FD3D: Exploiting Foreground Depth Map for Feature-Supervised Monocular 3D Object Detection

Zizhang Wu[1], Yuanzhu Gan[2], Yunzhe Wu[2], Ruihao Wang[2], Xiaoquan Wang[3], Jian Pu[1]*

[1] Fudan University
[2] ZongmuTech
[3] ExploAI

wuzizhang87@gmail.com, {yuanzhu.gan, nelson.wu, ruihao.wang}@zongmutech.com, rocky.wang@exploai.com, jianpu@fudan.edu.cn

Abstract

Monocular 3D object detection usually adopts direct or hierarchical label supervision. Recently, the distillation supervision transfers the spatial knowledge from LiDAR- or stereo-based teacher networks to monocular detectors, but remaining the domain gap. To mitigate this issue and pursue adequate label manipulation, we exploit ***F***oreground ***D***epth map for feature-supervised monocular ***3D*** object detection named ***FD3D***, which develops the high-quality instructive intermediate features to conduct desirable auxiliary feature supervision with only the original image and annotation foreground object-wise depth map (AFOD) as input. Furthermore, we build up our instructive feature generation network to create instructive spatial features based on the sufficient correlation between image features and pre-processed AFOD, where AFOD provides the attention focus only on foreground objects to achieve clearer guidance in the detection task. Moreover, we apply the auxiliary feature supervision from the pixel and distribution level to achieve comprehensive spatial knowledge guidance. Extensive experiments demonstrate that our method achieves state-of-the-art performance on both the KITTI and nuScenes datasets, with no external data and no extra inference computational cost. We also conduct experiments to reveal the effectiveness of our designs.

Introduction

3D object detection is crucial for the perception task in extensive applications such as autonomous driving and robotic manipulation (Reading et al. 2021; Liu, Wu, and Tóth 2020). Considering different scenarios, recent 3D detection approaches (Thomas et al. 2019; Nabati and Qi 2021; Huang et al. 2022; Sun et al. 2020) measure the objects' precise location from the different-modalities inputs, such as 3D point clouds, radar signals, monocular images or stereo images. In particular, the monocular setting adopting the deployment of a single RGB camera has attracted increasing attention.

The usual pipeline for monocular 3D object detection reveals to apply the direct label supervision with the well-designed model and constrains (Zhang, Lu, and Zhou 2021; Reading et al. 2021; Huang et al. 2022), or deliver the hierarchical label supervision on different-layer features (Lu et al. 2021). Recently, researchers have explored the distillation schema to train the LiDAR- or stereo-based teacher networks to transfer the learned spatial features' knowledge to monocular 3D detectors (Chong et al. 2022; Chen, Dai, and Ding 2022). It actually exhibits an auxiliary feature supervision to the original monocular baseline, with the elaborate manipulation of the labels, teacher models, and inputs like LiDAR or stereo sensors.

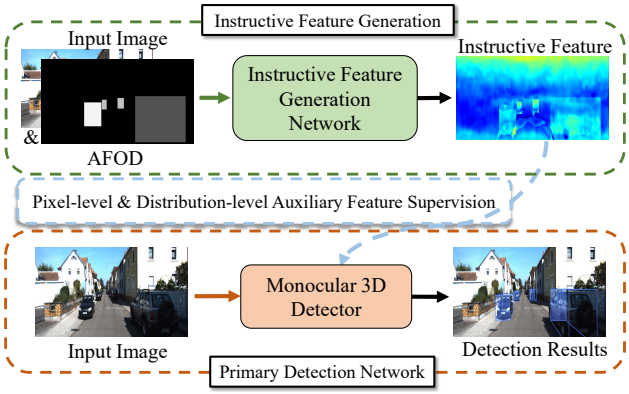

Figure 1: Illustration of our auxiliary feature supervision framework, where the instructive features are trained from the generation network with inputs of image and annotation foreground object-wise depth (AFOD).

Notably, these multi-modal sensors bring robust spatial guidance like feature-level (Chen, Dai, and Ding 2022) or object-level (Chong et al. 2022) adaptation but remain more expensive compared with the monocular image-only setting (Yin, Zhou, and Krahenbuhl 2021; Li and Zhao 2021). In addition, their teacher network's training stays restricted by the domain gap, which reveals the other-modal input, the heavier feature extractor and more challenging cross-modal feature distillation. Furthermore, they utilize indirect label supervision to make inaccurate predictions under the model limitation, ignoring the potential of ground truth as a direct indication to generate reliable spatial features.

To mitigate these issues and pursue adequate label manipulation, we exploit ***F***oreground ***D***epth map for feature-supervised monocular ***3D*** object detection named ***FD3D***, which develops the high-quality instructive intermediate

*Corresponding author

features to conduct desirable auxiliary feature supervision with only the original image and annotation foreground object-wise depth map (AFOD) as input. As shown in Fig. 2, it can be considered as an elegant auxiliary feature supervision with a novel input-label manipulation for excellent label absorbance. In specific, within the instructive feature generation network (IFGN), we cancel the heavy LiDAR or stereo feature extractor. Instead, we pre-process the foreground object-wise depth label to achieve alignment with the image domain, where AFOD provides the attention focus only on foreground objects to achieve clearer guidance in the detection task. Moreover, we propose a vision-depth association module (VDAM) to promote the semantic and spatial clues correlation with image and AFOD as input. Noticeably, the IFGN actually keeps the similar light monocular detection framework, and the VDAM fully manipulates the depth labels as a reliable indicator to generate more instructive intermediate features. Afterwards, we deliver the auxiliary feature supervision to release the cross-domain challenge and acquires efficient spatial knowledge migration with the channel-wise projection layer (pixel-level supervision) and adversarial scoring block (distribution-level supervision). After training, we reserve the primary detection network to inference tested images, with no external data and no extra inference computational cost. We summarize our contributions as follows:

- We propose a new framework **FD3D** for monocular 3D object detection, which sufficiently manipulates the image and annotation foreground object-wise depth map (AFOD) as input to focus on the foreground objects, thus to produce instructive intermediate features for further domain-free feature supervision.
- We develop a vision-depth association module to generate robust intermediate features, which projects features with semantic, depth, and geometric clues into 3D coordinates to deploy adequate feature fusion.
- We propose the auxiliary feature supervision to reach efficient pixel-level and distribution-level feature guidance with channel-wise projection layer and adversarial scoring block.

