2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

CrossDTR: Cross-view and Depth-guided Transformers for 3D Object Detection

Ching-Yu Tseng[1], Yi-Rong Chen[1], Hsin-Ying Lee[1], Tsung-Han Wu[1], Wen-Chin Chen[1], and Winston H. Hsu[1,2]

***Abstract*— To achieve accurate 3D object detection at a low cost for autonomous driving, many multi-camera methods have been proposed and solved the occlusion problem of monocular approaches. However, due to the lack of accurate estimated depth, existing multi-camera methods often generate multiple bounding boxes along a ray of depth direction for difficult small objects such as pedestrians, resulting in an extremely low recall. Furthermore, directly applying depth prediction modules to existing multi-camera methods, generally composed of large network architectures, cannot meet the real-time requirements of self-driving applications. To address these issues, we propose Cross-view and Depth-guided Transformers for 3D Object Detection, CrossDTR. First, our lightweight *depth predictor* is designed to produce precise object-wise sparse depth maps and low-dimensional depth embeddings without extra depth datasets during supervision. Second, a *cross-view depth-guided transformer* is developed to fuse the depth embeddings as well as image features from cameras of different views and generate 3D bounding boxes. Extensive experiments demonstrated that our method hugely surpassed existing multi-camera methods by 10 percent in pedestrian detection and about 3 percent in overall mAP and NDS metrics. Also, computational analyses showed that our method is 5 times faster than prior approaches. Our codes will be made publicly available at https://github.com/sty61010/CrossDTR.**

I. Introduction

Detecting instances of objects in the 3D space from sensor information, i.e. 3D object detection, is crucial for various intelligent systems, such as autonomous driving and indoor robotics. Previous work tends to rely on accurate depth information from different sensors, such as LiDAR signals [1]–[3] and binocular information [4], [5], to accomplish superior performance. In recent years, in order to achieve high-quality detection at a low cost, several methods based on commodity cameras have been proposed. Among them, since naive monocular detection [6]–[12] suffers from the problems of occlusion and deficiency of cross-view information, methods transforming camera information from multiple views into Bird-Eye-View [13]–[21], called multi-view methods, has received increasing attention.

Though these multi-view methods have made some progress with cross-view information and Bird-Eye-View representation [14]–[16], [22]–[24], we observe existing practices suffered from either extremely low recall in small objects due to inaccurate depth or an unaffordable computational burden because of complex depth prediction modules. Specifically, while some methods fusing information from multiple views [13]–[21] easily locate the pixel coordinate of small objects in images, they can hardly estimate precise distances of objects from the image plane. Consequently, these methods tend to predict a row of false positive bounding boxes along a ray of depth direction in candidate regions when detecting small objects (shown in Fig. 1), leading to low recall in perception and poor follow-up prediction and planning. In addition, some previous monocular approaches utilized complex depth prediction modules [25], [26] or large-scale depth-pretrained backbone [27]–[30] to provide depth cues. Nevertheless, directly applying them to existing multi-camera methods, generally composed of large network architectures, cannot meet the real-time requirements of the self-driving application (Tab. II). From the two observations above, we conclude that a module is needed to obtain depth hints from multiple cameras and fuse both depth and image information from different views in real time.

To achieve the goal, we proposed CrossDTR, a novel end-to-end Cross-view and Depth-guided Transformer network for multi-camera 3D object detection as shown in Fig.

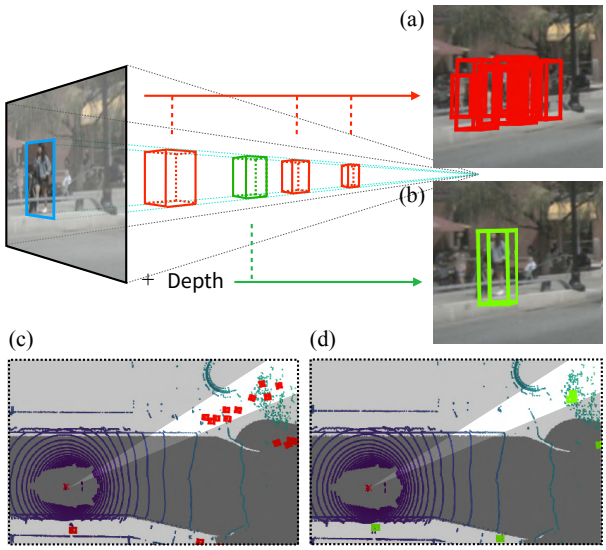

Fig. 1: **Multi-camera methods suffer from inaccurate depth estimation.** Red and green bounding boxes represent inaccurate and accurate predictions respectively. The above 2D-to-3D projection diagram mainly shows that (a) previous multi-view methods usually produce a row of false positives predictions alone a ray of depth, but (b) our method, guided with depth hints, can precisely predict only one bounding box. Plot (c) and (d) demonstrate the corresponding bird-eye view predictions of (a) and (b).

