Improving Distant 3D Object Detection Using 2D Box Supervision

Zetong Yang[1*] Zhiding Yu[2†] Chris Choy[2] Renhao Wang[3*] Anima Anandkumar[4*] Jose M. Alvarez[2]

[1]CUHK [2]NVIDIA [3]UC Berkeley [4]Caltech

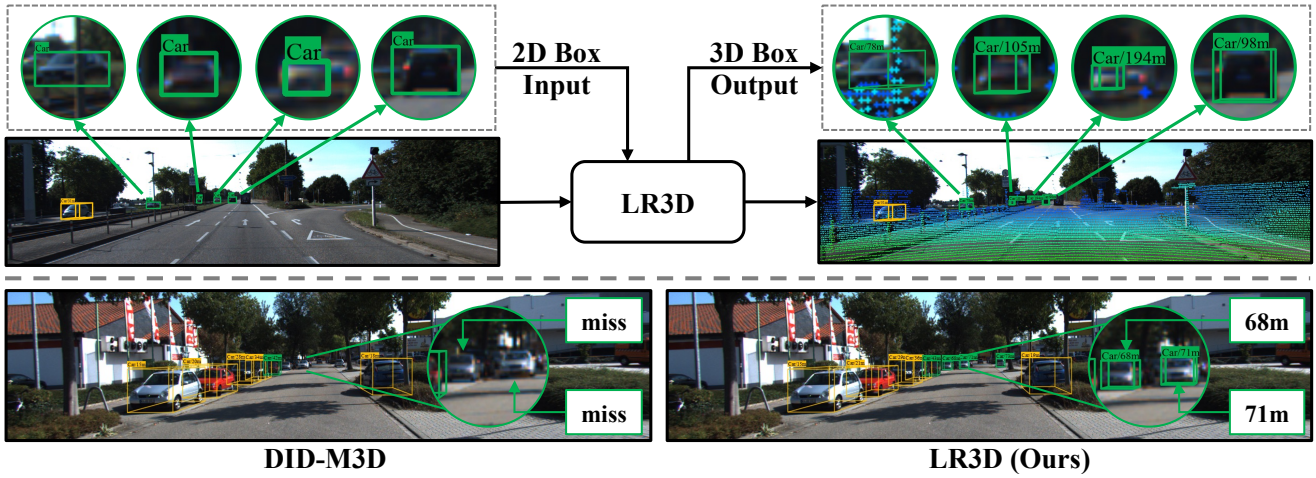

Figure 1. **Upper:** Without 3D annotations beyond 40m, LR3D enables the predictions of 3D boxes for extremely distant objects over 200m (right, in **green**) based on the inputs of image and 2D box (left). **Lower:** Existing methods fail to detect 3D objects beyond the 3D supervision range, *e.g.*, 40m for this case. With LR3D, these remote missing objects are well detected.

Abstract

Improving the detection of distant 3d objects is an important yet challenging task. For camera-based 3D perception, the annotation of 3d bounding relies heavily on LiDAR for accurate depth information. As such, the distance of annotation is often limited due to the sparsity of LiDAR points on distant objects, which hampers the capability of existing detectors for long-range scenarios. We address this challenge by considering only 2D box supervision for distant objects since they are easy to annotate. We propose LR3D, a framework that learns to recover the missing depth of distant objects. LR3D adopts an implicit projection head to learn the generation of mapping between 2D boxes and depth using the 3D supervision on close objects. This mapping allows the depth estimation of distant objects conditioned on their 2D boxes, making long-range 3D detection with 2D supervision feasible. Experiments show that without distant 3D annotations, LR3D allows camera-based methods to detect distant objects (over 200m) with comparable accuracy to full 3D supervision. Our framework is general, and could widely benefit 3D detection methods to a large extent.

*Work done during internship / affiliation with NVIDIA.
†Corresponding author: zhidingy@nvidia.com

1. Introduction

Camera-based 3D object detection [1–7] is an important task in autonomous driving aiming to localize and classify objects in 3D space with monocular or multi-view image input. Detecting distant objects with camera input is both important and challenging. On one hand, the ability to detect objects at distance is needed for planning, especially for highway scenarios. According to [8], at 60 miles/hour, the typical stopping distance for a vehicle is 73 meters and it grows significantly under harsh road conditions such as having wet surface [9–11]. On the other hand, the detection range of camera-based methods depends heavily on the distance range of 3D annotations. For example, these methods work well within the distance range with abundant 3D annotations (*e.g.*, ∼70 meters on KITTI [12] and nuScenes [13]), but often fail beyond this range where 3D annotations are missing. This indicates the importance to improve distant 3D object detection with longer 3D annotation range.

Though of great importance, increasing the range of 3D annotation is not easy. One main challenge is the sparsity of LiDAR points at longer distances. LiDAR is typically the main reference for axial depth information in 3D annotation. For distant objects with few or even no LiDAR points,

labeling their 3D bounding boxes is an ill-posed, noisy, and time-consuming task for human annotators. For this reason, distant 3D box annotations are rare, and most datasets [12–14] provide 3D annotations around 70 meters. A few datasets [15, 16] provide far-away annotations. They obtain long-range 3D annotations either by using costly LiDAR sensors with very long sensing ranges or by taking significant efforts in annotating 3D boxes with consistency among sequences of stereo images. Both of these strategies are more expensive and less scalable, compared to drawing precise 2D boxes on images. Thus, in this paper, we explore an efficient camera-based approach to achieve high-quality long-range 3D object detection, which uses only 2D annotations for long-range objects of which the 3D annotations are hardly available due to sparse or no interior LiDAR points.

Our approach: We present LR3D, a camera-based detection framework that detects distant 3D objects using only 2D bounding box annotations. The core of LR3D is an Implicit Projection Head (IP-Head) design that can be plugged into existing camera-based detectors and enables them to effectively predict 3D bounding boxes at all ranges, using the supervision from close 2D/3D and distant 2D box labels.

IP-Head learns to generate specific mapping for each instance from their 2D bounding box prediction to the corresponding depth, and thus is capable of estimating depth for distant objects relying on 2D supervision only. We also design a projection augmentation strategy to force the generated implicit function to correctly model the mapping between 2D boxes and 3D depth of target instances and estimate different depth outputs if the 2D box input changes.

We also notice issues in evaluating camera-based detectors for long-range 3D detection. Existing evaluation metrics [12, 13], based on mean average precision (mAP) with fixed thresholds, neglect the inaccuracy of depth estimation of distant objects from cameras and lead to meaningless numbers even for state-of-the-art methods, *e.g.*, only 0.43% mAP of DID-M3D [6] on KITTI Dataset for objects farther than 40m. This motivates us to design a novel metric, Long-range Detection Score (LDS), which sets a dynamic threshold for judging TP prediction regarding associated ground truth depth, for informative quatitative comparison.