Our approach achieves state-of-the-art performance on the KITTI (Geiger, Lenz, and Urtasun 2012) and nuScenes (Caesar et al. 2019) datasets.

Related Work

Monocular 3D object detection. The objective of monocular 3D objection detection was to recognize objects of interest and recover the corresponding 3D bounding box information from monocular images. It was an ill-posed problem due to the lack of direct depth information measurements for solving 2D-3D projection ambiguity (Ma et al. 2021; Reading et al. 2021). Recent approaches (Zhang, Lu, and Zhou 2021; Reading et al. 2021; Park et al. 2021; Huang et al. 2022) adopted the convolutional neural networks to encode high-level semantic features from image inputs, and designed geometric constraints based on calibration projection or utilized additional depth supervision with LiDAR measuring to decode target-level responses. To mitigate the issue of depth measuring absence, PatchNet (Ma et al. 2020) adopted dense depth estimation pretraining (Fu et al. 2018) and performed the task of regression from patched depth maps. MonoDTR (Huang et al. 2022) proposed a depth-aware transformer to encode long-range semantic and depth dependencies. MonoDistill (Chong et al. 2022) utilizes the projected LiDAR signals as the inputs for the teacher model to educate the student model with spatial information. Our approach reveals the monocular detector to further motivate the auxiliary spatial feature supervision.

Auxiliary Learning. Auxiliary learning (Zhang, Tang, and Jia 2018; Liu, Davison, and Johns 2019; Ye et al. 2021) aimed at jointly training a primary task alongside auxiliary tasks to improve the primary model robustness to unseen data. Works (Flynn et al. 2016; Zhou et al. 2017) accomplished unsupervised monocular depth estimation via developing image synthesis networks that predicted the relative pose of multiple cameras for auxiliary learning. For the primary task of 2D object detection, Mordan et al. (Mordan et al. 2018) proposed the generic ROCK residual block to train auxiliary scene classification, depth estimation, and normal estimation. In terms of monocular 3D object detection, DD3D (Park et al. 2021) proposed to pre-train the detector with the auxiliary task of depth estimation to assist in monocular 3D localization. MonoCon (Liu, Xue, and Wu 2022) proposed to recover the 2D-3D relationship via applying geometric constraints, which trained the key-point estimations of foreground objects as the auxiliary task. Our auxiliary feature supervision can be regarded as an auxiliary learning task to improve the primary detection task.

Methodology

Overview

As shown in Fig. 2, we illustrate the overview of our FD3D framework for monocular 3D object detection. Firstly, we propose an instructive feature generation network (IFGN) by developing a vision-depth association module (VDAM), which considers the long-range dependencies of foreground object-wise depth maps (AFOD) from labels and their encoded semantic features. We supervise it with labels and restore the instructive intermediate features for feature supervision. Afterwards, we adopt MonoDLE (Ma et al. 2021) monocular detector as our baseline model, and train it for the primary task of 3D detection, together with the auxiliary feature supervision between intermediate features, where we design the channel-wise projection layer and an adversarial scoring block to promote spatial knowledge migration. After training, we propose the tested images into the primary detection network to receive detection results with no external data. Next, we will introduce our framework with equations compared with the previous ones, and reveal more details of our contributions.

Label Manipulation for 3D Object Detection

In the Fig. 3, we elaborate the label manipulation comparison. Firstly, we introduce the usual label manipulation: Fig. 3(a) with direct label supervision and Fig. 3(b) with multi-layer losses to realize hierarchical label supervision,

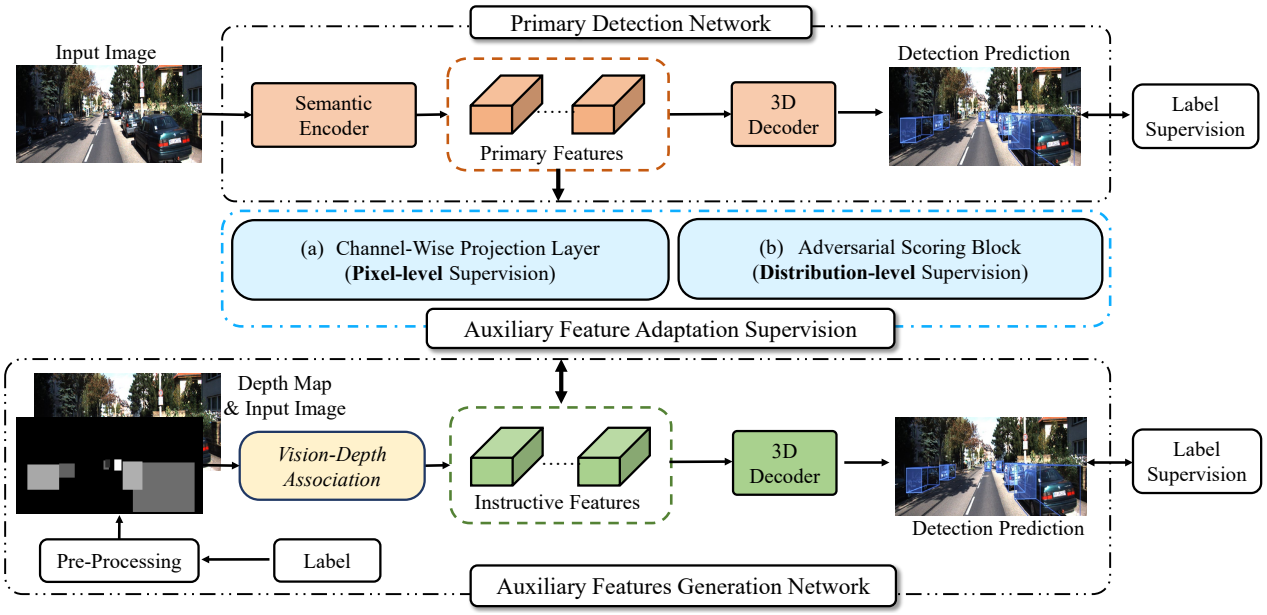

Figure 2: Overview of our FD3D framework for monocular 3D object detection.