[1]National Taiwan University, [2]Mobile Drive Technology

2. To efficiently obtain depth hints for downstream 3D object detection, we leverage a lightweight *depth predictor* to produce precise depth maps for each view (Sec. III-D). Specifically, inspired by previous depth-aware methods [25], [26], [31]–[33], the depth predictor is supervised with our generated object-wise sparse depth maps without extra depth dataset (Sec. III-C). Then, to fused the depth and image information from multi-view cameras effectively, we propose a novel *cross-view and depth-guided transformer* (Sec. III-E). In short, the Transformer Encoder is used to compress high-resolution depth maps into low-dimensional depth embeddings, and the Transformer Decoder performs the cross-attention mechanism among depth as well as image information from multi-views.

Experimental results demonstrated that our depth-guided method resolves the problem of false positive predictions on small objects (Fig. 1) and achieves overall improvement with the limited computational burden. Compared with existing multi-camera methods on the nuScenes dataset [34], we increased by 10 percent Average precision (AP) in pedestrian detection and about 3 percent in mAP and NDS metrics. Also, computational analyses demonstrated that our lightweight method is 5 times faster than prior methods under similar network backbones. To sum up, the overall contributions of this work can be summarized as follows:

- We build up a novel cross-view and depth-guided perception framework, CrossDTR, to insert accurate depth cues into multi-view detection methods.
- Our proposed depth-guided module can alleviate the problem of false positive predictions along the direction of depth for small objects.
- Our framework achieves state-of-the-art 3D detection performance on the nuScenes dataset [34] with fewer computational budgets compared with existing multi-view or depth-guided methods.

II. Related Work

A. Monocular 3D Object Detection

Monocular 3D object detection [6], [8]–[12] has received lots of attention due to the low cost of commodity cameras. It originates from Orthographic Feature Transform (OFT) [6], which projects camera features into ego pose coordinate uniformly, voxelized the features, and uses the detector from PointPillar [2]. OFT [6] is the first method that deals with camera features in a LiDAR-based [1]–[3] technique, but OFT [6] predicts the depth value of each pixel uniformly and thus results in inaccurate depth estimation. Extend from OFT, Pseudo-lidar [8], [9] and CaDDN [10] methods use a convolutional neural network to predict depth distribution and auxiliary loss to enhance performance. Apart from the above methods, the other monocular methods [11], [12] directly regression 3D representation in camera coordinate. FCOS3D [11] is built on FCOS [35], which is a single-stage 2D object detection framework. Furthermore, PGD [12] is an improved version of FCOS3D [11] with extra depth information. In conclusion, depth information is critical for monocular methods to enhance performance.

B. Multi-Camera 3D Object Detection

Multi-camera methods [13]–[21] have been proved to solve the problem of occlusion through temporal and spatial information. DETR-based methods [13], [16]–[18] were first proposed. DETR3D [13] was built based on DETR [36] and utilized the attention mechanism to select features from different camera. Unlike DETR [36] in the 2D space, we can not assure the candidate area of objects in the 3D space, so DETR3D [13] initialize object query randomly with uniform distribution and pick features by projecting object position into camera coordinate. As an improved version of DETR3D [13], PETR [17] enhance performance by initializing object queries with the method from Anchor DETR [37] and applying 3D positional embedding to embed 3D coordinate frustum into multi-head attention. PETRv2 [18] optimized PETR by adding temporal information. Additionally, some researchers regard that Bird-Eye-View (BEV) [22]–[24], [38]–[41] representations can provide better space concepts in 3D coordinates. BEVDet methods [14], [15] transform image features into BEV according to Lift-Splat-Shoot (LSS) Method [22] and propose BEV data augmentation [14] to avoid over-fitting. BEVFormer [16] proposed BEV query and use the deformable attention [42] to suppress computation. BEVFormer [16] also applies temporal information by cross-attention between BEV queries from different time stamps. However, those methods still lack accurate depth estimation.

C. Depth-guided Monocular Methods

Monocular methods can not achieve competitive performance in comparison with LiDAR-based methods [1]–[3] and binocular methods [4], [5]. The main reason is inaccurate depth estimation, and thus the depth-aware methods [25], [26], [31]–[33], [43] are proposed recently. ASTransformer [43] first proposed to use depth maps supervision to enhance the performance of depth estimation. Besides, both MonoDTR [32] and MonoDETR [31] compose depth-aware embedding from depth maps and are supervised by ground truth depth maps to enhance the performance of monocular object detection. However, all the above methods still build on monocular methods and will suffer from complex post-processing between cameras to aggregate all information and remove repeated predicted bounding boxes. Multi-camera method with a depth-guided module is still missing so far.

III. Method

A. Problem Definition

In this work, we aim to precisely detect instances of objects in 3D space given multiple scanned RGB images [44]. Specifically, let $\mathbf{I}_{sensors} = \{\mathcal{I}_1, ..., \mathcal{I}_{N_{cameras}}\}$ represent a set of scanned multi-view images, and $\mathbf{B}_{LiDAR} = \{\mathcal{B}_1, ..., \mathcal{B}_{N_{box}}\}$ denotes a set of ground truth bounding boxes, a 3D bounding box $\hat{\mathcal{B}}_i \in \mathbf{B}_{LiDAR}$ is formulated as a vector with 7 degree of freedom:

$$\hat{\mathcal{B}}_i = (x_c, y_c, z_c, l, w, h, \theta), \tag{1}$$

where (x_c, y_c, z_c) denotes the center of each bounding box. (l, w, h) represents the length, width, and height of the cuboid