We conduct experiments on five popular 3D detection datasets including KITTI [12], nuScenes [13], and Waymo [14] for objects within 80m, as well as Cityscapes3D [16] and Argoverse 2 [15] for objects farther than 200m. We remove the 3D annotations of farther objects in the train set and evaluate the performance on 3D objects at all ranges.

Experiments show that, with LR3D, state-of-the-art detectors gain significant improvement in detecting distant objects without 3D annotation, *i.e.*, 14.8% improvement of DID-M3D [6] on KITTI [12], 15.9% improvement of BEVFormer [2] on nuScenes [13], 34.5% and 14.17% improvement of FCOS3D [1] on Cityscapes3D and Argoverse 2, 7.09% improvement of MV-FCOS3D++ [7] on Waymo. With LR3D, these methods yield competitive performance even compared to fully 3D supervised counterparts. Notably, LR3D enables them to detect extremely distant 3D objects as shown in Figure 1.

2. Related Work

LiDAR-based Detectors. LiDAR methods detect 3D objects from point clouds. According to inputs, these methods can be divided into point, voxel, and range-view detectors. Point methods [17–21] make 3D predictions from raw point clouds [22–27]. Voxel methods [28–31] transform point clouds into voxels and extract features by convolutions [32–37]. Other methods project point clouds into range-view [38–42] and process them as images [43]. Albeit impressively performed, LiDAR methods are 3D-label-greedy, and limited by the perspective range of LiDAR sensors, *i.e.*, usually fail for areas with few or no LiDAR points.

Camera-based Detectors. Camera-based methods do 3D detection from images. Monocular methods predict 3D boxes in single image directly [1, 44–47]. Stereo methods based on multi-view images formulate a 3D volume [48–50]. Recent methods [2, 51–57] build a bird-eye-view (BEV) representation for detection.

The advantages of Camera-based methods lie in the unbounded perception range of cameras, which makes them suitable for distant 3D detection. Some methods, *e.g.*, Far3D [58], are designed for long-range 3D detection, but still heavily rely on abundant high-quality distant 3D annotations, which are hard to collect. Some datasets [15, 16] provide distant annotations. However, their labeling is time-consuming and requires huge human efforts to take hints from multi-modal sensors and temporal consistency to label distant objects with few or no interior LiDAR points. The workload of the labeling process and the 3D-label-greedy character of existing detectors limit the applications of long-range 3D detection. Instead, we propose to detect distant 3D objects without long-range 3D box supervision. It is the simplest setting and is practical for scalability.

Object Distance Estimation. Another related topic is monocular distance estimation, which estimates the object distance from an RGB image. SVR [59] and DisNet [60] estimate object distance through pixel height and width. They base on an assumption of projected 2D box size determined by the distance of object only. However, other factors, like size and orientation, also affect the projected 2D box size.

Method of [61] develops a FastRCNN [62] pipeline for end-to-end distance estimation by directly regressing the distance from RoI features. R4D [9] utilizes reference objects for further improvements. These methods are limited by their requirements of abundant annotations. Given the difficulty in labeling distant 3D objects, in this paper, we develop a simpler 2D setting that even achieves competitive

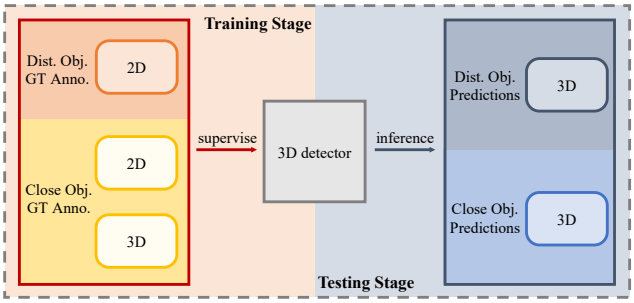

Figure 2. Illustration of LR3D which detects 3D boxes for both close and distant objects using the supervision of close 2D/3D and distant 2D bounding box annotations.

Distant 3D Groundtruth?	Location Err.	Size Err.	Orientation Err.
√	0.09	0.21	0.37
-	**0.34 (+0.25)**	0.23 (+0.02)	0.41 (+0.04)

Table 1. Performance comparison between FastRCNN3D trained with and without distant 3D supervision.

performance to their 3D supervised counterparts.

3. Method

LR3D is a long-range 3D detection framework that detects 3D bounding boxes, including locations, sizes, and orientations, of distant objects using only their 2D supervision (Figure 2). In this section, we first analyze the main challenge of directly utilizing existing 3D detectors on LR3D in Section 3.1 and then introduce our solution in Section 3.2.

3.1. Analysis

Albeit important and critical, long-range 3D object detection is seldom explored. One difficulty is to obtain sufficient high-quality 3D bounding box annotations for distant objects. We discard this high-human-labor-cost setting and propose a new framework, LR3D, to predict 3D boxes for distant objects only from their 2D image supervision.

Can Existing Methods Use only 2D Supervision? We first analyze if existing methods can use only 2D supervision for detecting distant 3D objects. To this end, we use FastRCNN3D, a FastRCNN-like [62] 3D detector where the original head for 2D detection is replaced by the FCOS3D head [1] for 3D detection. For training, we first manually assign objects over 40m as distant objects and remove their 3D annotations. Then, we train FastRCNN3D with close 2D/3D and distant 2D box labels as Figure 2. At test time, we use the 2D ground truth of distant objects as proposals, predict their 3D boxes, and evaluate their errors in location, size, and orientation. Note that we use relative distance as location error, IoU with aligned location and orientation as size error, and absolute difference as orientation error.

Table 1 compares errors between the LR3D configuration and traditional full supervision. As shown, errors in size and orientation are comparable. The main gap comes from the location error, in which depth estimation plays the essential role. Due to the lack of depth information in 2D box supervision, directly regressing the depth as previous methods cannot produce accurate depth prediction for distant objects and leads to a significant drop. Therefore, we seek a new method to implicitly estimate the depth.

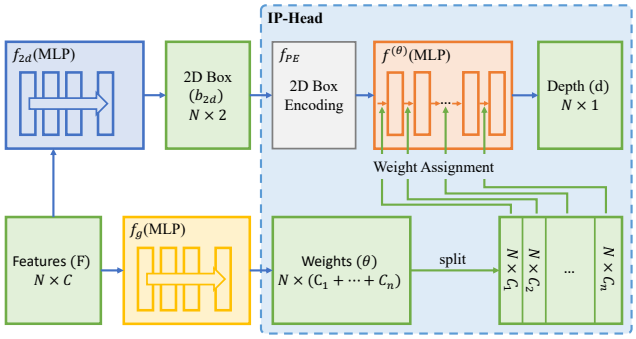

Figure 3. Illustration of IP-Head. We use an MLP $f^{(\theta)}$ to fit the implicit function from 2D box to 3D depth, of which the weights θ are dynamically determined by instance features including information of size and orientation.