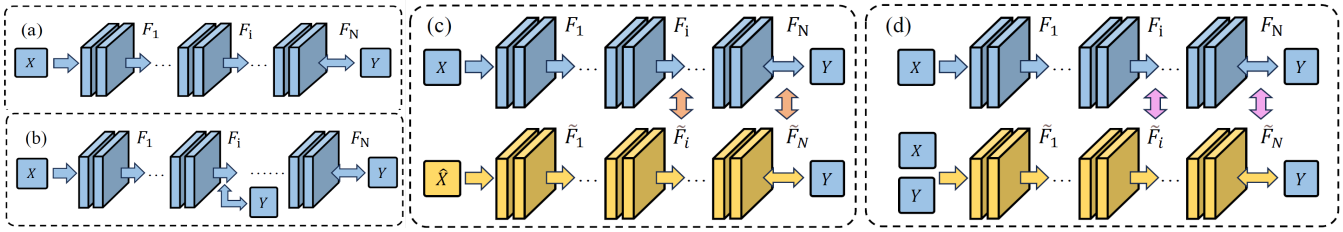

Figure 3: The usual label manipulation for monocular 3D object detection with the direct label (a) or hierarchical label (b) supervision. The distillation schema in (c) with other-modal input $\hat{X}$, which applies cross-domain intermediate feature supervision. Our strategy in (d) replaces the $\hat{X}$ with image X, adopts the similar monocular model $\tilde{F}$ and leverages the ground truth Y as a spatial indicator to achieve more instructive intermediate features.

where X denotes the monocular **image** inputs and F_i denotes the i-th layer of model.

$$\begin{aligned} &L_1(I(F_N;X);Y)+ \\ &L_{31}(I(G_N;\hat{X});Y)) + \sum L_{32}(O(I(F_i;X));I(G_i;\hat{X})) \\ =&L_1(...) + \sum L_3(I(F_i;X);I(T_i(G_N,\hat{X};Y),\hat{X})) \end{aligned} \tag{1}$$

Furthermore, $Equ.(1)$ reflects the distillation schema in Fig. 3(c) which trains the guidance network G in L_{31} with other-modal input $\hat{X}$ like LiDAR pointcloud, and applies the feature supervision loss L_{32} between intermediate features, where $I(A;B)$ denotes the inference result of model layers A with input B; $L(A;B)$ denotes the loss between prediction A with label B; O denotes the operation on features. Actually, we could simplify the L_{31} and L_{32} to the L_3 where $T(G_N,\hat{X};Y)$ indicates the manipulation (model training) process of the three elements $(G_N,\hat{X},Y)$, thus could achieve the i-th intermediate T_i features via the input $\hat{X}$. Considering the limitations of previous frameworks, we propose our stylish pipeline:

$$\begin{aligned} &L_1(I(F_N;X);Y) + L_{41}(I(\tilde{F}_N;X,Y);Y)) \\ &+\sum L_{42}(O(I(F_i;X));I(\tilde{F}_i;X,Y)) \\ =&L_1(...) + \sum L_4(I(F_i;X);I(T_i(\tilde{F}_N,X,Y;Y);X,Y)) \end{aligned} \tag{2}$$

As shown in Fig. 3 (d), we firstly replace the $\hat{X}$ with the same image domain input X and replace the other-modal-based guidance network G with the similar monocular model $\tilde{F}$ to release the domain gap. In addition, we delve into label manipulation where we leverage the ground truth Y as the strong geometrical clue indicator to achieve more instructive intermediate features $I(T_i;X,Y)$. We regard our approach as a sufficient approach to absorb and manipulate the three elements $(\tilde{F},X,Y)$, which decreases the training complexity.

Instructive Feature Generation Network

Prior to the details, we briefly introduce some important design considerations. First of all, we shall avoid the usage