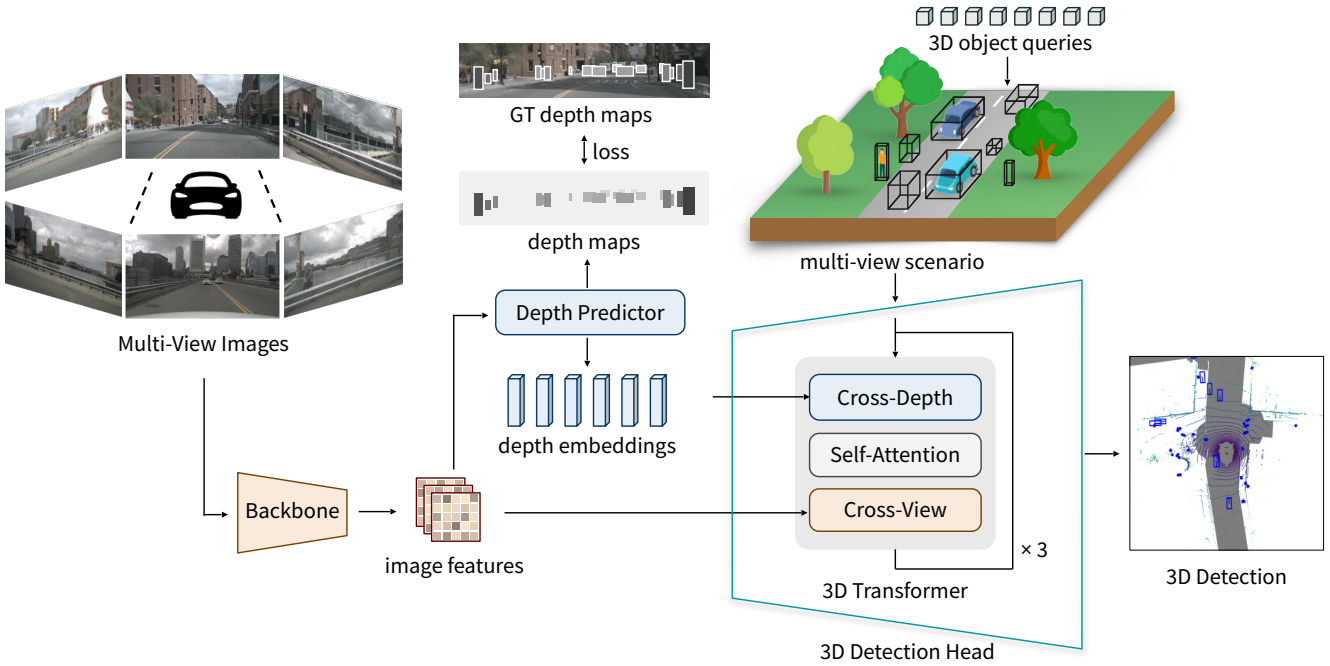

Fig. 2: **The overall framework of our proposed CrossDTR.** First, Multi-view images are fed into a feature extractor backbone to generate image features. Then, image features are fed into our Depth Predictor (in Sec. III-D) to produce depth embeddings and predicted depth maps. Lastly, given image features, depth embeddings, and 3D object queries, our Cross-view and Depth-guided Transformer (in Sec. III-E) conducts cross-view attention and cross-depth attention to generate 3D bounding boxes. Note that we minimize the difference between predicted depths and our generated object-wise sparse depth maps (in Sec. III-C) under supervision during training. Best viewed in color.

respectively. θ means orientation (yaw) of each object. Formally, given a set of predicted bounding boxes $\hat{\mathbf{B}}_{LiDAR}$, a multi-view 3D object detector $\boldsymbol{f}_{Det}$ is defined as follows:

$$\hat{\mathbf{B}}_{LiDAR} = \boldsymbol{f}_{Det}(\mathbf{I}_{sensors}). \tag{2}$$

B. Overall Architecture

Fig. 2 illustrates our architecture. First, we feed our multi-view images $\{\mathcal{I}_1, ..., \mathcal{I}_{N_{cameras}}\}$ into our model. Without external data, [45] is applied as our backbone to extract features $\mathcal{F}_{view}$ for each view. Then, we feed image features $\mathcal{F}_{view}$ into Depth Predictor (in Sec. III-D). Given single-view image features $\mathcal{F}_{view}$, the Depth Predictor produces low-dimensional depth embeddings and depth maps by a Transformer encoder. During training, we minimize the difference between the predicted depth maps and our generated sparse depth maps (in Sec. III-C) in a supervised manner. Lastly, given image features, depth embeddings, and 3D object queries, our Cross-view and Depth-guided Transformer (in Sec. III-E) conducts cross-view attention and cross-depth attention to generate 3D bounding boxes.

C. Object-wise Sparse Depth Map

Unlike some prior depth-guided monocular methods requiring costly dense depth maps during training, we extend [31] to multi-view scenarios and leverage only the sparse depth hints provided by the raw LiDAR data, which is more cost-effective. We first detail our depth generation process below.