3.2. Implicit Projection Head

Here, we introduce Implicit Projection head (IP-Head), our proposal to estimate depth of distant objects using only 2D supervision. Given a 3D object with fixed depth, size, and orientation [1], through the camera calibration matrix, it is easy to obtain the corresponding projected 2D bounding box (described by its width w_{2d} and height h_{2d}) on the target image. We use a function f, determined by the calibration matrix, to indicate the mapping between depth (d), size (s), orientation (o) and 2D box ($b_{2d} = (w_{2d}, h_{2d})$) as

$$f(d, s, o) = b_{2d}. \tag{1}$$

Eq. (1) shows the ubiquitous relation between d and b_{2d} if the object size s and orientation o are fixed – for objects with the same size and orientation, the further these objects locate, the smaller their projected 2D boxes on image are. Inspired by this fact, we ask if it is possible to estimate the inverse function f^{-1} to transfer the 2D bounding box to the corresponding depth conditioned by s and o, formulated as

$$f^{-1}(b_{2d}|s, o) = d. \tag{2}$$

With the power of neural networks to fit complicated functions, we utilize a small-size network with a multi-layer perceptron (MLP) to estimate the implicit inverse function f^{-1}. For simplicity, we use $f^{(\theta)}$ to represent this network, of which the parameter weights are represented as θ.

Since the implicit inverse function f^{-1} depends on the size and orientation of the specific 3D objects, $f^{(\theta)}$ should also be different across multiple objects, which means weights θ should be dynamic.

[1]The mentioned orientation is the relative one to the camera, i.e., the observed orientation on the image. We refer the readers to FCOS3D [1] and Stereo-RCNN [63] for more details on obtaining relative orientation.

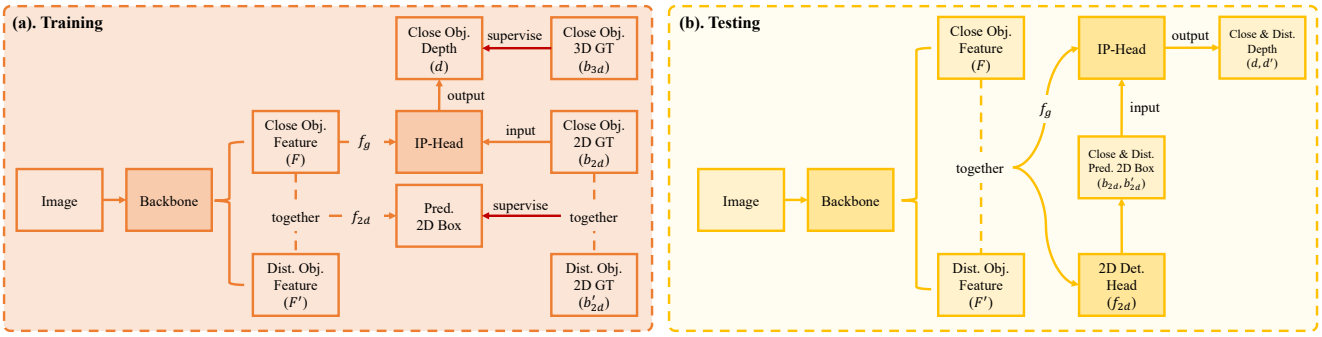

Figure 4. Illustration of the training and testing pipeline of IP-Head. **(a). Training:** During training, we use 2D/3D annotation pairs of close objects to supervise f_g to generate dynamic weights of MLP $f^{(\theta)}$ which models the transformation of target 3D object from 2D box to corresponding depth in Eq. (3). **(b). Testing:** During testing, we use a 2D detection head (2D Det. Head) f_{2d} to generate 2D detection results for all objects. They are then transferred to corresponding depth by IP-Head.

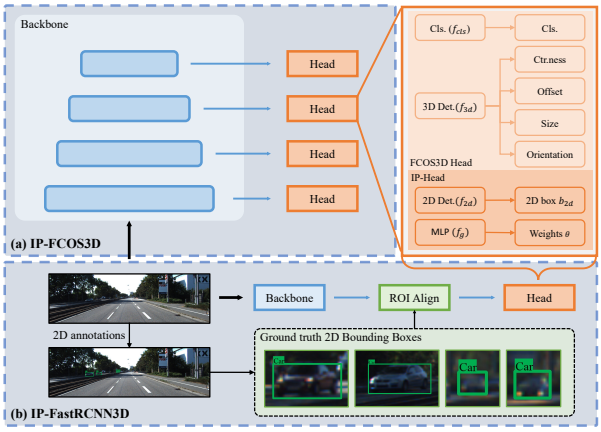

Figure 5. Illustration of deploying IP-Head to monocular 3D detectors: **(a)** FCOS3D; **(b)** FastRCNN3D.

With these considerations, rather than utilizing a shared θ for all objects, we use a trainable MLP f_g to generate a set of dynamic weights θ_i according to the features F_i of each object i. We then utilize those θ_i as the weights of network $f^{(\theta)}$ to estimate the corresponding depth of the i-th 2D box. This process generates the specific **I**mplicit inverse function of each object to **P**roject its 2D box to 3D depth. We illustrate this procedure in Figure 3, in which "2D Box Encoding", f_{PE}, is a positional encoding function [64] to encode 2-channel 2D box descriptors (with width w_{2d} and height h_{2d}) into informative high-dimensional features and f_{2d} is a 2D detection network for predicting 2D bounding boxes on the image. The overall procedure is formulated as

$$d_i = f^{(f_g(F_i))}(f_{\text{PE}}(b_{2d_i})), \tag{3}$$

where f_g estimates the weights of $f^{(\theta_i)}$ from instance features F_i to transfer the i-th 2D bounding boxes to its corresponding depth. Its condition information, including size s_i and orientation o_i, is included in feature F_i, from which information can be obtained like that of [1, 65, 66].