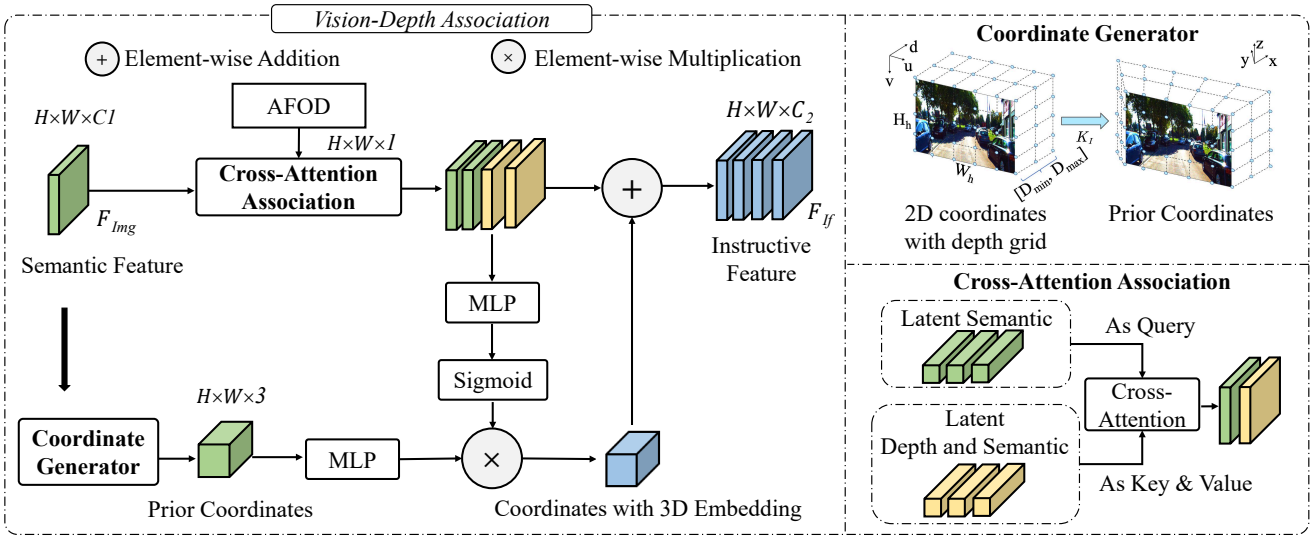

Figure 4: The structure of our vision-depth fusion module is only employed in our teacher model.

of additional inputs like LiDAR signals or paired stereo-images during the network training or inference, so that our results could be fairly compared against the prior monocular research. Secondly, noting that we aim at transferring the abundant spatial knowledge to the baseline, we shall ensure that the instructive intermediate features from the generation network are accurate and reliable. Thirdly, to alleviate the cost of performing dimensional alignments, despite the vision-depth association module, we maintain the generation network architecture as that of the baseline model.

Specifically, as shown in Fig. 4, we create the annotation foreground object-wise depth map (AFOD), denoted as $\mathcal{D}$, via applying the calibration projection of object-wise depth labels $\mathcal{A}$. We adopt the AFOD for its overall superior performance with clearer instructive features about the foreground objects. LiDAR depth reaches the ambiguous feature guidance on the long-distance and occluded cases, which is affected by the disruption of background noise, such as buildings and traffic devices. More details are demonstrated in the experiment section.

Then pixels within projected 2D boxes are assigned with center depth values of 3D boxes, while the occlusion areas are assigned with depth values of closer targets. As shown in Fig. 4, we designed the vision-depth association module based on the cross-attention association (Vaswani et al. 2017) and the coordinate embedding (Liu et al. 2022) mechanisms. While, instead of generating 3D coordinate feature maps with semantic features only, we further restrain the pixel localization error by multiplying it with encoded depth features from $\mathcal{D}$. We denote the image feature extractor as F_{Img}, the multi-layer-perceptron (MLP) operation as F_M, the cross-attention operation as ψ, and the coordinate generator as ϕ. Hence in function, we formulate the vision-depth association module as

$$\begin{aligned} F_{If} =& \sigma(F_M(\psi(F_{Img}(X), D)) \\ & \times F_M(\phi(X)) + \psi(F_{Img}(X), D), \end{aligned} \tag{3}$$

where F_{If} refers to the associated vision-depth features, i.e. our wanted instructive features, and σ refers to the sigmoid operation. Within the coordinate generator ϕ, we first create height and width arrays with lengths of the input image size, and leverage the LID distribution to create the depth array for each pixel location. Hence, we multiplied the camera-axis coordinates with the inverse of the intrinsic parameters, together with the camera to LiDAR extrinsic parameters to obtain our initial 3D coordinates of shape $H \times W \times 3$. Our cross-attention association sets the encoded semantic features as query and the concatenation of depth and semantic map as key and value, so that the generator has access to the precise information of distance measuring.

Afterwards, we train the IFGN with 3D labels in a supervised manner and evaluate the depth-measuring quality of instructive intermediate features based on its 3D detection performance on the validation split.

Auxiliary Feature Supervision

In this subsection, we are going through the auxiliary feature supervision process of instructive intermediate features. Unlike previous auxiliary learning approaches for monocular 3D object detection (Park et al. 2021, 2023; Peng et al. 2022) that require the pre-training of a large number of samples with dense depth map labels, we train our auxiliary supervision on the detection dataset only.

Channel-Wise Projection Layer (CPL) Motivated by the MLP-only architecture (Touvron et al. 2022), we conduct the channel-wise interaction in a residual manner with the MLP operations. As shown in Fig. 5, the proposed CPL consists of the residual structure with channel attention generation. Instead of directly adopting the proposed designs (Touvron et al. 2022), we replace the self-attention module with linear layers (Li.), affine transformation (Aff.) and other operations, which abandons the traditional multi-head attention computation and achieves GPU savings and stable training. Next, we deliver the global features through the channel attention generation, to obtain the channel attention weights.

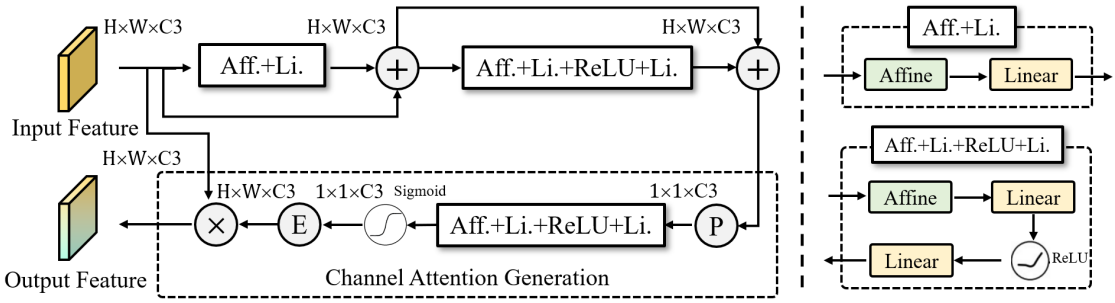

Figure 5: Overview of the channel-wise projection layer (CPL).