Given a camera matrix $T \in \mathbb{R}^{3\times4}$ and a point $p \in \mathbb{R}^3$ in the LiDAR coordinate, we define the transformation function $\mathcal{T}$ from the LiDAR coordinate to the camera coordinate as follows:

$$\begin{aligned} \mathcal{T}(T,p) &= \begin{bmatrix} u & v & d \end{bmatrix}^T, \\ \text{where } d \cdot \begin{bmatrix} u & v & 1 \end{bmatrix}^T &= T \cdot (p \oplus 1), \end{aligned} \tag{3}$$

and $\oplus$ denotes tensor concatenation. We transform the center point and corners of each bounding box into each camera view by (3). Let $p^{m,i}_{centers}, d^{m,i}_{centers}, \mathbf{P}^{m,i}_{corners}$ be the center point, the depth value of the center, and the set of corner points of the bounding box $\mathcal{B}_i$ in the m-th camera coordinate, then

$$p^{m,i}_{centers} = \begin{bmatrix} u_c & v_c & d_c \end{bmatrix}^T = \mathcal{T}(T_m, p_i), \tag{4}$$

$$\mathbf{P}^{m,i}_{corners} = \{\mathcal{T}(T_m, p) \mid p \in \mathcal{C}(\mathcal{B}_i)\}, \tag{5}$$

where $T_m \in \mathbf{T}$, $p_i = \begin{bmatrix} x_c & y_c & z_c \end{bmatrix}^T$ and $x_c, y_c, z_c \in \mathcal{B}_i$. $\mathcal{C}(\mathcal{B})$ is the function returning 8 corners of the given 3D bounding box $\mathcal{B}$. We further extract the 2D bounding box $\mathcal{B}^{2d}_{m,i} = (u^i_{min}, u^i_{max}, v^i_{min}, v^i_{max})$ with respect to $\mathcal{B}_i$ in the m-th camera from $\mathbf{P}^{m,i}_{corners}$ by getting the minimum and maximum (u, v) value in $\mathbf{P}^{m,i}_{corners}$.

Next, we collect all valid 2D bounding boxes $\mathcal{B}^{2d}_{m,i}$ for each camera and their depth values $d^{m,i}_{centers}$ to a new set $\boldsymbol{\mathcal{V}}_m$. We set the pixel value to the depth of the object center point if the pixel lies in an object's bounding box. If the pixel lies in multiple bounding boxes, we set it to the nearest one. Lastly, the object-wise sparse depth map is obtained by adopting linear-increasing discretization (LID) [46].

D. Depth Predictor

Inspired by previous depth-guided methods [31], [32] and other methods [10], [14], [15], [22], [23], we utilize the Depth Predictor from [31] to learn depth information from object-wise sparse depth maps. To save the memory of our model, we use lightweight architecture built by convolution layers to predict depth distribution and match the number of depth bins as 3D positional embedding [17], [18]. Given image features $\mathcal{F}_{view}$, we use light-weight network $\boldsymbol{f}_{ddn}$ to predict depth logits $\hat{\mathcal{D}}$ and depth probability $\hat{\mathcal{D}}_{prob}$ among each depth map. Besides, we utilize Transformer encoder ψ_i to encode image features $\mathcal{F}_{view}$ into depth embeddings $\mathcal{F}_{depth}$ with multi-head attention ψ_i:

$$\begin{aligned} &\mathcal{E}_0 = \mathcal{F}_{view}, \\ &\mathcal{E}_i = \psi_i(\mathcal{E}_{i-1}, \psi_{i-1}), i = 1, ..., I, \\ &\mathcal{F}_{depth} = \mathcal{E}_I, \end{aligned} \tag{6}$$

where $\mathcal{E}_i \in \mathbb{R}^{L \times C}$ represents the depth embeddings from the Transformer encoder in the depth predictor. C is the size of the embeddings. $L = H_d \times W_d$ is the length of depth embeddings. The number i denotes the i-th layer in the encoder, and the encoder contains total I layers. The final depth embedding is utilized in Sec. III-E.

E. Cross-view and Depth-guided Transformer

As the Transformer has successfully been used to fuse different modalities, we adopt it to combine both image features and depth embeddings. We adopt PETR [17] methods to conduct attention between different views. Besides, inspired by MonoDETR [31], we insert depth embedding into multi-view attention from PETR [17].

Cross-view Attention. We follow cross-view attention as methods from PETR [17], and we feed image features $\mathcal{F}_{view}$ as keys and values. We utilize 3D positional embedding [17] as query positional embedding.

Cross-depth Attention. Previous methods [17], [18] only use visual information and thus lack depth cues for the detector. Inspired by [31], [32] and methods [25], [26], [31]–[33], [43], we suggest inserting depth hints as depth embedding into the detector to provide more detailed information for small objects. After we obtain depth embeddings $\mathcal{F}_{depth} \in \mathbb{R}^{B \times N \times C \times H_d \times W_d}$ from Sec. III-D, we flatten depth embeddings into $\mathcal{F}_{depth} \in \mathbb{R}^{\acute{L} \times B \times C}$ where $\acute{L} = N \times H_d \times W_d$ indicates the flatten depth embeddings $\mathcal{F}_{depth}$ among all camera views. Then, we follow the previous Sec. III-E and select depth embeddings as keys and values for multi-head attention. Those depth embeddings can not only learn pixel-level depth hints in a single view in Depth Predictor (in Sec. III-D) but also consider depth messages from the other views during cross attention mechanism.

F. 3D Detection Head and Loss

3D Detection Head. To learn information between the camera view features and 3D position, we adopt the 3D Detection Head from PETR [17]. The 3D Detection Head from PETR [17] initialize 3D reference points among 3D space and adopt multi-layer perception to learn the candidate area in the ego-pose coordinate. Then, the 3D Detection Head generates 7 degrees of freedom to represent bounding boxes in the ego-pose coordinate.

Depth Distribution Network Loss. To conduct the depth-guided method on a predefined depth map in Sec. III-C, we borrow the depth-guided method from [31] and refer to CaDDN [10] and adopt Depth Distribution Network Loss (DDN Loss) to regularize the predicted depth map values and predicted depth map logits. Following CaDDN [10], we build our loss as the following equation:

$$\boldsymbol{L}_{ddn} = \frac{1}{W_d \cdot H_d} \sum_{u=1}^{W_d} \sum_{v=1}^{H_d} \boldsymbol{FL}(\mathcal{D}(u,v), \hat{\mathcal{D}}(u,v)), \tag{7}$$

where W_d and H_d represent the size of depth map logits, and (u, v) denotes the position of pixels. ***FL*** means adopted Focal Loss [47].

IV. Experiments

A. Setup

Dataset. In this paper, we use the nuScenes dataset [34] as our benchmark, which provides camera, radar, and LiDAR sensor data with 3D bounding box annotations. Its data is mainly composed of camera data and provides only sparse LiDAR data as an auxiliary. As it lacks data to provide depth information like dense LiDAR or depth maps, we generate object-wise sparse depth maps (in Sec. III-C for the cross-view and depth-guided method. The nuScenes dataset [34] contains 1000 scenes and each scene is 20 seconds in length and annotated at 2HZ.

Implementation Details. Following training policies from [13], [17], [18], we use features from backbones [29], [30], [45], [49] with downsample scale of $\frac{1}{16}$ and $\frac{1}{32}$. And we feed the features into both depth predictor and cross-view attention. Besides, we follow [17], [18] to sample 64 points for depth in 3D positional embeddings and also for depth predictor to estimate depth distribution. Specifically, we supervise generated depth maps only during training. We use the AdamW optimizer with a learning rate of 2e-4 to train our model. We train our model for 24 epochs on 4 Nvidia 3090 GPUS with a total batch size of 8 for 48 hours. We use input images at resolution 512×1408 for our baseline model.

Baselines. We compare CrossDTR with both monocular and multi-camera approaches. CenterNet [7], FCOS3D [11], and PGD [12] represent monocular approaches, while DETR3D [13], PETR [17], and BEVDet [14] serve as the baselines of multi-camera ones. For fair comparison, we only adopt the performance of these approaches without tricks such as test-time augmentation [11], [12], CBGS [48], and oversampling [48]. Besides, the comparison with the lightweight BEV-Former [16] (from their official repository) without encoding extra temporal information is also included.