During training, we make use of 2D/3D annotation pairs of close objects to supervise IP-Head for obtaining a reliable dynamic weight generator f_g. Specifically, for a close object, after obtaining its corresponding dynamic weights,

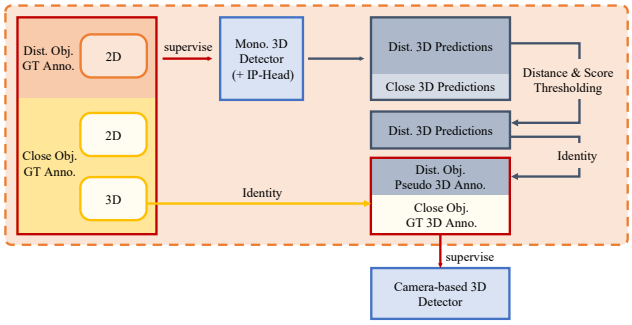

Figure 6. Illustration of extending IP-Head to all camera-based 3D detectors through a teacher-student pipeline.

we transfer its 2D ground truth box for depth prediction by IP-Head, and optimize f_g by computing the loss between the predicted depth and the 3D annotation.

During inference, we first use the backbone network to extract instance features F'_i. Then, we extract dynamic weights using f_g and use a detection network f_{2d}, supervised by 2D bounding box ground truth, to predict the associated 2D bounding box b'_{2d_i} as $b'_{2d_i} = f_{2d}(F'_i)$. Finally, we use Eq. (3) to obtain depth predictions for each instance in the image. The overall process is illustrated in Figure 4.

Using neural networks to directly estimate a hard-to-optimize function was successful in other research areas. For example, NeRF [67] uses MLP networks to learn a shared mapping from positions to colors and densities; classification models use a shared CNN model to learn the function from pixels to classification confidence. In contrast to previous methods, we first study the estimation of transformation function from 2D boxes on images to 3D depths. Moreover, instead of learning a shared mapping, IP-Head generates dynamic mappings from 2D box sizes to 3D depth according to the 3D size and orientation of target objects, which equips IP-Head with great generalization ability and potential applied to complex autonomous driving scenarios.

Projection Augmentation. The implicit inverse function $f^{(\theta)}$ needs to model the relation between the 2D box and corresponding depth, and generate different depth predictions based on 2D box input. To further improve effective-

ness, we also propose an augmentation strategy, called projection augmentation. The idea is to generate more depth d and 2D box b_{2d} pairs for each close object during training, so as to enable f_g to estimate a more accurate b_{2d}-d relation.

These extra b_{2d}-d training pairs come from Eq. (1). Given an object with fixed size and orientation, we randomly choose different depth values d^*, calculate their corresponding 2D boxes b^*_{2d} through Eq. (1), and utilize these augmented b^*_{2d}-d^* pairs, along with the ground truth b_{2d}-d pair, to train the IP-Head for higher performance.

Long-range Teacher The proposed IP-Head can be used in existing monocular 3D detectors, like FCOS3D [1], to boost their performance in LR3D. As illustrated in Figure 5, only two additional branches, i.e., a 2D detection branch f_{2d} and a weight generation MLP f_g, are needed to utilize IP-Head in FCOS3D. Apart from monocular methods, BEV methods become popular due to their strong performance and multi-tasking ability. We extend IP-Head to BEV methods to alleviate their demand of 3D annotations.

Our solution is to utilize a monocular method equipped with IP-Head as the detector of the LR3D model which then serves as the long-range teacher to generate pseudo distant 3D annotations. With close 2D/3D annotations and distant 2D annotations, we first apply LR3D to generate 3D predictions for distant objects. Then, we treat these predictions as pseudo 3D box labels, which are, together with close 3D ground truth, taken as the whole supervision to train BEV methods. Our pipeline is shown in Figure 6. Experiments show that, it works decently for various 3D detectors on long-range objects without distant 3D box annotation.

4. Long-range Detection Score

Existing detection metrics are built upon mAP based on either IoU [12, 14] or absolute distance error [13]. They use fixed IoU or error thresholds as criterion to calculate AP.

However, the fixed threshold is not suitable for distant objects due to its neglect of the increasing error in depth estimation when objects go farther. For example, though DID-M3D [6] is the state-of-the-art camera-based detector, it only achieves 0.43% mAP under fixed IoU criterion on objects over 40m. Fixed threshold metrics hardly bring informative results. In fact, further objects should have a larger tolerance to closer ones, since estimating their depth is more ill-posed. Similar ideas were taken in tasks like dense depth estimation [68] and object distance estimation [9]. They all prefer relative distance errors as their metrics.

To address this limitation, we introduce Long-range Detection Score (LDS), a new metric for long-range 3D object detection. LDS is based on the widely applied nuScenes Detection Score (NDS) [13] and it is defined as

$$\text{LDS} = \frac{1}{6}[3\text{mAP} + \text{Rec} \times \sum_{\text{mTP} \in \text{mTP}} (1 - \min(1, \text{mTP}))], \quad (4)$$

where Rec is the recall rate and mTP represents the mean True Positive metric.

LDS improves NDS in two main aspects: the criterion of mAP and the multiplication of Rec and summation of mTP.

Improvements on mAP. In LDS, we compute mAP based on the relative distance error of

$$\text{Rel. Dist. Err.} = \frac{\|P_c - G_c\|}{G_d}, \quad (5)$$

where P_c, G_c and G_d represent the center of predicted 3D box, center of ground truth 3D box and the distance of ground truth 3D box towards ego vehicle, respectively. Predictions with a relative error smaller than a threshold r are counted as true positive, and false positive otherwise, for computing AP. We choose 4 thresholds $\mathbb{R} = \{0.025, 0.05, 0.1, 0.2\}$ and take average over these thresholds and the class set $\mathbb{C}$. Finally, we obtain mAP as

$$\text{mAP} = \frac{1}{|\mathbb{C}||\mathbb{R}|} \sum_{c \in \mathbb{C}} \sum_{r \in \mathbb{R}} \text{AP}_{c,r}. \quad (6)$$

Improvement on mTP. Also, we multiply the recall rate and mTP before adding to mAP. The mTP is utilized to measure errors on the location (mATE), size (mASE) and orientation (mAOE) for TP prediction, whose relative distance to the ground truth is smaller than $r = 0.1$ during matching. mATE is computed as the relative distance, normalized by 0.1 to ensure range falling within 0 and 1.

mASE and mAOE are the same as those in nuScenes Dataset [13]. The intuition of multiplying the recall rate to the mTP is simple. The larger the recall rate is, the more predictions are involved in the statistics of mTP. Compared to simply setting a recall threshold [13], the multiplication improvement adjusts the weight of mTP to LDS according to its comprehensiveness, and thus brings a more informative quantitative result.