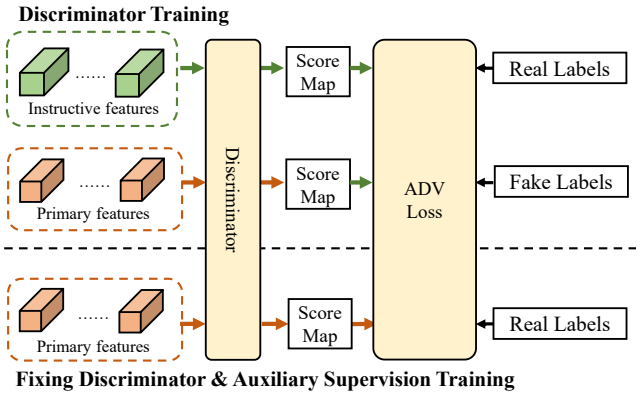

Figure 6: Training of the discriminator for the adversarial scoring block (ASB).

Therefore, the total pixel-wise loss $\mathcal{L}_{pix}$ can be denoted as:

$$\mathcal{L}_{pix} = \lambda_m^F \mathcal{L}_{mse}\Big(\mathrm{CPL}(F_S), F_T\Big), \tag{4}$$

where λ_m^F is the hyper-parameters for loss balancing the training. The MSE loss refers to the mean squared errors.

Adversarial Scoring Block The above MSE loss cares about the pixel-wise diversity, which may mislead the auxiliary supervision with the unbalanced feature distribution: a small partition of foreground object targets but a large partition of background (Chong et al. 2022). Therefore, we adopt the training procedure of generative adversarial network (Goodfellow et al. 2020), and propose the adversarial scoring block (ASB), as shown in Fig. 6. It allocates the distribution consistency between paired features, and leverages the discriminator network to distinguish them. When the primary features succeed in fooling the discriminator, we receive a similar distribution between the primary and instructive features. Specifically, we design the discriminator network with linear layers. When training the discriminator, we assign the real labels to the auxiliary instructive features and fake labels to the primary ones.

We minimize the binary cross-entropy loss $\mathcal{L}_D$ of the score map from discriminator D on the features as follows:

$$\begin{aligned}\mathcal{L}_D^F = -\frac{1}{N}\sum_i \sum_{h,w} & y_i \log\Big(\mathrm{D}(F_{in})^{(h,w)}\Big) \\ & + (1-y_i)\log\Big(\mathrm{D}(F_{pr})^{(h,w)}\Big)\end{aligned} \tag{5}$$

Method	mAP↑	mATE↓	mAOE↓	NDS↑
CenterNet	0.338	0.658	0.629	0.400
FCOS3D	0.358	0.690	0.452	0.428
DD3D	0.418	0.572	0.368	0.477
PETR	0.391	0.647	0.433	0.455
BEVFormer	0.409	0.650	0.439	0.462
BEVDet	0.398	0.556	0.414	0.463
Ours	**0.431**	**0.569**	**0.365**	**0.485**
Imp.	+1.3%	+0.3%	+0.3%	+0.8%

Table 1: Single-frame nuScenes detection test set evaluation. 'Imp.' indicates our performance improvement over the base model DD3D.

where y_i=1 when the discriminator input is instructive intermediate features F_{in}, and y_i=0 when the input reveals the primary features F_{pr}.

After the discriminator's training, we fix the discriminator and begin the auxiliary supervision training for the primary network. As shown in the bottom of Fig. 6, by constraining the score map from the auxiliary training and assigning real labels with binary cross-entropy loss, the auxiliary learning model can effectively generate appropriate features and prediction outputs with higher quality, which acts as one distribution agreement different from the pixel-wise MSE loss. The adversarial loss $\mathcal{L}_{adv}$ could be formulated as:

$$\mathcal{L}_{adv}^F = -\frac{1}{N}\sum_i \sum_{h,w} \log\Big(\mathrm{D}(F_{pr})^{(h,w)}\Big), \tag{6}$$

As a result, features from the primary network can fool the discriminator by maximizing the probability of the feature or prediction similarity.

Primary and Auxiliary Supervision

For brevity, we use $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$ to denote the detection loss. We first train our primary network with the loss:

$$\mathcal{L}_{pr} = \mathcal{L}_{cls} + \mathcal{L}_{reg} \tag{7}$$

Then, we train our auxiliary feature supervision with the pre-trained primary network and instructive feature generation network. To achieve end-to-end network training, within one batch training, we first train the discriminator with $\mathcal{L}_D^F$ (Equ. 5), then increase the auxiliary feature supervision (Equ. 8).

$$\mathcal{L}_S = \mathcal{L}_{pr} + \mathcal{L}_{pix} + \lambda_a^F \mathcal{L}_{adv}^F \tag{8}$$

Monocular Method	Extra Data	Time(ms)	AP_{3D} (Car test)			AP_{BEV} (Car test)		
			Easy	**Mod.**	Hard	Easy	**Mod.**	Hard
MonoDTR (Huang et al. 2022)	LiDAR	37	21.99	15.39	12.73	28.59	20.38	17.14
MonoDistill (Chong et al. 2022)	LiDAR	40	22.97	16.03	13.60	31.87	22.59	19.72
DCD (Li et al. 2022)	LiDAR	-	23.81	15.90	13.21	32.55	21.50	18.25
SGM3D (Zhou et al. 2022)	Stereo	30	22.46	14.65	12.97	31.49	21.37	18.43
OPA-3D (Su et al. 2023)	KITTI-Depth	40	24.60	17.05	14.25	33.54	22.53	19.22
NeurOCS (Min et al. 2023)	G.T. Fore. Mask	-	29.89	18.94	15.90	37.27	24.49	20.89
PGD (Wang et al. 2021b)	None	21	19.05	11.76	9.39	26.89	16.51	13.49
MonoDLE (Ma et al. 2021)	None	40	17.23	12.26	10.29	24.79	18.89	16.00
MonoEF (Zhou et al. 2021)	None	30	21.29	13.87	11.71	29.03	19.70	17.26
GUPNet (Lu et al. 2021)	None	34	20.11	14.20	11.77	-	-	-
HomoLoss (Gu et al. 2022)	None	-	21.75	14.94	13.07	29.60	20.68	17.81
MonoJSG (Lian, Li, and Chen 2022)	None	42	24.69	16.14	13.64	32.59	21.26	18.18
MonoCon (Liu, Xue, and Wu 2022)	None	26	22.50	16.46	13.95	31.12	22.10	19.00
Ours	None	40	**25.38**	**17.12**	**14.50**	**34.20**	**23.72**	**20.76**
Improvements	-	-	+0.69	+0.66	+0.55	+1.61	+1.62	+1.76

Table 2: Comparison on the KITTI test set. 'Improvements' indicates our performance gain over the previous best results without extra data. G.T. Fore. Mask denotes ground truth foreground mask (Min et al. 2023).