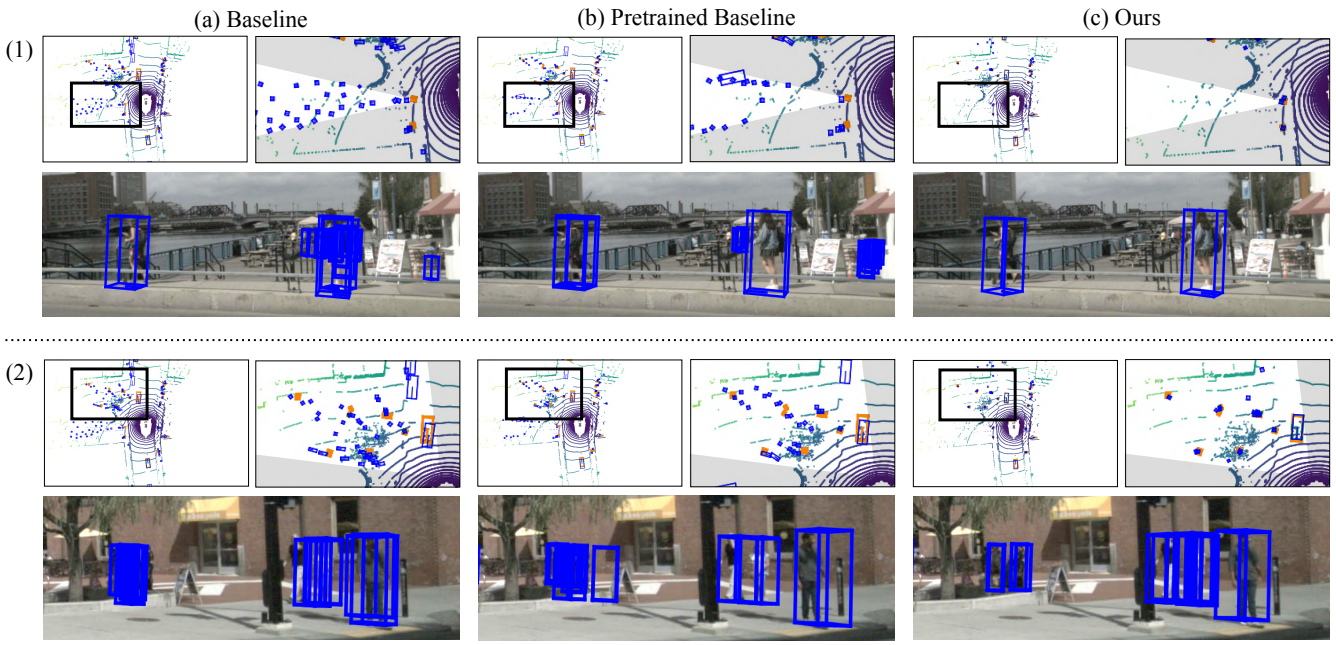

Fig. 3: **Visualization of false positive predictions on the nuScenes validation set.** We provide two qualitative examples, (1) and (2), with bird-eye-view (BEV) and camera-view representations. In the first row, the left and right images illustrate the focused areas of the global and zoomed-in BEVs, where blue and orange bounding boxes represent predictions and ground truth respectively. In the second row, the images in the camera view show predictions from models. Under a fair comparison with the same network backbone (PETR [17] with ResNet50 backbone [45]), our cross-view and depth-guided method (c) effectively mitigates the false positive issue in prior multi-view baselines (a), i.e. **does not produce repeated bounding boxes along the ray of depth**. Also, we even surpass methods equipped with a heavy-weight pretrained depth prediction module (b). Best viewed in color and zoom-in.

TABLE I: **Comparison with SOTA methods on the nuScene validation set.** PETR [17] are trained with CBGS [48]. The best results are shown in **bold**.

Methods	Backbone	Img Size	#param.	FPS	GFLOPs	mAP	NDS	mATE	mASE	mAOE	mAVE	mAAE
CenterNet [7]	DLA	1600*900	-	-	-	0.306	0.328	0.716	0.264	0.609	1.426	0.658
FCOS3D [11]	ResNet101	1600*900	52.5M	1.7	2008.2	0.295	0.372	0.806	0.268	0.511	1.315	**0.170**
PGD [12]	ResNet101	1600*900	53.6M	1.4	2223.0	0.335	0.409	0.732	**0.263**	**0.423**	1.285	0.172
Detr3D [13]	ResNet101	1600*900	**51.3M**	2.0	1016.8	0.303	0.374	0.860	0.278	0.437	0.967	0.235
BEVDet [14]	Swin-T	1408*512	126.6M	1.9	2962.6	0.349	0.417	**0.637**	0.269	0.490	0.914	0.268
PETR [17]	ResNet101	1408*512	59.2M	5.3	504.6	0.357	0.421	0.710	0.270	0.490	0.885	0.224
CrossDTR	ResNet101	1408*512	53.3M	**5.8**	**483.9**	**0.370**	**0.426**	0.773	0.269	0.482	**0.866**	0.203

TABLE II: **Comparison with lightweight multi-view methods.** We utilize **bold** to highlight the best results.

Methods	Backbone	#param.	FPS	GFLOPs	mAP
DETR3D [13]	ResNet101	51.3M	2.0	1016.8	0.303
BEVDet [14]	ResNet50	54.1M	9.3	452.0	0.299
BEVFormer [16]	ResNet50	68.7M	2.3	1303.5	0.252
PETR [17]	ResNet50	36.6M	10.4	297.2	0.317
CrossDTR	ResNet50	**31.8M**	**10.6**	**268.1**	**0.326**

TABLE III: **Ablation study of depth-guided module**. DE denotes Depth Embedding and DDN Loss denotes Depth Distribution Network Loss. Our method was built on PETR [17] with DDN Loss and DE.

Methods	DE	DDN Loss	mAP	NDS
PETR [17]			0.357	0.421
CrossDTR	✓		0.366	0.423
CrossDTR	✓	✓	**0.370**	**0.426**

TABLE IV: **Study of false positive predictions for the pedestrian class.** We choose PETR [17] with ResNet50 [45] and PETR [17] with depth-pretrained VoVNetV2 [30] as baselines and compare them with our method, CrossDTR, with different distance thresholds. We utilize **bold** to highlight the best results.

Methods	Depth-pretrained	Backbone	AP (Pedestrain) @ Dist.		
			[0.5]	[1.0]	[4.0]
PETR [17]		ResNet50	0.09	0.401	0.809
	✓	VoVNetV2	0.102	0.426	0.870
CrossDTR		ResNet50	**0.320**	**0.689**	**0.875**

Evaluation Metrics. We report mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), mean Average Attribute Error (mAAE), mean Average Precision (mAP), and nuScenes detection score (NDS). mAP estimates the distance between the centers of a predicted and a ground-truth 3D bounding box. To evaluate the efficacy of our method in solving false positive predictions, we report the Average Precision of *pedestrian* as

our metrics. We also take Frame Per Second (FPS) and Giga Flops (GFLOPs) into consideration to evaluate the real-time ability of multi-camera models.