5. Experiments

We evaluate our method on five popular 3D detection datasets of KITTI [12], nuScenes [13], Cityscapes3D [16], Waymo Open Dataset [14] and Argoverse 2 [15], which have plenty of high-quality 3D annotations for both close and distant objects. **Due to the space limitation, we show the experimental results on Cityscapes3D, Waymo and Argoverse 2 datasets in the supplementary material.**

5.1. Results on KITTI

Data Preparation. KITTI Dataset [12] provides high-quality 2D/3D annotations for objects in "Car", "Pedestrian" and "Cyclist" within the range of ∼80m. For extremely distant objects, due to lack of points, it only labels their 2D ground truth boxes on images and marks them as

Method	Distant 3D Groundtruth?	Overall (0m-Inf)		Close (0m-40m)		Distant (40m-Inf)	
		LDS (%)	mAP (%)	LDS (%)	mAP (%)	LDS (%)	mAP (%)
FCOS3D [1]	√	48.0	44.7	49.9	47.5	38.2	32.6
FCOS3D [1]	-	42.1	38.9	50.5	47.1	4.9	3.3
LR3D (IP-FCOS3D)	-	**50.0**	**46.7**	52.1	49.4	**36.2**	**31.0**
Long-range Teacher							
ImVoxelNet [52]	√	44.8	45.2	48.1	48.4	26.2	24.7
ImVoxelNet [52]	-	40.3	39.9	48.6	48.1	4.4	4.1
+LR3D teacher	-	**44.9**	**45.1**	47.9	47.9	**26.9**	**25.3**
CaDDN [53]	√	51.1	50.1	57.8	56.3	19.3	18.5
CaDDN [53]	-	48.9	48.6	58.5	57.9	5.5	5.7
+LR3D teacher	-	**51.1**	**50.6**	57.6	56.8	**20.9**	**19.8**
MonoFlex [65]	√	55.5	52.4	57.7	55.2	40.6	34.8
MonoFlex [65]	-	50.1	47.3	59.3	56.1	8.4	6.6
+LR3D teacher	-	**54.3**	**51.6**	57.3	54.8	**35.8**	**32.9**
GUPNet [66]	√	49.3	47.5	54.0	52.1	26.7	25.1
GUPNet [66]	-	46.8	45.2	54.0	52.3	17.6	16.5
+LR3D teacher	-	**49.4**	**47.8**	54.2	52.8	**26.4**	**24.4**
DID-M3D [6]	√	56.1	55.1	58.5	57.2	38.9	36.7
DID-M3D [6]	-	52.5	51.4	58.9	57.8	24.2	22.5
+LR3D teacher	-	**56.3**	**55.0**	58.7	57.2	**39.0**	**36.2**

Table 2. Comparison on state-of-the-art methods with and without IP-Head or LR3D teacher supervised by distant 2D ground truth only on the KITTI val dataset. Their fully supervised counterparts (with distant 3D ground truth) are also illustrated.

"DontCare". Given these labels, we accordingly conduct quantitative and qualitative evaluations on KITTI Dataset.

Quantitative evaluation is conducted on objects with 3D annotations. Specifically, based on the official provided 3D annotations, we mark annotations over 40m as distant ones following [9], only using their 2D labels for training. For those closer than 40m, we keep both 2D and 3D labels for training. We report LDS on both close and distant objects for statistical results. The split of train and val sets follows [69]. Due to the lack of "Pedestrian" and "Cyclist" labels beyond 40m, we only report results on class "Car".

Qualitative evaluation is designed for "DontCare" objects which are extremely far away and without 3D annotations due to no LiDAR points. We visualize the 3D detection results of our model conditioned by their 2D ground truth boxes. Through this evaluation, we manifest the capacity of our model for extremely distant 3D detection.

Model Setting. For quantitative evaluation, we utilize FCOS3D [1] as our baseline for its simplicity and implementation platform. FCOS3D is implemented on the MMDetection3D [70] platform including multiple methods and datasets. We extend FCOS3D with our IP-Head (IP-FCOS3D, Figure 5) as the detector in the LR3D framework. For qualitative evaluation, we utilize FastRCNN3D with IP-Head (Figure 5) as LR3D detector, to utilize 2D conditions.

Quantitative Results. We compare our LR3D, using IP-FCOS3D as its detector, and the baseline FCOS3D [1] for distant 3D detection without 3D annotations in Table 2. It is obvious that our LR3D outperforms FCOS3D on detecting distant objects by a large margin, i.e., 31.3% and 27.7% on LDS and mAP, respectively. Even compared to FCOS3D with distant 3D ground truth supervision, LR3D still shows competitive performance on distant objects.

We further evaluate the performance of using LR3D as the teacher model to generate pseudo long-range 3D labels to train state-of-the-art camera-based detectors. The results are listed in Table 2. Surprisingly, models trained with the combination of close 3D ground truth and distant 3D pseudo labels even achieve comparable or even better performance to their fully 3D supervised counterparts.

Quantitative results in Table 2 demonstrate the effectiveness of our LR3D framework, as well as the potential of our method in relieving the great demand of accurate 3D bounding box annotations for long-range object detection.

Qualitative Results. We visualize the 3D prediction results of LR3D conditioned with "DontCare" objects in Figure 7. It is also clear that supervised by close 3D annotations only (maximum to 40m away), our method infers reasonable 3D bounding boxes (3rd & 4th rows) for those extremely distant objects much beyond the 3D supervision range according to the corresponding 2D bounding boxes (1st & 2nd rows). These qualitative results further demonstrate the potential to extend LR3D to the labeling process. Annotators can first simply label 2D boxes for distant objects and then generate associated 3D annotations by LR3D.

5.2. Results on nuScenes

Data Preparation. NuScenes [13] is a large-scale dataset with 1,000 autonomous driving sequences labeled with 3D boxes in 10 classes. In nuScenes, we set the distance threshold of marking long-range objects as 40m. We mark those annotations beyond the threshold as distant objects, remove the 3D annotations and only use 2D box labels for training. For those closer than 40m, we keep their both 2D/3D labels.