Method	Modal	AP_{3D}(Car val)			AP_{BEV}(Car val)		
		Easy	Mod.	Hard	Easy	Mod.	Hard
(i) MV3D (Chen et al. 2017)	LiDAR	71.19	56.60	55.30	86.18	77.32	76.33
(ii) Instructive Fea. Gen. Network	RGB + All LiDAR Depth	69.40	55.43	43.54	82.45	67.56	55.60
(iii) Instructive Fea. Gen. Network	RGB+ Foreground LiDAR Depth	69.50	55.57	43.60	82.58	67.64	55.58
(iv) **Instructive Fea. Gen. Network**	**RGB+ AFOD (Ours)**	**68.39**	**60.47**	**52.24**	**78.39**	**73.05**	**64.33**

Table 3: Evaluation of the instructive feature generation network on the KITTI validation split on 'Car' category. Our approach (iv) could achieve a comparable detection performance with the LiDAR-based detector (i) (Chen et al. 2017) and obtains better overall performance compared with LiDAR-depth-related (ii, iii).

where λ_a^F denote the loss balancing hyper-parameters of the adversarial loss.

Experiments

Settings

Dataset. We evaluate our approach on two widely used datasets: KITTI (Geiger, Lenz, and Urtasun 2012) and nuScenes (Caesar et al. 2020) benchmarks. The KITTI dataset consists of 7,481 samples for training and 7,518 for testing. Following (Reading et al. 2021), we divide training samples into a training set with 3,712 samples and a validation set with 3,769 samples. Ablation studies are all conducted on the validation split with models trained on the training split. There are three object classes (Car, Pedestrian and Cyclist) and each class is divided into three difficulty levels based on occlusion, truncation and size. The large-scale dataset nuScenes (Caesar et al. 2020) contains a full 360-degree field of view provided by 6 cameras, 1 Lidar and 5 radars, which consists of 1000 driving scenes, with 700, 150 and 150 scenes for training, validation, and testing, respectively. The corresponding sequences are sampled to frames with the resolution of 1600×900 at 2Hz.

Evaluation metric. For KITTI dataset, following prior works (Reading et al. 2021), the 3D Average Precision (AP_{3D}) and BEV Average Precision (AP_{BEV}) are two vital evaluation metrics. They are calculated using class-specific thresholds with 40 recall positions based on the intersection-over-union (IoU) of 2D BEV and 3D bounding boxes. The Car, Pedestrian and Cyclist categories have 0.7, 0.5, 0.5 IoU threshold. For nuScenes, we follow (Park et al. 2021) and adopt the evaluation metrics including nuScenes Detection Score (NDS) and mean Average Precision (mAP), along with two true-positive metrics ATE and AOE.

Implementation details. We select monocular 3D detection method MonoDLE (Ma et al. 2021) as our base model for KITTI dataset following (Chong et al. 2022), and DD3D (Park et al. 2021) as the base model for nuScenes dataset, which both reveal one-stage center-based detection approaches. The weight performs λ_m^F=0.9 for pixel-wise loss, and λ_a^F=0.9 for the adversarial loss. The settings for the optimizer and batch size follow base models (Ma et al. 2021; Park et al. 2021).

Evaluation of Our Framework

Comparisons on KITTI dataset. In Table 2, we present the benchmark evaluation on the KITTI test split. Compared with the previous best results without extra data, our framework outperforms it with a certain margin. Furthermore, our framework realizes the inference time of $40ms$, which does not introduce additional computational costs in the inference stage and is industrially implementable.

Ablation	AP_{3D}@IoU=0.7			AP_{BEV}@IoU=0.7		
	Easy	Mod.	Hard	Easy	Mod.	Hard
(i) direct.	60.57	57.32	49.50	71.00	67.68	59.30
(ii) cro.	61.31	57.45	49.61	71.72	67.69	59.75
(iii) +coord.	62.03	58.54	50.10	72.39	68.72	60.83
(iv) +coord.$\times\sigma(\mathcal{F}_M(vis.))$	66.58	63.52	54.82	77.49	75.40	63.32
(v) +coord.$\times\sigma(\mathcal{F}_M(dep.))$	67.38	62.97	54.03	78.91	76.06	64.64
(vi) +coord.$\times\sigma(\mathcal{F}_M(cro.))$	**73.48**	**67.10**	**56.99**	**83.93**	**78.76**	**69.39**

Table 4: Ablation study of different instructive feature generation network designs on the KITTI validation set. 'vis.' denotes the visual features. 'dep.' denotes the object-wise annotation depth. 'direct.' denotes the direct association from the MLP layer with concatenation input of 'vis.' and 'dep.'. 'cro.' denotes the our cross-attention association strategy. 'coord.' denotes the preset 3D coordinates. $\mathcal{F}_M$ and σ denotes the MLP and sigmoid operation.

	Feature	C-P	A-S	AP_{3D}@IoU=0.7			AP_{BEV}@IoU=0.7		
				Easy	Mod.	Hard	Easy	Mod.	Hard
(i)	-	-	-	17.45	13.66	11.68	24.97	19.33	17.01
(ii)	✓	✓	-	26.78	19.43	16.41	35.79	26.21	22.71
(iii)	✓	-	✓	24.72	18.16	15.47	34.48	24.47	21.16
(iv)	✓	✓	✓	**28.22**	**20.23**	**17.04**	**36.98**	**26.77**	**23.16**

Table 5: Ablation study for auxiliary feature supervision. 'C-P' denotes the channel-wise projection layer for pixel-wise loss. 'A-S' denotes the adversarial scoring block for distribution-level loss.

Comparisons on nuScenes dataset. In Table 1, we compare with monocular approaches CenterNet (Zhou, Wang, and Krähenbühl 2019), FCOS3D (Wang et al. 2021a) and DD3D (Park et al. 2021) on the nuScenes dataset, where our approach outperforms the base model DD3D with a 1.3% improvement in mAP and 0.8% improvement in NDS.

Evaluation of the instructive feature generation network. The generation network should achieve adequate detection accuracy to ensure reliable instructive intermediate features. As shown in Table 3, we conduct the evaluation comparison with (i) the LiDAR-based method MV3D (Chen et al. 2017), and (ii-iv) the IFGN with all LiDAR depth, foreground LiDAR depth and our AFOD as input. Our approach (iv) could achieve a comparable detection performance with the LiDAR-based detector (i) and generates clearer guidance features on the long-distance and occluded objects, i.e. the 'Mod.' and 'Hard' cases. The settings (ii) and (iii) gain higher score on the 'Easy' cases with denser and more refined LiDAR depth clues, but receive the mixed and ambiguous instructive features on the long-distance and occluded cases. In addition, the full-range LiDAR depth (ii) is disturbed by the background objects like buildings and road devices. Overall, our (iv) obtains better overall performance.

Ablation Study

In this section, we investigate the effects of each component of our framework on the KITTI validation split.

Ablation study for instructive feature generation network. We explore some different designs of the instructive feature generation network, specifically for the prior coordinates with 3D embedding, as shown in Table 4. Firstly, the cross-attention association (setting (ii)) performs better compared to the direct association of concatenation and MLP (setting (i)), for the long-range dependencies consideration. Based on setting (ii), we add the fixed 3D grid coordinates, yet have a limited improvement (setting (iii)), as the fixed 3D grid is a presetting schema and lacks the content-aware and geometry-aware bias. If we insert the visual information (setting (iv)) or depth knowledge (setting (v)), as the attention weight, into the originally fixed coordinates, the performance advances. We hence inject both contents simultaneously (setting (vi)) and achieve our strongest features generation network for producing instructive intermediate features.

Ablation study for auxiliary feature supervision. As shown in Table 5, we investigate our designs of auxiliary feature supervision: the channel-wise projection layer (CPL) and adversarial scoring block (ASB). Setting (ii) proves the effects of our CPL to enhance the pixel-wise guidance. Setting (iii) also improves the baseline, but it performs sub-optimal results without applying the pixel-wise constraint. We hence further jointly take both supervisions, and it turns to generate convincing improvements in setting (iv).

Conclusion

In this paper, we propose a new monocular 3D object detection framework named ***FD3D***, which develops high-quality instructive intermediate features to conduct auxiliary feature supervision with only the image and annotation foreground object-wise depth map (AFOD) as input. To obtain representative instructive features with depth-positional cues, we develop a vision-depth association within the generation network that interacts with the AFOD with semantic features to realize long-range aware embedding. We proceed with auxiliary feature supervision from both the pixel and distribution levels. Our pipeline is shown effective and efficient on KITTI and nuScenes datasets.

References

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2019. nuScenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A multimodal dataset for autonomous driving. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 11621–11631. IEEE.

Chen, X.; Ma, H.; Wan, J.; Li, B.; and Xia, T. 2017. Multi-view 3d object detection network for autonomous driving. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1907–1915.

Chen, Y.-N.; Dai, H.; and Ding, Y. 2022. Pseudo-Stereo for Monocular 3D Object Detection in Autonomous Driving. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 887–897.

Chong, Z.; Ma, X.; Zhang, H.; Yue, Y.; Li, H.; Wang, Z.; and Ouyang, W. 2022. MonoDistill: Learning spatial features for monocular 3D object detection. In *Int. Conf. on Learning Representations (ICLR)*.

Flynn, J.; Neulander, I.; Philbin, J.; and Snavely, N. 2016. DeepStereo: Learning to predict new views from the world's imagery. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 5515–5524.

Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; and Tao, D. 2018. Deep ordinal regression network for monocular depth estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2002–2011.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 3354–3361. IEEE.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Gu, J.; Wu, B.; Fan, L.; Huang, J.; Cao, S.; Xiang, Z.; and Hua, X.-S. 2022. Homography Loss for Monocular 3D Object Detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Huang, K.-C.; Wu, T.-H.; Su, H.-T.; and Hsu, W. H. 2022. MonoDTR: Monocular 3D Object Detection with Depth-Aware Transformer. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Li, P.; and Zhao, H. 2021. Monocular 3d detection with geometric constraint embedding and semi-supervised training. *IEEE Robotics and Automation Letters (RA-L)*, 6(3): 5565–5572.

Li, Y.; Chen, Y.; He, J.; and Zhang, Z. 2022. Densely Constrained Depth Estimator for Monocular 3D Object Detection. In *European Conf. on Computer Vision (ECCV)*. Springer.

Lian, Q.; Li, P.; and Chen, X. 2022. MonoJSG: Joint Semantic and Geometric Cost Volume for Monocular 3D Object Detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1070–1079.

Liu, S.; Davison, A.; and Johns, E. 2019. Self-supervised generalisation with meta auxiliary learning. In *Conf. and Workshop on Neural Information Processing Systems (NeurIPS)*, 1677–1687.

Liu, X.; Xue, N.; and Wu, T. 2022. Learning auxiliary monocular contexts helps monocular 3D object detection. In *AAAI Conf. on Artificial Intell. (AAAI)*, 1810–1818.