B. Quantitative Results

Comparison with State-of-the-art. As shown in Tab. I, our method surpasses other previous methods and achieves the state-of-the-art performance of mAP and NDS on the validation dataset [34]. To begin with, CenterNet [7], FCOS3D [11], and PGD [12] are classic monocular baseline. Our method exceeds by more than **3 percent on mAP and 2 percent on NDS**. Additionally, compared with the SOTA multi-camera methods (starting from the fifth row), our method still surpasses all of them by at least **1.3 percent on mAP and 0.5 percent on NDS**. Swin-T represents Swin-Transformer [49], which is the strongest backbone among the Tab. I. Our method with ResNet101 also beats BEVDet [14] with Swin-T [49]. Then, our method needs the least computational resource **(483.9 GFLOPs and 5.8 FPS)**. Our method is lightweight and can conduct real-time 3D detection on the nuScenes [34] dataset.

Comparison with lightweight multi-view methods. Tab. II shows the comparison between our method and previous multi-camera methods. Note that all the scores are from their official repositories. Since we conduct our experiments on the validation set with limited computation resources, we choose a smaller backbone ResNet50 [45] to extract features from input images at resolution 512×1408. Our proposed method overtakes all previous multi-camera methods, even against DETR3D [13] with stronger ResNet101 backbone [45] and BEVFormer [16] with temporal information. Our method surpasses the second best method by **0.9 percent on mAP and 1.1 percent on NDS**. Besides, our model contains the least parameters as shown in Tab. II and attains the highest score **(10.6) on FPS**. The result shows that our model can conduct real-time detection.

C. Ablation Study

Tab. III shows the effectiveness of our cross-view and depth-guided module. We conduct an ablation study to verify the effectiveness of depth embedding (DE) and Depth Distribution Network Loss (DDN Loss). We take PETR [17] as baseline model . We find the performance is improved by 0.9 percent on mAP and 0.2 percent on NDS when we plug depth embeddings into the cross-attention, and the full model achieves the best performance increasing by **1.3 percent on mAP and 0.5 percent on NDS**.

D. False Positive Predictions Results

To verify whether our method can resolve the false positive problem, we consider Average Precision (AP). Tab. IV shows the AP of the pedestrian class with different distance thresholds on the validation set. Our method surpasses our baseline with a depth-pretrained backbone by **over 10 percent on average** among each threshold. Moreover, Fig. 4 illustrates our overall performance predominantly exceeds the baseline on all thresholds and thus resolves the false positive issue.

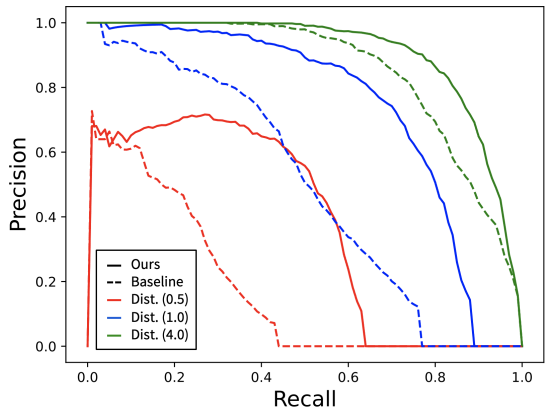

Fig. 4: **The precision-recall curve of pedestrian class.** Fig. 4 shows the comparison of AP between baseline (dotted lines) and our method (solid lines). The red, blue, and green colors represent the distance threshold at 0.5, 1.0, and 4.0 respectively. Regardless of distance, our method hugely outperforms the baseline on small object (e.g. pedestrian) detection.

Red, blue, and green represent the distance threshold at 0.5, 1.0, and 4.0 respectively.

E. Qualitative Results

Fig. 3 shows the qualitative result. Orange and blue bounding boxes represent ground truth and predictions respectively. As shown in Fig. 3, both PETR [17] with ResNet50 backbone [45] and PETR [17] with depth-pretrained VoVNetV2 [27]–[30] still predict a row of false positive predictions along the direction of depth for small objects. Since depth-pretrained backbones are generally pretrained on the external dataset and contain different settings on camera matrices, we suggest that those backbones can narrowly deal with the false positive problem due to weak depth estimation. Nevertheless, our method can predominately alleviate this problem due to referred depth information from the internal dataset [34].

V. Conclusion

In this paper, we design an end-to-end Cross-view and Depth-guided Transformer, called CrossDTR, for 3D object detection. To address the false positive bounding boxes commonly existing in prior multi-view approaches, a lightweight Depth Predictor, supervised by our produced object-wise sparse depth maps, is proposed to generate low-dimensional depth embeddings. Furthermore, to combine image and depth hints from different views, a Cross-view and Depth-guided Transformer is developed to fuse this information efficiently. We are optimistic that our proposed method would pave a new way for developing a cost-effective and real-time 3D object detector.

Acknowledgement

This work was supported in part by National Science and Technology Council, Taiwan, under Grant NSTC 111-2634-F-002-022, and Mobile Drive Technology Co., Ltd (MobileDrive). We are grateful to the National Center for High-performance Computing.

REFERENCES

[1] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[2] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[3] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793.

[4] Y. Chen, S. Liu, X. Shen, and J. Jia, "Dsgn: Deep stereo geometry network for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 536–12 545.

[5] P. Li, X. Chen, and S. Shen, "Stereo r-cnn based 3d object detection for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7644–7652.

[6] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3d object detection," *arXiv preprint arXiv:1811.08188*, 2018.

[7] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[8] X. Weng and K. Kitani, "Monocular 3d object detection with pseudo-lidar point cloud," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[9] Y. You, Y. Wang, W.-L. Chao, D. Garg, G. Pleiss, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving," *arXiv preprint arXiv:1906.06310*, 2019.

[10] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8555–8564.

[11] T. Wang, X. Zhu, J. Pang, and D. Lin, "Fcos3d: Fully convolutional one-stage monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 913–922.

[12] T. Wang, Z. Xinge, J. Pang, and D. Lin, "Probabilistic and geometric depth: Detecting objects in perspective," in *Conference on Robot Learning*. PMLR, 2022, pp. 1475–1485.

[13] 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 *Conference on Robot Learning*. PMLR, 2022, pp. 180–191.

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

[15] J. Huang and G. Huang, "Bevdet4d: Exploit temporal cues in multi-camera 3d object detection," *arXiv preprint arXiv:2203.17054*, 2022.

[16] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," *arXiv preprint arXiv:2203.17270*, 2022.

[17] Y. Liu, T. Wang, X. Zhang, and J. Sun, "Petr: Position embedding transformation for multi-view 3d object detection," *arXiv preprint arXiv:2203.05625*, 2022.

[18] Y. Liu, J. Yan, F. Jia, S. Li, Q. Gao, T. Wang, X. Zhang, and J. Sun, "Petrv2: A unified framework for 3d perception from multi-camera images," *arXiv preprint arXiv:2206.01256*, 2022.

[19] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao, "Graph-detr3d: Rethinking overlapping regions for multi-view 3d object detection," *arXiv preprint arXiv:2204.11582*, 2022.

[20] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu, "Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving," *arXiv preprint arXiv:2205.09743*, 2022.

[21] D. Rukhovich, A. Vorontsova, and A. Konushin, "Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2397–2406.

[22] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *European Conference on Computer Vision*. Springer, 2020, pp. 194–210.

[23] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, and A. Kendall, "Fiery: Future instance prediction in bird's-eye view from surround monocular cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 273–15 282.

[24] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 760–13 769.

[25] Y. Zhang, X. Ma, S. Yi, J. Hou, Z. Wang, W. Ouyang, and D. Xu, "Learning geometry-guided depth via projective modeling for monocular 3d object detection," *arXiv preprint arXiv:2107.13931*, 2021.

[26] M. Ding, Y. Huo, H. Yi, Z. Wang, J. Shi, Z. Lu, and P. Luo, "Learning depth-guided convolutions for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 1000–1001.

[27] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2485–2494.

[28] D. Park, R. Ambrus, V. Guizilini, J. Li, and A. Gaidon, "Is pseudo-lidar needed for monocular 3d object detection?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3142–3152.

[29] Y. Lee, J.-w. Hwang, S. Lee, Y. Bae, and J. Park, "An energy and gpu-computation efficient backbone network for real-time object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[30] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 906–13 915.

[31] R. Zhang, H. Qiu, T. Wang, X. Xu, Z. Guo, Y. Qiao, P. Gao, and H. Li, "Monodetr: Depth-aware transformer for monocular 3d object detection," *arXiv preprint arXiv:2203.13310*, 2022.

[32] K.-C. Huang, T.-H. Wu, H.-T. Su, and W. H. Hsu, "Monodtr: Monocular 3d object detection with depth-aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4012–4021.

[33] A. Kumar, G. Brazil, E. Corona, A. Parchami, and X. Liu, "Deviant: Depth equivariant network for monocular 3d object detection," *arXiv preprint arXiv:2207.10758*, 2022.

[34] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[35] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[36] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[37] Y. Wang, X. Zhang, T. Yang, and J. Sun, "Anchor detr: Query design for transformer-based detector," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2567–2575.

[38] C. Lu, M. J. G. van de Molengraft, and G. Dubbelman, "Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 445–452, 2019.

[39] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4867–4873, 2020.

[40] A. Saha, O. Mendez, C. Russell, and R. Bowden, "Translating images into maps," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 9200–9206.

[41] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng, "Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs," *arXiv preprint arXiv:2203.04050*, 2022.

[42] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[43] W. Chang, Y. Zhang, and Z. Xiong, "Transformer-based monocular depth estimation with attention supervision," 2021.

[44] J. Mao, S. Shi, X. Wang, and H. Li, "3d object detection for autonomous driving: A review and new outlooks," *arXiv preprint arXiv:2206.09474*, 2022.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[46] Y. Tang, S. Dorn, and C. Savani, "Center3d: Center-based monocular 3d object detection with joint depth understanding," in *DAGM German Conference on Pattern Recognition*. Springer, 2020, pp. 289–302.

[47] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[48] B. Zhu, Z. Jiang, X. Zhou, Z. Li, and G. Yu, "Class-balanced grouping and sampling for point cloud 3d object detection," *arXiv preprint arXiv:1908.09492*, 2019.

[49] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.