACCV.