

# MSF3DDETR: Multi-Sensor Fusion 3D Detection Transformer for Autonomous Driving

Gopi Krishna Erabati\* and Helder Araujo

Institute of System and Robotics

University of Coimbra, Portugal

Email: {gopi.erabati, helder}@isr.uc.pt

**Abstract**—3D object detection is a significant task for autonomous driving. Recently with the progress of vision transformers, the 2D object detection problem is being treated with the set-to-set loss. Inspired by these approaches on 2D object detection and an approach for multi-view 3D object detection DETR3D, we propose MSF3DDETR: Multi-Sensor Fusion 3D Detection Transformer architecture to fuse image and LiDAR features to improve the detection accuracy. Our end-to-end single-stage, anchor-free and NMS-free network takes in multi-view images and LiDAR point clouds and predicts 3D bounding boxes. Firstly, we link the object queries learnt from data to the image and LiDAR features using a novel MSF3DDETR cross-attention block. Secondly, the object queries interacts with each other in multi-head self-attention block. Finally, MSF3DDETR block is repeated for  $L$  number of times to refine the object queries. The MSF3DDETR network is trained end-to-end on the nuScenes dataset using Hungarian algorithm based bipartite matching and set-to-set loss inspired by DETR. We present both quantitative and qualitative results which are competitive to the state-of-the-art approaches.

## I. INTRODUCTION

3D object detection is one of the fundamental and the most crucial element of the visual perception system of autonomous vehicles. Object detection is a twofold problem of classifying and localizing the objects in the scene. A 3D LiDAR point cloud provides the depth information of the scene but the point cloud data is very sparse at long range. On the other hand, a RGB sensor provides dense pixel data of the scene but it does not perceive the depth and it is very sensitive to illumination changes in the scene. Therefore, high-resolution RGB and 3D LiDAR data can be fused to leverage their complementary nature.

The well researched 2D CNN based object detection techniques [1], [2] are not well suited for 3D LiDAR point clouds due to their sparse and unstructured nature. In order to tackle this challenge, researchers employed various approaches such as voxelization or discretization of point clouds. Several approaches [3], [4] discretize the point cloud into 3D voxels to apply regular 3D CNNs, but they are computationally expensive. Other approaches [5], [6] uses spherical or cylindrical projection to obtain Birds-Eye View (BEV) representation of point clouds and apply 2D CNNs. In our method we test two approaches: an efficient BEV representation of 3D LiDAR point clouds obtained using PointPillars [7], and SparseConv [8] to obtain 3D sparse convolutions on the voxels.

\*Corresponding author

We propose Multi-Sensor Fusion 3D DEtection TRansformer (MSF3DDETR) a single-stage, anchor-free and post-processing (NMS) free network for 3D object detection leveraging the fusion of multi-view RGB and LiDAR data. The multi-view RGB features are extracted by a shared ResNet backbone [9] and LiDAR features are extracted by SECOND backbone [4] from the VoxelNet [10] or PointPillars [7]. We introduce a novel MSF3DDETR cross-attention block which fuses the RGB and LiDAR features by the attention mechanism using the sparse set of object queries which are learnt end-to-end. To our knowledge, this is the first approach to fuse the RGB and LiDAR features by attention mechanism leveraging the learnt object queries. Similar to [11], the fused features interact with each other through a multi-head self-attention layer [12]. The 3D bounding box parameters are regressed from every layer of MSF3DDETR block and we use bipartite matching with a set-to-set loss inspired by DETR [13] to optimize the model during training.

We train our network on the publicly available autonomous driving nuScenes dataset [14] and obtained competitive results (without NMS) to the state-of-the-art methods (with NMS).

Our main contributions are as follows:

- We propose an end-to-end, single-stage, anchor-free and NMS-free MSF3DDETR network to detect 3D objects. Similar to DETR [13], our method does not require any post-processing such as NMS, however it obtains competitive results with existing NMS-based approaches. We also employ knowledge distillation with teacher and student model to improve the accuracy.
- Our approach fuses the multi-view RGB and LiDAR features leveraging the learnt object queries using the cross-attention mechanism. To our knowledge, this is the first attempt to fuse multi-view RGB and LiDAR features using a novel cross-attention block at the object query level and posing multi-sensor fusion as a 3D set-to-set prediction.
- We release our code and models to facilitate further research.

## II. RELATED WORK

**CNN-based object detection** The CNN-based methods span from two-stage approaches to anchor-free approaches. RCNN [15] and its variants Fast RCNN [16] and Faster RCNN [17] are two-stage 2D object detection methods which are

typically slow in practice. Point RCNN [18] and Faster Point RCNN [19] are two-stage 3D object detection approaches which aggregate RoI specific features for object proposals using PointNet set abstraction layer [20]. Region proposal network [17] is removed in the single-stage approaches [1], [2], thus making the detection faster than two-stage approaches. VoxelNet [10] is a single-stage 3D object detection approach which uses 3D convolutions, which are bottleneck in computation efficiency. Some approaches such as PIXOR [21] and PointPillars [7] project the point cloud to BEV representation and use 2D CNNs to predict 3D bounding boxes. However, these methods contain the anchor boxes which are statistically obtained from the dataset. Few approaches like FCOS [22], CenterNet [23] and CornerNet [24] shift from per-anchor detection to per-pixel or per-keypoint 2D object detection. FCOS3D [25] is a monocular 3D object detection approach which employs per-pixel prediction. PillarOD [26] and CenterPoint [27] are anchor-free 3D object detection approaches which employ per-pillar (BEV pixel) prediction. However, these methods still predict redundant bounding boxes which are removed by post-processing methods, such as NMS.

**Transformer-based object detection** To remove the post-processing (like NMS), DETR [13] defines the object detection problem as a set-to-set problem. It uses bipartite matching using Hungarian algorithm to match the predictions with the ground-truth boxes, thus removing the post-processing to remove redundant boxes. To accelerate the training Deformable DETR [28] employs deformable self-attention. DETR3D [11] and Object DGCNN [29] are 3D object detection approaches based on set-to-set prediction problem. Our method also defines 3D object detection as a set-to-set prediction problem, but with a novel cross-attention block to fuse RGB and LiDAR features.

**Fusion-based 3D object detection** approaches fuse data of 3D LiDAR and RGB camera. MV3D [5] and AVOD [6] fuses the projected LiDAR and RGB data using a two-stage CNN approach to predict the 3D bounding boxes. The objects are fused at the proposal stage using RoIPooling. Frustum PointNet [30] uses the 2D object detection on RGB images to narrow the search space in the 3D scene. It improves efficiency but limited by the accuracy of 2D object detection. The LiDAR and RGB features are shared across the backbones in Deep Continuous Fusion [31]. MVXNet [32] and PointPainting [33] fuse the points with the corresponding image features. However, all these approaches use CNNs to fuse the LiDAR and RGB features. In our method, we fuse the RGB and LiDAR features in a novel MSF3DDETR cross-attention block using attention mechanism leveraged by learnt object queries.

### III. MSF3DDETR NETWORK

Our end-to-end 3D object detection network (MSF3DDETR) inputs multi-view RGB and LiDAR data, and outputs 3D bounding boxes for different objects in a scene.

Our architecture comprises of three main components as shown in Figure 1. Initially, we extract RGB features from multi-view RGB images using a shared ResNet [9], and BEV

LiDAR features using VoxelNet [10] or PointPillars [7] and SECOND [4] backbone networks. The RGB and BEV LiDAR features are enhanced using a Feature Pyramid Network (FPN) [34]. Secondly, a novel transformer head is designed to not only link the computed RGB and LiDAR features with 3D object predictions inspired by [11] but also to efficiently fuse the RGB and LiDAR features by cross-attention mechanism [12] leveraging the learnt object queries in the MSF3DDETR cross-attention block. The transformer head starts with object queries which are transformed into 3D object locations, which are projected to the RGB and BEV planes to sample the RGB and BEV LiDAR features via bilinear interpolation. The sampled RGB and BEV LiDAR features are fused with attention weights learnt from object queries. Multi-head self-attention block incorporates fused feature interactions and thereby refine object queries. We repeat the MSF3DDETR block multiple times with alternating self-attention, cross-attention and MLP blocks respectively. Finally, we transform the object query features into 3D object parameters and train the network end-to-end with bipartite matching and set-to-set loss [13].

#### A. CNN Backbone and Neck

Our network inputs multi-view RGB images  $\mathcal{I} = \{I_1, \dots, I_6\} \subset \mathbb{R}^{H \times W \times 3}$  and LiDAR point cloud  $\mathcal{P} \subset \mathbb{R}^{P \times 3}$ . We extract multi-scale, multi-view RGB features using a shared ResNet101 [9] backbone with deformable convolutions [35]. The features are further enhanced by FPN to obtain multi-scale features  $\mathcal{F}_1^{img}, \mathcal{F}_2^{img}, \mathcal{F}_3^{img}, \mathcal{F}_4^{img}$ , where  $\mathcal{F}_x^{img} = \{f_x^{img1}, \dots, f_x^{img6}\} \subset \mathbb{R}^{h \times w \times c}$ . To encode the LiDAR point clouds, we use VoxelNet [10] with 0.1m voxel size or to accelerate object detection for point clouds, we scatter point clouds into BEV pillars using PointPillars [7] which maps sparse point clouds into dense BEV pillar map. We use SECOND [4] to extract sparse voxel or BEV pillar features and further enhance them using FPN to obtain four sets of features  $\mathcal{F}_1^{bev}, \mathcal{F}_2^{bev}, \mathcal{F}_3^{bev}, \mathcal{F}_4^{bev}$ . These features are passed to MSF3DDETR cross-attention block to link with object queries in the transformer head.

#### B. Transformer Head

Earlier approaches such as [6], [25], [26] employs dense set of anchor boxes or produces dense per pixel predictions which results in redundant bounding boxes removed by post-processing methods like NMS. However, NMS-based redundancy removal is non-parallelizable and amounts to increase of inference time. We tackle this issue using a transformer head which uses  $L$  layers of MSF3DDETR block. The MSF3DDETR Block alternates between multi-head self-attention block, MSF3DDETR cross-attention block and MLP block with layer norms and skip connections.

**MSF3DDETR cross-attention** The details are illustrated in Figure 2. Inspired by DETR [13], the transformer head starts with a set of object queries  $\mathcal{Q} = \{q_1, \dots, q_{N_q}\} \subset \mathbb{R}^d$ . The

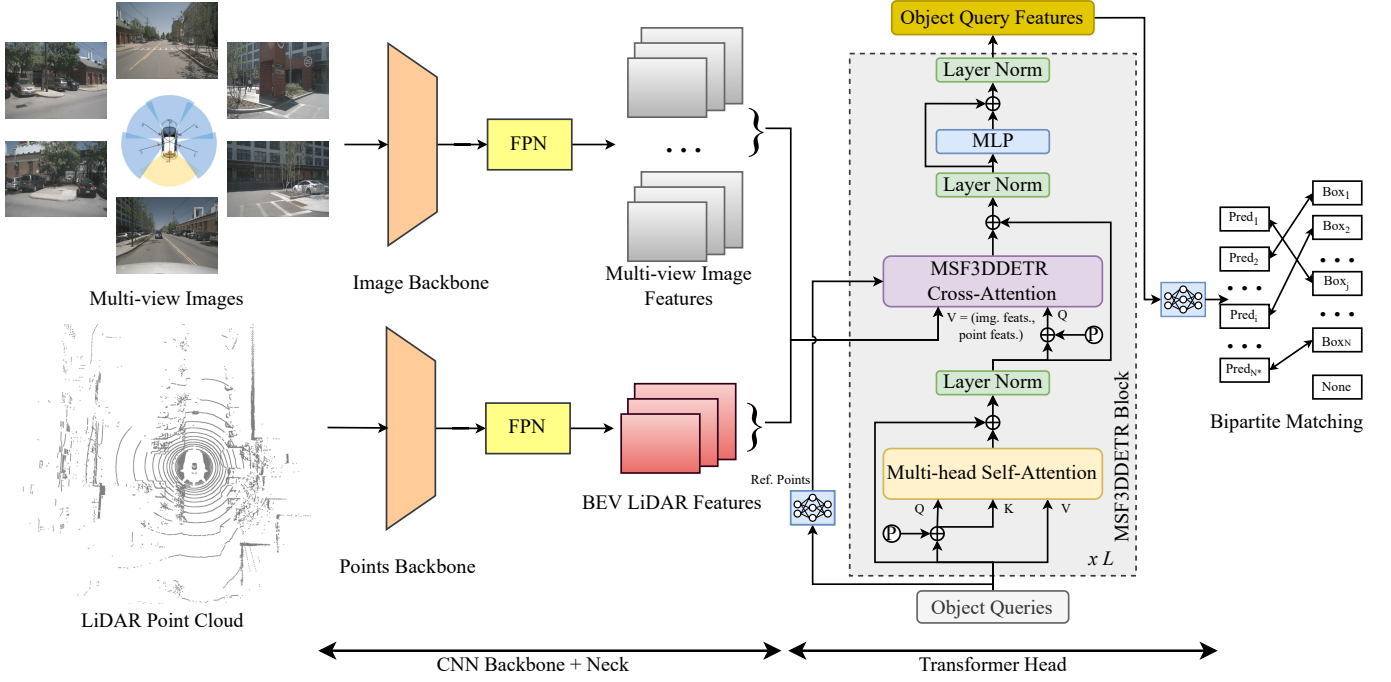


Fig. 1: An overview of MSF3DDETR architecture. It is an end-to-end network which inputs multi-view RGB images and LiDAR point cloud to predict 3D bounding boxes. CNN backbone extracts image and BEV LiDAR features which are linked to object queries in the transformer head. Specifically, the novel MSF3DDETR cross-attention block fuses image and LiDAR features by attention mechanism leveraging the learnt object queries.

queries are transformed to a set of reference points  $\mathcal{R} \subset \mathbb{R}^3$  as:

$$\mathcal{R} = \phi(Q) \quad (1)$$

where  $\phi$  is a single layer fully connected (FC) network. The reference points are projected to multi-view image planes using sensor transformation matrices to obtain image projected 3D reference points  $\mathcal{R}_{lv}^{img}$  for each layer  $l$  and camera view  $v$ . Simultaneously the reference points are also projected to the BEV plane to obtain BEV projected 3D reference points  $\mathcal{R}_l^{pts}$ .

The image and BEV features are collected by

$$\begin{aligned} \mathbf{f}_{lxv}^{img} &= f^{bilinear}(\mathcal{F}_{xv}^{img}, \mathcal{R}_{lv}^{img}) \\ \mathbf{f}_{lx}^{bev} &= f^{bilinear}(\mathcal{F}_x^{bev}, \mathcal{R}_l^{pts}) \end{aligned} \quad (2)$$

where  $\mathbf{f}_{lxv}^{img}$  is the image features for  $x$ -th level of  $v$ -th camera view at  $l$ -th layer and  $\mathbf{f}_{lx}^{bev}$  is the BEV features for  $x$ -th level at  $l$ -th layer.

A neural network is applied on object queries to obtain attention weights for cross-attention of sampled image and BEV features. We further obtain cross-attention image and BEV features which are concatenated and further a neural network is applied to obtain fused image and BEV features as shown in Figure 2.

The fused features interact with each other in the multi-head self-attention block where the queries are refined iteratively. Finally, we predict bounding box  $\hat{B}_l$  and its class  $\hat{C}_l$  from

the refined object queries using a two layer FC networks  $\phi_l^{reg}$  and  $\phi_l^{cls}$ , where  $\hat{B}_l = \{\hat{b}_{l1}, \dots, \hat{b}_{lj}, \dots, \hat{b}_{lN_q}\} \subset \mathbb{R}^9$  and  $\hat{C}_l = \{\hat{c}_{l1}, \dots, \hat{c}_{lj}, \dots, \hat{c}_{lN_q}\} \subset \mathbb{Z}$ . We compute the loss from the predictions for every layer during training but we only use outputs from last layer during inference.

### C. Loss

Inspired by [13], we use bipartite matching with hungarian algorithm and set-to-set loss to calculate the error between predicted set  $(\hat{B}, \hat{C})$  and ground-truth set  $(B, C) = (\{b_1, \dots, b_j, \dots, b_M\}, \{c_1, \dots, c_j, \dots, c_M\})$ . The number of ground-truth boxes  $M$  is smaller than number of predicted boxes  $N_q$ , so we pad the ground-truth boxes with  $\emptyset$ s (no object) up to  $M$ . The one-to-one matching between prediction and ground-truth is established with the Hungarian algorithm [36],  $\sigma^* = \arg \min_{\sigma \in \mathcal{P}} \sum_{i=1}^M -\mathbb{1}_{\{c_j \neq \emptyset\}} \hat{p}_{\sigma(i)}(c_i) + \mathbb{1}_{\{c_i = \emptyset\}} \mathcal{L}_{box}(b_i, \hat{b}_{\sigma(i)})$ , where  $\mathcal{L}_{box}$  is the  $L_1$  loss for bounding box regression and focal loss [37] for class labels. Finally, the set-to-set loss is defined as  $\mathcal{L} = \sum_{i=1}^N -\log \hat{p}_{\sigma^*(i)}(c_i) + \mathbb{1}_{\{c_i = \emptyset\}} \mathcal{L}_{box}(b_i, \hat{b}_{\sigma^*(i)})$ .

## IV. EXPERIMENTS

### A. Dataset

We evaluate our network on large-scale publicly available nuScenes dataset [14]. The dataset consists of 1K sequences of  $\sim 20$ s duration which are split into 700, 150, 150 sequences for training, validation and testing respectively. The

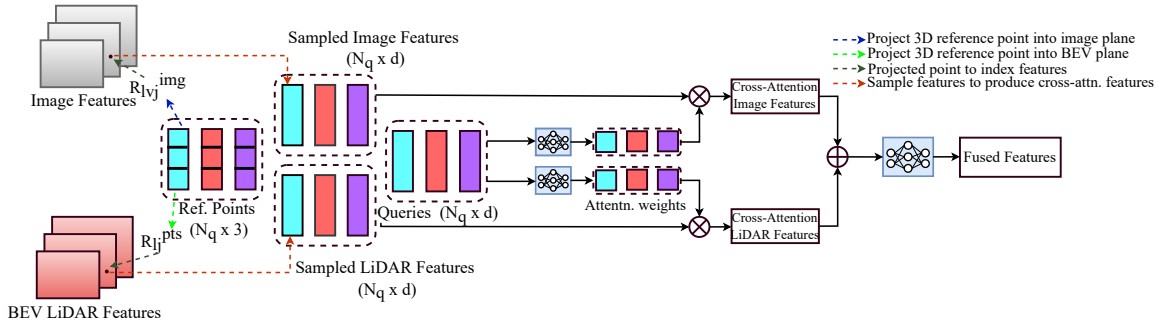


Fig. 2: An overview of MSF3DDETR cross-attention block

sequences are annotated at 2Hz, leading to 28K, 6K, 6K annotated samples for training, validation and testing respectively. Each sample consists of RGB images from 6 cameras [front, front\_right, front\_left, back, back\_left, back\_right] and 32-beam LiDAR point cloud with 30K points per sample. 10 categories are available to compute the metrics.

### B. Metrics

We evaluate our model following the official evaluation protocol defined by nuScenes [14]. We evaluate the True Positive (TP) metrics, such as average translation error (ATE), average scale error (ASE), average orientation error (AOE), average velocity error (AVE), average attribute error (AAE), computed in the physical unit. In addition to TP metrics, we also evaluate the main metrics mean average precision (mAP) and the consolidated nuScenes detection score (NDS).

### C. Network Architecture

Our model inputs RGB images from 6 camera views and extract RGB features using a shared ResNet101 [9] backbone and FPN [34] and thus it produces feature maps of strides 4, 8, 16, 32. We scale the RGB images to 0.8 of input dimensions ( $1600 \times 900$ ) which achieves a good trade-off between accuracy and efficiency. The LiDAR point cloud features are extracted using SparseConv [8] to obtain 3D sparse convolutions that refine voxel features. The features are further extracted using a backbone having [5, 5] convolutional layers with dimensions [128, 256] respectively. The BEV feature maps are combined using a FPN and further a convolutional layer is added at the end to produce four BEV feature maps. The transformer head consists of 6 (layers) MSF3DDETR blocks, where each block has multi-head self-attention, MSF3DDETR cross-attention and MLP with hidden dimension 256. The number of object queries is set to 900.

### D. Training Details

We train our network using AdamW optimizer with an initial learning rate of  $2 \times 10^{-4}$ . The weight decay is  $10^{-2}$ . We use Cosine Annealing as learning rate scheduler with a linear warmup until 2K iterations and with minimum learning rate of  $2 \times 10^{-7}$ . The RGB backbone is initialized with a pretrained FCOS3D [25] network and BEV backbone is initialized with

a pretrained VoxelNet [10] network on the same dataset. We train for 24 epochs on two RTX 3090 GPUs with a batch size of 1. We take the top 300 objects with highest category score as final predictions and we do not use any post-processing method such as NMS during inference.

### E. Results

1) *Quantitative Results:* We compare the existing state-of-the-art methods on the nuScenes [14] dataset as shown in Table I. Our network surpassed the camera-only CNN and transformer based approaches FCOS3D [25] and DETR3D [11] respectively, by approximately 74% (0.26 mAP), proving our hypothesis that fusion of data improves the accuracy. Our method achieve 4.4% and 8.6% improvement in mAP over LiDAR based approaches Object DGCNN [29] and CenterPoint [27] respectively. The data fusion approaches like CenterFusion [38] combines camera and radar data to produce dense per-pixel predictions, PointPainting [33] annotates each LiDAR point with the CNN features from the camera images and it employs NMS to remove redundant boxes in each view and in overlap regions, and MVP [39] employs 2D segmentation and augments virtual points based on 2D segmentation. The result of our camera and LiDAR data fusion approach shows significant improvement in terms of mAP and NDS compared to CenterFusion and PointPainting without bells and whistles. This observation verifies the significance of our cross-attention based fusion approach. Our method obtains 0.051 less mAP than MVP, but the MVP is a two-stage approach where the prediction result depends on the accuracy of 2D segmentation and it employs NMS to remove redundant boxes which results in significant inference overhead. Moreover, MVP predicts 3D objects for individual multi-view images, whereas our approach predicts the objects in a single-stage by processing multi-view images all at once without NMS. Our network achieve close-to-real-time inference speed (6.3 FPS) on a desktop GPU (RTX 3090), which is approximately 18% improvement in inference speed over MVP.

We employ set-to-set knowledge distillation (KD) with a teacher and student model similar to [29] to improve the accuracy. As our approach is NMS-free, we can easily distill the information from teacher to student with homogeneous

TABLE I: Quantitative results comparison of recent works on nuScenes [14] validation dataset. Modality - C: Camera, L: LiDAR, R: Radar

Method	Modality	NDS $\uparrow$	mAP $\uparrow$	mATE $\downarrow$	mASE $\downarrow$	mAOE $\downarrow$	mAVE $\downarrow$	mAAE $\downarrow$	NMS
FCOS3D [25]	C	0.415	0.343	0.725	0.263	0.422	1.292	0.153	$\checkmark$
DETR3D [11]	C	0.434	0.349	0.716	0.268	0.379	0.842	0.200	$\times$
CenterPoint (voxel) [27]	L	0.648	0.564	-	-	-	-	-	$\checkmark$
Object DGCNN (voxel) [29]	L	0.661	0.587	0.333	0.263	0.288	0.251	0.190	$\times$
CenterFusion [38]	C+R	0.449	0.326	0.631	0.261	0.516	0.614	<b>0.115</b>	$\checkmark$
PointPainting [33]	C+L	0.581	0.464	-	-	-	-	-	$\checkmark$
MVP [39]	C+L	<b>0.705</b>	<b>0.664</b>	-	-	-	-	-	$\checkmark$
MSF3DDETR (ours)	C+L	0.667	0.606	0.334	0.258	0.288	0.283	0.193	$\times$
MSF3DDETR (ours w/ distillation)	C+L	0.672	0.613	<b>0.333</b>	<b>0.256</b>	<b>0.287</b>	<b>0.281</b>	0.191	$\times$

TABLE II: Average Precision (AP) by object category. CV - Construction Vehicle, Motor - Motorcycle, Ped - Pedestrian, TC - Traffic Cone

Modality	Car	Truck	Trailer	Bus	CV	Bicycle	Motor	Ped	TC	Barrier
Camera-only [11]	0.54	0.28	0.16	0.34	0.08	0.27	0.34	0.42	0.52	0.47
LiDAR-only [29]	0.84	0.54	0.40	0.66	0.20	0.44	0.66	0.81	0.64	0.62
Camera + LiDAR (Ours)	0.86	0.58 $\uparrow$ 0.04	0.40	0.71 $\uparrow$ 0.05	0.21	0.53 $\uparrow$ 0.09	0.67	0.83	0.67	0.61

detection heads. We achieve an increase of 0.007 mAP by KD as shown in Table I.

2) *Qualitative Results*: The qualitative results of our MSF3DDETR architecture are illustrated in Figure 3. The 3D bounding boxes are projected on to the BEV and 6 camera view planes. The MSF3DDETR network was able to detect even the small objects and the objects which are not annotated in the ground-truth, for instance in front of the ego vehicle (right-side of LIDAR\_BEV\_PRED) and also front towards the right (bottom-right of LIDAR\_BEV\_PRED) a car and few persons were not annotated in ground-truth but our network was able to detect them as shown in CAM\_FRONT and CAM\_FRONT\_RIGHT respectively. Some failure cases include as some barriers in CAM\_FRONT\_LEFT were not detected. A short demo video of the 3D object predictions of MSF3DDETR architecture can be watched at <https://youtu.be/F0rxAPI-a24>.

3) *Analysis*: We analyse the performance of our approach by object categories, object distances and object sizes in comparison to single modality approaches. We report the performance of our approach for different number of layers employed in the transformer detection head.

**Object category.** We present the Average Precision (AP) of camera-only, LiDAR-only and our fusion based approach for every object category in Table II. LiDAR-only model [29] significantly outperform camera-only model [11] in all the object categories. The camera data helps the LiDAR-only model as in our approach which significantly improves performance on bicycle, bus, motorcycle categories.

**Object distance.** The ground-truth bounding boxes are split into three subsets basing on distance of object centers to ego vehicle: [0m, 20m], [20m, 30m] and [30m,  $+\infty$ ]. Table III shows the mAP of different modalities by object distance. The objects that are near to ego vehicle (up to 20 meters distance) there is not much improvement by fusing camera

data to LiDAR model, but the performance on objects that are far from ego vehicle (above 30 meters distance) significantly improves by equipping camera data to LiDAR model. This is due to the fact that LiDAR data gets very sparse as the distance from ego vehicle increases.

TABLE III: mAP by object distance

Modality	[0m, 20m]	[20m, 30m]	[30m, $+\infty$ ]
Camera-only [11]	0.479	0.257	0.103
LiDAR-only [29]	0.732	0.555	0.303
Camera + LiDAR (Ours)	0.738	0.575 $\uparrow$ 0.02	0.333 $\uparrow$ 0.03

**Object Size.** The ground-truth bounding boxes are split into two subsets based on longer edge of the 3D bounding box: [0m, 4m] and [4m,  $+\infty$ ]. Table IV shows the mAP of different modalities by object size. The performance of our approach on smaller objects (less than 4 meters) is significantly better than larger objects (more than 4 meters). This is due to the fact that camera data has high resolution than LiDAR data which helps in detecting the smaller objects.

TABLE IV: mAP by object size

Modality	[0m, 4m]	[4m, $+\infty$ ]
Camera-only [11]	0.227	0.137
LiDAR-only [29]	0.360	0.254
Camera + LiDAR (Ours)	0.376 $\uparrow$ 0.016	0.263 $\uparrow$ 0.009

**Detection Layers** The quantitative performance of our approach for different layers in the detection head is shown in Table V. The performance of our approach significantly improves as the number of layers in detection head increases, because the object queries are iteratively refined in every layer of the detection head. However, after 6 layers of MSF3DDETR block in detection head there is not much improvement in the performance of our network, so we fix to 6 layers of

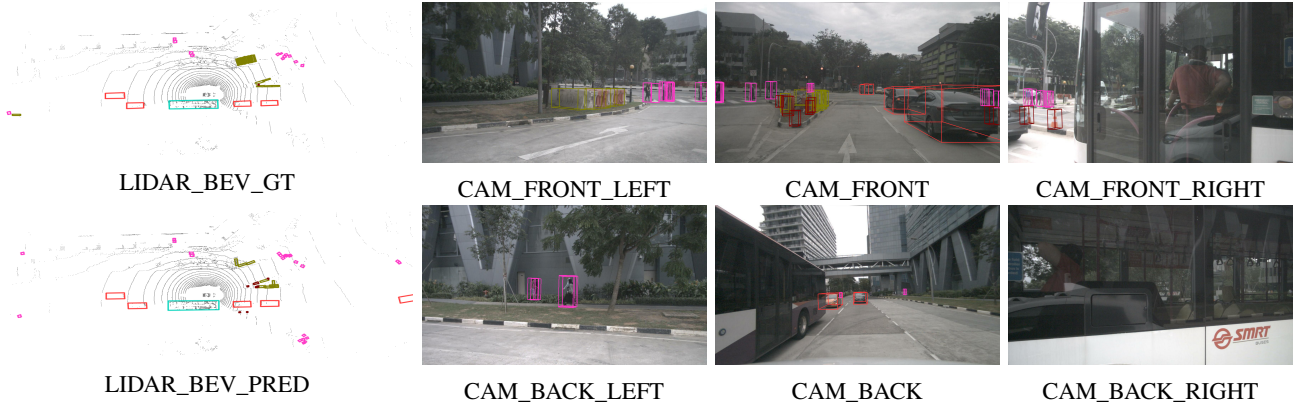


Fig. 3: Qualitative results of MSF3DDETR network both in BEV plane and image planes. Different objects are shown in different colors. Image best viewed in color and zoom.

TABLE V: Performance by different layers in detection head

Layer	NDS $\uparrow$	mAP $\uparrow$	mATE $\downarrow$	mASE $\downarrow$	mAOE $\downarrow$	mAVE $\downarrow$	mAAE $\downarrow$
0	0.607	0.517	0.409	0.273	0.336	0.305	0.187
1	0.640	0.561	0.359	0.264	0.302	0.289	0.188
2	0.658	0.593	0.340	0.263	0.293	0.287	0.193
3	0.665	0.603	0.336	0.262	0.286	0.285	0.195
4	0.667	0.605	0.335	0.260	0.284	0.285	0.188
5	0.667	0.606	0.334	0.258	0.288	0.283	0.193

MSF3DDETR block in the detection head as trade-off between performance and efficiency.

#### F. Ablation studies

We report the performance of MSF3DDETR using different number of queries. We show the results for pillar [7] and voxel [10] backbones for LiDAR data and ResNet101 [9] backbone for camera data. As shown in Table VI, the number of queries in the transformer detection head has slight impact on the performance of our approach. However, we use 900 queries in other experiments.

TABLE VI: Ablation study on the number of object queries

No. of queries	MSF3DDETR (pillar)		MSF3DDETR (voxel)	
	mAP	NDS	mAP	NDS
300	0.524	0.585	0.587	0.641
600	0.536	0.601	0.595	0.652
900	<b>0.545</b>	<b>0.614</b>	<b>0.606</b>	<b>0.667</b>
1000	0.538	0.605	0.591	0.648

We report the performance of MSF3DDETR using different camera-LiDAR backbone combinations as shown in Table VII. We test the performance of MSF3DDETR with ResNet50 and ResNet101 for camera backbones, and PointPillars [7] with 0.2m pillar size and VoxelNet [10] with 0.1m voxel size for LiDAR backbones. We can observe from the Table VII that the deep convolutional network ResNet101 with more powerful LiDAR backbone VoxelNet achieves the best performance. However, our approach is more flexible to plugin various backbones for camera and LiDAR data depending on the trade-

off between accuracy and efficiency required for the specific application.

TABLE VII: Ablation study on the backbones

Camera	LiDAR	NDS	mAP
ResNet50	PointPillar 0.2m	0.531	0.596
ResNet50	VoxelNet 0.1m	0.595	0.660
ResNet101	PointPillar 0.2m	0.545	0.614
ResNet101	VoxelNet 0.1m	<b>0.606</b>	<b>0.667</b>

## V. CONCLUSION

We present MSF3DDETR a single-stage, anchor-free and NMS-free network which inputs multi-view RGB images and LiDAR point clouds to predict 3D bounding boxes. The novel MSF3DDETR cross-attention block links the object queries learnt from the data to RGB and LiDAR features, and thereby helps to fuse the data at object query level, which to the best of our knowledge is the first of its kind. We train on nuScenes dataset and present quantitative and qualitative competitive results to other state-of-the-art approaches. We also employ set-to-set knowledge distillation by teacher and student model to improve the accuracy.

## ACKNOWLEDGMENT

This work has been supported by European Union's H2020 MSCA-ITN-ACHIEVE with grant agreement No. 765866, Fundação para a Ciência e a Tecnologia (FCT) under the project UIDB/00048/2020 and FCT Portugal PhD research grant with reference 2021.06219.BD.

## REFERENCES

- [1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37. 1, 2
- [2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788. 1, 2
- [3] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 922–928. 1
- [4] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018. 1, 2
- [5] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915. 1, 2
- [6] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8. 1, 2
- [7] 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. 1, 2, 6
- [8] B. Graham, M. Engelcke, and L. Van Der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9224–9232. 1, 4
- [9] 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. 1, 2, 4, 6
- [10] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499. 1, 2, 4, 6
- [11] 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. 1, 2, 4, 5
- [12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 1, 2
- [13] 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. 1, 2, 3
- [14] 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. 1, 3, 4, 5
- [15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587. 1
- [16] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448. 1
- [17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015. 1, 2
- [18] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 770–779. 2
- [19] Y. Chen, S. Liu, X. Shen, and J. Jia, "Fast point r-cnn," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9775–9784. 2
- [20] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017. 2
- [21] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3d object detection from point clouds," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7652–7660. 2
- [22] 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. 2
- [23] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6569–6578. 2
- [24] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 734–750. 2
- [25] 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. 2, 4, 5
- [26] Y. Wang, A. Fathi, A. Kundu, D. A. Ross, C. Pantofaru, T. Funkhouser, and J. Solomon, "Pillar-based object detection for autonomous driving," in *European Conference on Computer Vision*. Springer, 2020, pp. 18–34. 2
- [27] 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. 2, 4, 5
- [28] 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. 2
- [29] Y. Wang and J. M. Solomon, "Object dgcnn: 3d object detection using dynamic graphs," *Advances in Neural Information Processing Systems*, vol. 34, 2021. 2, 4, 5
- [30] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 918–927. 2
- [31] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 641–656. 2
- [32] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvx-net: Multimodal voxelnet for 3d object detection," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7276–7282. 2
- [33] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612. 2, 4, 5
- [34] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125. 2, 4
- [35] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773. 2
- [36] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955. 3
- [37] 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. 3
- [38] R. Nabati and H. Qi, "Centerfusion: Center-based radar and camera fusion for 3d object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1527–1536. 4, 5
- [39] T. Yin, X. Zhou, and P. Krähenbühl, "Multimodal virtual point 3d detection," *Advances in Neural Information Processing Systems*, vol. 34, 2021. 4, 5