We evaluate different models among all 10 classes, and do not set the detection range upper-bound for different

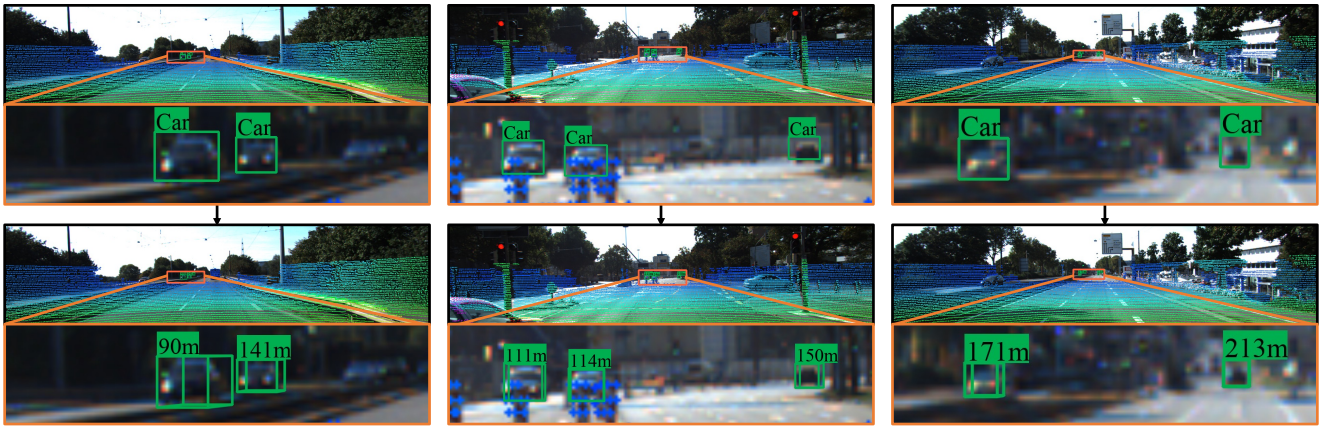

Figure 7. Qualitative results on KITTI Dataset in detecting extremely far away 3D objects. LiDAR points are projected to images shown with different colors related to depth. The distances of 3D bounding boxes are marked on the top-left. **1st & 2nd rows:** 2D annotations of extremely distant objects; **3rd & 4th rows:** 3D box prediction of corresponding objects.

Method	Distant 3D Groundtruth?	Overall		Close (0m-40m)		Distant (40m-51.2m)		Distant (51.2m-Inf)	
		LDS (%)	mAP (%)	LDS (%)	mAP (%)	LDS (%)	mAP (%)	LDS (%)	mAP (%)
FCOS3D [1]	√	26.6	27.4	29.1	30.1	16.4	12.7	11.3	8.6
FCOS3D [1]	-	22.6	23.4	29.9	30.3	1.8	1.4	0.0	0.0
LR3D (IP-FCOS3D)	-	**24.1**	**24.4**	28.5	28.4	**16.1**	**11.3**	**6.4**	**4.3**
Long-range Teacher									
BEVFormer-S [2]	√	36.9	37.2	38.3	38.7	19.6	15.2	-	-
BEVFormer-S [2]	-	33.6	34.2	38.0	38.5	2.4	1.7	-	-
+LR3D teacher	-	**36.3**	**36.8**	38.0	38.9	**18.3**	**10.9**	-	-
BEVFormer-S† [2]	√	33.7	34.0	37.4	37.9	19.4	14.5	14.4	10.2
BEVFormer-S† [2]	-	29.1	29.9	38.0	38.7	2.4	1.0	0.0	0.0
+LR3D teacher	-	**33.0**	**33.6**	37.5	38.0	**17.9**	**10.7**	**12.7**	**6.4**

Table 3. Comparison on state-of-the-art methods with and without IP-Head or LR3D teacher supervised by distant 2D ground truth only on the nuScenes val dataset. Their fully supervised counterparts (with distant 3D ground truth) are also illustrated.

classes as the official benchmark [13]. This setting helps keep distant predictions. Following [13], we train our model on 700 training scenes and test it on 150 validation scenes.

Model Setting. We utilize FCOS3D [1] as our monocular baseline and enhance it with IP-Head (IP-FCOS3D) for LR3D with 2D supervision only. To demonstrate the effectiveness of LR3D as a long-range teacher on nuScenes, we choose BEVFormer [2] as the student for its effectiveness, and test its performance without distant 3D supervision.

Main Results. The experimental results are illustrated in Table 3. With IP-Head, LR3D outperforms its baseline FCOS3D by 14.3% and 9.9% performance improvement in terms of LDS and mAP for distant objects from 40m to 51.2m. It is significant. Compared to the fully 3D supervised FCOS3D with distant 3D ground truth, LR3D still yields competitive capacity on long-range 3D detection.

We also list the results of BEVFormer [2] with and without LR3D teacher on detecting distant 3D objects without 3D supervision. As illustrated, with the LR3D teacher, BEVFormer-S yields strong performance, comparable to its fully-supervised counterparts with 3D box labels.

Given the limited perception range of official BEVFormer (-51.2m to 51.2m), we enlarge the perception range of BEVFormer from 51.2m to 76.8m for further verification, denoted as BEVFormer-S†. Still, as shown in Table 3, with LR3D teacher, BEVFormer-S† achieves impressive improvement compared to its baseline, and yields comparable results to those of the fully-supervised model.

5.3. Ablation Studies

All ablation studies are conducted on KITTI Dataset, using IP-FCOS3D as the detector of the LR3D model.

Analysis on the IP-Head. IP-Head learns to estimate implicit inverse function f^{-1}, which is the mapping from 2D bounding box on image to corresponding depth, through two MLP networks f_g and $f^{(\theta)}$. We evaluate its effectiveness by comparing the mapping curve from 2D box size to the depth of a fixed 3D object generated by ground truth transformation and estimated transformation.

The generation of the ground truth curve follows Eq. (1), from which we obtain multiple d-b_{2d} pairs by modifying depth d. The estimated curve is generated through Eq. (3) by changing the 2D box inputs. As illustrated in Figure 8, our IP-Head can estimate specific implicit inverse functions for different objects (by comparison across rows). The estimated function by MLP network $f^{(\theta)}$ well models the mapping from 2D bounding box size to the corresponding depth of target 3D object (through comparison inside each row).

$f^{(\theta)}$ Weight Learning Strategy	Overall (0m-Inf) LDS (%)	Close (0m-40m) LDS (%)	Distant (40m-Inf) LDS (%)
Shared	45.0	46.6	32.5
Dynamic	**50.0**	**52.1**	**36.2**

(a) Effect of parameter learning in $f^{(\theta)}$.

Pos. Enc.	Overall (0m-Inf) LDS (%)	Close (0m-40m) LDS (%)	Distant (40m-Inf) LDS (%)
None	19.9	20.2	13.3
Sin. Cos. [64]	**50.0**	**52.1**	**36.2**

(b) Effect of positional encoding in $f^{(\theta)}$.

2D Box (b_{2d}) Descriptors	Overall (0m-Inf) LDS (%)	Close (0m-40m) LDS (%)	Distant (40m-Inf) LDS (%)
box width (w_{2d})	48.5	51.7	30.6
box height (h_{2d})	49.4	51.8	34.9
both (w_{2d}, h_{2d})	**50.0**	**52.1**	**36.2**

(c) Effect of 2D box descriptors.