Liu, Y.; Wang, T.; Zhang, X.; and Sun, J. 2022. PETR: Position Embedding Transformation for Multi-View 3D Object Detection. In *European Conf. on Computer Vision (ECCV)*.

Liu, Z.; Wu, Z.; and Tóth, R. 2020. SMOKE: Single stage monocular 3D object detection via keypoint estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 996–997.

Lu, Y.; Ma, X.; Yang, L.; Zhang, T.; Liu, Y.; Chu, Q.; Yan, J.; and Ouyang, W. 2021. Geometry uncertainty projection network for monocular 3D object detection. In *IEEE Int. Conf. on Computer Vision (ICCV)*, 3111–3121.

Ma, X.; Liu, S.; Xia, Z.; Zhang, H.; Zeng, X.; and Ouyang, W. 2020. Rethinking pseudo-lidar representation. In *European Conf. on Computer Vision (ECCV)*, 311–327. Springer.

Ma, X.; Zhang, Y.; Xu, D.; Zhou, D.; Yi, S.; Li, H.; and Ouyang, W. 2021. Delving into localization errors for monocular 3D object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 4721–4730.

Min, Z.; Zhuang, B.; Schulter, S.; Liu, B.; Dunn, E.; and Chandraker, M. 2023. NeurOCS: Neural NOCS Supervision for Monocular 3D Object Localization. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 21404–21414.

Mordan, T.; Thome, N.; Henaff, G.; and Cord, M. 2018. Revisiting multi-task learning with rock: a deep residual auxiliary block for visual detection. In *Conf. and Workshop on Neural Information Processing Systems (NeurIPS)*, 1317–1329.

Nabati, R.; and Qi, H. 2021. CenterFusion: Center-based radar and camera fusion for 3d object detection. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 1527–1536.

Park, D.; Ambrus, R.; Guizilini, V.; Li, J.; and Gaidon, A. 2021. Is pseudo-lidar needed for monocular 3D object detection? In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 3142–3152.

Park, D.; Li, J.; Chen, D.; Guizilini, V.; and Gaidon, A. 2023. Depth is all you need for monocular 3d detection. In *IEEE Int. Conf. on Robotics and Automation (ICRA)*.

Peng, L.; Liu, F.; Yu, Z.; Yan, S.; Deng, D.; Yang, Z.; Liu, H.; and Cai, D. 2022. Lidar point cloud guided monocular 3d object detection. In *European Conf. on Computer Vision (ECCV)*, 123–139. Springer.

Reading, C.; Harakeh, A.; Chae, J.; and Waslander, S. L. 2021. Categorical depth distribution network for monocular 3D object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 8555–8564.

Su, Y.; Di, Y.; Zhai, G.; Manhardt, F.; Rambach, J.; Busam, B.; Stricker, D.; and Tombari, F. 2023. OPA-3D: Occlusion-aware pixel-wise aggregation for monocular 3d object detection. *IEEE Robotics and Automation Letters (RA-L)*, 8(3): 1327–1334.

Sun, J.; Chen, L.; Xie, Y.; Zhang, S.; Jiang, Q.; Zhou, X.; and Bao, H. 2020. Disp R-CNN: Stereo 3D object detection via shape prior guided instance disparity estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 10548–10557.

Thomas, H.; Qi, C. R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; and Guibas, L. J. 2019. KPConv: Flexible and deformable convolution for point clouds. In *IEEE Int. Conf. on Computer Vision (ICCV)*, 6411–6420.

Touvron, H.; Bojanowski, P.; Caron, M.; Cord, M.; El-Nouby, A.; Grave, E.; Izacard, G.; Joulin, A.; Synnaeve, G.; Verbeek, J.; et al. 2022. ResMLP: Feedforward networks for image classification with data-efficient training. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Conf. and Workshop on Neural Information Processing Systems (NeurIPS)*, 30.

Wang, T.; Zhu, X.; Pang, J.; and Lin, D. 2021a. FCOS3D: Fully convolutional one-stage monocular 3d object detection. In *IEEE Int. Conf. on Computer Vision (ICCV)*, 913–922.

Wang, T.; Zhu, X.; Pang, J.; and Lin, D. 2021b. Probabilistic and geometric depth: Detecting objects in perspective. In *Conf. on Robot Learning (CoRL)*. PMLR.

Ye, J.; Batra, D.; Das, A.; and Wijmans, E. 2021. Auxiliary tasks and exploration enable objectgoal navigation. In *IEEE Int. Conf. on Computer Vision (ICCV)*, 16117–16126.

Yin, T.; Zhou, X.; and Krahenbuhl, P. 2021. Center-based 3d object detection and tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 11784–11793.

Zhang, Y.; Lu, J.; and Zhou, J. 2021. Objects are Different: Flexible monocular 3D object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 3289–3298.

Zhang, Y.; Tang, H.; and Jia, K. 2018. Fine-grained visual categorization using meta-learning optimization with sample selection of auxiliary data. In *European Conf. on Computer Vision (ECCV)*, 233–248.

Zhou, T.; Brown, M.; Snavely, N.; and Lowe, D. G. 2017. Unsupervised learning of depth and ego-motion from video. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1851–1858.

Zhou, X.; Wang, D.; and Krähenbühl, P. 2019. Objects as points. *arXiv preprint arXiv:1904.07850*.

Zhou, Y.; He, Y.; Zhu, H.; Wang, C.; Li, H.; and Jiang, Q. 2021. Monocular 3D object detection: An extrinsic parameter free approach. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 7556–7566.

Zhou, Z.; Du, L.; Ye, X.; Zou, Z.; Tan, X.; Zhang, L.; Xue, X.; and Feng, J. 2022. SGM3D: stereo guided monocular 3d object detection. *IEEE Robotics and Automation Letters (RA-L)*, 7(4): 10478–10485.