No. of layers	Overall (0m-Inf) LDS (%)	Close (0m-40m) LDS (%)	Distant (40m-Inf) LDS (%)
1	48.6	51.3	32.4
2	**50.0**	**52.1**	**36.2**
3	49.7	52.0	35.2
4	49.0	51.4	34.9

(d) Effect of the number of layers in $f^{(\theta)}$.

No. of channels	Overall (0m-Inf) LDS (%)	Close (0m-40m) LDS (%)	Distant (40m-Inf) LDS (%)
8	48.4	50.9	34.5
16	**50.0**	52.1	**36.2**
32	49.9	**52.3**	35.6
64	49.7	51.9	36.1

(e) Effect of the channel number in $f^{(\theta)}$.

Aug. Strategy	Overall (0m-Inf) LDS (%)	Close (0m-40m) LDS (%)	Distant (40m-Inf) LDS (%)
FCOS3D [1]	42.1	50.5	4.9
+Copy-Paste	N/A	N/A	N/A
LR3D	48.3	**52.7**	26.5
+Proj. Aug.	**50.0**	52.1	**36.2**

(f) Effect of the projection augmentation.

Table 4. Ablation studies on IP-Head structure and projection augmentation. Default settings are highlighted in lightcyan.

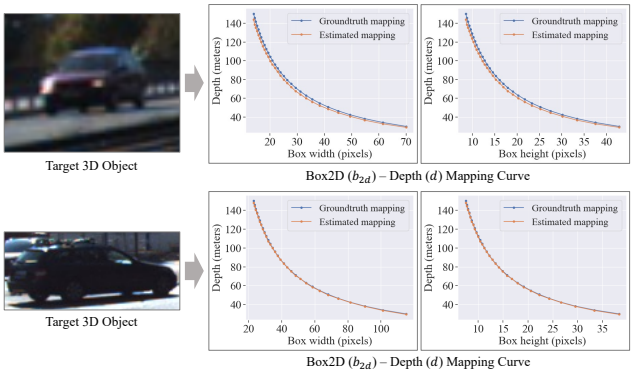

Figure 8. Illustration of the ground truth and estimated b_{2d}-d mappings. Each row indicates the target 3D object and its mapping from 2D box width and height to the depth.

We further provide quantitative analysis of using MLP network f_g to dynamically determine weights of $f^{(\theta)}$ to estimate specific implicit inverse functions of different 3D objects. As listed in Table 4a, compared to the design of shared weights, using dynamic weights assignment in $f^{(\theta)}$ well models the difference of implicit inverse functions due to size and orientation of 3D objects, and thus outperforms the model using a shared weight θ across 3D objects.

Analysis on the IP-Head Structure. We first describe the default structure of IP-Head. $f^{(\theta)}$ is a 2-layer perceptron with channels of 16 and 1. The 2D boxes b_{2d} are described as height (h_{2d}) and width (w_{2d}). The positional encoding function f_{PE} is the sine-cosine function illustrated in [64].

In Table 4b, we evaluate the contribution of positional encoding. With positional encoding [64], the 2D descriptors are well encoded into more informative features, which thus outperforms the model without positional encoding ("None"). In Table 4c, we compare the model with different 2D box descriptors. Utilizing both 2D box width w_{2d} and height h_{2d} as descriptors yields the best performance. In Table 4d and Table 4e, we conduct ablation studies on MLP structures of network $f^{(\theta)}$ with different numbers of layers and channels. As illustrated, the model with 2-layer and channel number 16 achieves the best results.

Analysis on the Projection Augmentation. The projection augmentation ensures the estimated implicit inverse function $f^{(\theta)}$ models the relation between 2D box sizes and depths by generating more training b_{2d}-d pairs. We analyze it by comparing the performance of LR3D trained with and without this augmentation. As shown in Table 4f, compared to FCOS3D baseline (1st row), LR3D can model mapping functions from 2D sizes to depths and produces good results on distant objects without 3D supervision. Moreover, equipped with projection augmentation, LR3D further shows a huge performance improvement on distant objects.

A similar augmentation to projection augmentation is copy-paste augmentation [71, 72], which simulates distant training samples by cropping close objects with image patches and pasting them to distant areas by Eq. (1). However, this augmentation brings noises in image distribution, is hard to optimize, and thus leads to a great performance drop (Table 4f (2nd row)). In contrast, projection augmentation is simple to implement without changing training images, and able to augment to any depth without extra effort.

5.4. Results on Other Datasets

Please find additional results on Cityscapes3D [16], Waymo [14], and Argoverse 2 [15] in the supplementary materials.

6. Conclusion

In this paper, we proposed the LR3D framework for long-range 3D detection using 2D supervision and designed the IP-Head as the key component. We developed a teacher-student pipeline to benefit all camera-based detectors. Our LDS, a metric with relative distance criterion, helps obtain informative quantitative results for distant objects. With these designs, we demonstrate the feasibility of using 2D supervision only for long-range 3D detection.

References

[1] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: Fully convolutional one-stage monocular 3d object detection. In *ICCV Workshops*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[2] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 2, 7

[3] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection. In *AAAI*, 2023.

[4] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position Embedding Transformation for Multi-View 3D Object Detection. In *ECCV*, 2022.

[5] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, Jie Zhou, and Jifeng Dai. BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision. In *CVPR*, 2023.

[6] Liang Peng, Xiaopei Wu, Zheng Yang, Haifeng Liu, and Deng Cai. Did-m3d: Decoupling instance depth for monocular 3d object detection. In *ECCV*, 2022. 2, 5, 6

[7] Tai Wang, Qing Lian, Chenming Zhu, Xinge Zhu, and Wenwei Zhang. MV-FCOS3D++: Multi-View camera-only 4d object detection with pretrained monocular backbones. *arXiv preprint*, 2022. 1, 2

[8] Braking distance. https://en.wikipedia.org/wiki/Braking_distance. 1

[9] Yingwei Li, Tiffany Chen, Maya Kabkab, Ruichi Yu, Longlong Jing, Yurong You, and Hang Zhao. R4D: utilizing reference objects for long-range distance estimation. In *ICLR*, 2022. 1, 2, 5, 6

[10] Federal Motor Carrier Safety Administration. Long stopping distances, Federal Motor Carrier Safety Administration. https://www.fmcsa.dot.gov/ourroads/long-stopping-distances, 2016. Accessed: Sep-09-2021.

[11] Myra Blanco and Jonathan M Hankey. Visual performance during nighttime driving in fog. Technical report, FHWA-HRT-04-137, 2005. 1

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *I. J. Robotics Res.*, 2013. 1, 2, 5

[13] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *CVPR*, 2020. 1, 2, 5, 6, 7

[14] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 2, 5, 8

[15] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS Datasets and Benchmarks 2021*, 2021. 2, 5, 8

[16] Nils Gählert, Nicolas Jourdan, Marius Cordts, Uwe Franke, and Joachim Denzler. Cityscapes 3d: Dataset and benchmark for 9 dof vehicle detection. *arXiv preprint arXiv:2006.07864*, 2020. 2, 5, 8

[17] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*, 2019. 2

[18] Charles Ruizhongtai Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum pointnets for 3d object detection from RGB-D data. *CVPR*, 2018.

[19] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector, 2020.

[20] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. STD: sparse-to-dense 3d object detector for point cloud. *ICCV*, 2019.

[21] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. IPOD: intensive point-based object detector for point cloud. *CoRR*, 2018. 2

[22] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 2

[23] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, 2017.

[24] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *NIPS*, 2018.

[25] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021.

[26] Zetong Yang, Li Jiang, Yanan Sun, Bernt Schiele, and Jiaya Jia. A unified query-based paradigm for point cloud understanding. In *CVPR*, 2022.

[27] Zetong Yang, Yanan Sun, Shu Liu, Xiaojuan Qi, and Jiaya Jia. CN: channel normalization for point cloud recognition. In *ECCV*, 2020. 2

[28] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. *CVPR*, 2021. 2

[29] Zetong Yang, Yin Zhou, Zhifeng Chen, and Jiquan Ngiam. 3d-man: 3d multi-frame attention network for object detection. In *CVPR*, 2021.

[30] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. *CVPR*, 2019.

[31] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: point-voxel feature set abstraction for 3d object detection. In *CVPR*, 2020. 2

[32] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. 2

[33] Ben Graham. Sparse 3d convolutional neural networks. In *BMVC*, 2015.

[34] Christopher B. Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019.

[35] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 2018.

[36] Bin Yang, Wenjie Luo, and Raquel Urtasun. PIXOR: real-time 3d object detection from point clouds. In *CVPR*, 2018.

[37] Li Jiang, Zetong Yang, Shaoshuai Shi, Vladislav Golyanik, Dengxin Dai, and Bernt Schiele. Self-supervised pre-training with masked shape prediction for 3d scene understanding. In *CVPR*, 2023. 2

[38] Alex Bewley, Pei Sun, Thomas Mensink, Dragomir Anguelov, and Cristian Sminchisescu. Range conditioned dilated convolutions for scale invariant 3d object detection. In Jens Kober, Fabio Ramos, and Claire J. Tomlin, editors, *CoRL*, 2020. 2

[39] Pei Sun, Weiyue Wang, Yuning Chai, Gamaleldin Elsayed, Alex Bewley, Xiao Zhang, Cristian Sminchisescu, and Dragomir Anguelov. RSN: range sparse net for efficient, accurate lidar 3d object detection. In *CVPR*, 2021.

[40] Bo Li, Tianlei Zhang, and Tian Xia. Vehicle detection from 3d lidar using fully convolutional network. *arXiv preprint arXiv:1608.07916*, 2016.

[41] Lue Fan, Xuan Xiong, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Rangedet: In defense of range view for lidar-based 3d object detection. In *ICCV*, 2021.

[42] Yuning Chai, Pei Sun, Jiquan Ngiam, Weiyue Wang, Benjamin Caine, Vijay Vasudevan, Xiao Zhang, and Dragomir Anguelov. To the point: Efficient 3d object detection in the range image with graph convolution kernels. In *CVPR*, 2021. 2

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[44] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Probabilistic and Geometric Depth: Detecting objects in perspective. In *CoRL*, 2021. 2

[45] Garrick Brazil and Xiaoming Liu. M3D-RPN: monocular 3d region proposal network for object detection. In *ICCV*, 2019.

[46] Andrea Simonelli, Samuel Rota Bulò, Lorenzo Porzi, Manuel Lopez-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *ICCV*, 2019.

[47] Nils Gählert, Jun-Jun Wan, Nicolas Jourdan, Jan Finkbeiner, Uwe Franke, and Joachim Denzler. Single-shot 3d detection of vehicles from monocular RGB images via geometry constrained keypoints in real-time. *arXiv preprint arXiv:2006.13084*, 2020. 2

[48] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Dsgn: Deep stereo geometry network for 3d object detection. *CVPR*, 2020. 2

[49] Xiaoyang Guo, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector. In *ICCV*, 2021.

[50] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *ICCV*, 2019. 2

[51] Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv*, 2022. 2

[52] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *WACV*, 2022. 6

[53] Cody Reading, Ali Harakeh, Julia Chae, and Steven L. Waslander. Categorical depth distributionnetwork for monocular 3d object detection. *CVPR*, 2021. 6

[54] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying Voxel-based Representation with Transformer for 3D Object Detection. In *NeurIPS*, 2022.

[55] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, 2020.

[56] Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[57] Haisong Liu, Haiguang Wang, Yang Chen, Zetong Yang, Jia Zeng, Li Chen, and Limin Wang. Fully sparse 3d panoptic occupancy prediction, 2023. 2

[58] Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, and Xiangyu Zhang. Far3d: Expanding the horizon for surround-view 3d object detection. *arXiv preprint arXiv:2308.09616*, 2023. 2

[59] Fatih Gökçe, Göktürk Üçoluk, Erol Şahin, and Sinan Kalkan. Vision-based detection and distance estimation of micro unmanned aerial vehicles. *Sensors*, 2015. 2

[60] Muhammad Abdul Haseeb, Jianyu Guan, Danijela Ristić-Durrant, and Axel Gräser. DisNet: A novel method for distance estimation from monocular camera. *IROS*, 2018. 2

[61] Jing Zhu and Yi Fang. Learning object-specific distance from a monocular image. In *ICCV*, 2019. 2

[62] Ross B. Girshick. Fast R-CNN. In *ICCV*, 2015. 2, 3

[63] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *CVPR*, 2019. 3

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NIPS*, 2017. 4, 8

[65] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *CVPR*, 2021. 4, 6

[66] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. *arXiv preprint arXiv:2107.13774*, 2021. 4, 6

[67] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 4

[68] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 5

[69] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. *CVPR*, 2018. 6

[70] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 6

[71] Wenwei Zhang, Zhe Wang, and Chen Change Loy. Multi-modality cut and paste for 3d object detection. *arXiv preprint arXiv:2012.12741*, 2020. 8

[72] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *CVPR*, 2021. 8