Attention-Based Depth Distillation with 3D-Aware Positional Encoding for Monocular 3D Object Detection

Zizhang Wu[1*], Yunzhe Wu[1*], Jian Pu[2], Xianzhi Li[3], Xiaoquan Wang[1]

[1] ZongmuTech
[2] Fudan University
[3] Huazhong University of Science and Technology
wuzizhang87@gmail.com, {nelson.wu, rockywin.wang}@zongmutech.com, jianpu@fudan.edu.cn, xzli@hust.edu.cn

Abstract

Monocular 3D object detection is a low-cost but challenging task, as it requires generating accurate 3D localization solely from a single image input. Recent developed depth-assisted methods show promising results by using explicit depth maps as intermediate features, which are either pre-computed by monocular depth estimation networks or jointly evaluated with 3D object detection. However, inevitable errors from estimated depth priors may lead to misaligned semantic information and 3D localization, hence resulting in feature smearing and suboptimal predictions. To mitigate this issue, we propose **ADD**, an **A**ttention-based **D**epth knowledge **D**istillation framework with 3D-aware positional encoding. Unlike previous knowledge distillation frameworks that adopt stereo- or LiDAR-based teachers, we build up our teacher with identical architecture as the student but with extra ground-truth depth as input. Credit to our teacher design, our framework is seamless, domain-gap free, easily implementable, and is compatible with object-wise ground-truth depth. Specifically, we leverage intermediate features and responses for knowledge distillation. Considering long-range 3D dependencies, we propose *3D-aware self-attention* and *target-aware cross-attention* modules for student adaptation. Extensive experiments are performed to verify the effectiveness of our framework on the challenging KITTI 3D object detection benchmark. We implement our framework on three representative monocular detectors, and we achieve state-of-the-art performance with no additional inference computational cost relative to baseline models. Our code is available at https://github.com/rockywind/ADD.

Introduction

3D object detection is a primary foundation in 3D perception and has made evident contributions to autonomous driving, robot navigation and augmented reality. To reach accurate 3D localization, some methods use LiDAR sensors (Shi, Wang, and Li 2019; Lang et al. 2019; Shi et al. 2020; He et al. 2020) or stereo cameras (Wang et al. 2019b; Li, Chen, and Shen 2019; Chen et al. 2020; Sun et al. 2020) to provide accurate range measurements. However, cost and complexity increases caused by using extra sensors result in limited generalization in downstream tasks. To solve these
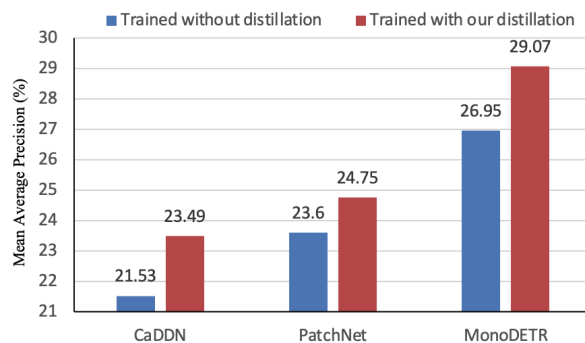
Figure 1: Moderate level BEV average precision (%) on different baselines with (red bar) or without (blue bar) training the networks with our distillation framework on the KITTI validation split. Significant improvements are observed on three representative networks. More comparison results are presented in the Experiment Section.

problems, the community has begun to pay more attention to monocular 3D object detection, which is much lower-cost and easily implementable.

To address the task of monocular 3D object detection, most deep-learning-based works adopt 2D-3D correspondences (Shi et al. 2021; Lu et al. 2021; Zhang, Lu, and Zhou 2021; Li et al. 2020; Liu et al. 2021; Liu, Xue, and Wu 2022; Chen et al. 2021a; Lian, Li, and Chen 2022) to constrain the learning of monocular images and directly estimate 3D bounding boxes. However, because of the lack of explicit 3D measurements, their localization accuracy degenerates dramatically compared with LiDAR- or stereo-based methods. To address this problem, some prior works (Wang et al. 2019b; Ma et al. 2019, 2020) take monocular depth estimation networks (Fu et al. 2018) to precompute depth estimations as extra inputs. Some other works (Park et al. 2021; Reading et al. 2021; Zhang et al. 2022) take depth estimation as an auxiliary task to introduce depth-aware features for object detection. However, inevitable depth estimation errors result in easily affected high-level features, thus leading to feature smearing and suboptimal predictions.

To mitigate this issue, we propose an attention-based depth knowledge distillation (KD) framework with 3D-aware positional encoding (PE) to provide students with

mighty and **dynamic** depth-positional cues. Unlike previous knowledge distillation frameworks that adopt stereo- or LiDAR-based teachers, we build up our teacher with identical architecture as depth-assisted student but with extra ground-truth (GT) depth as input. Leveraging teachers based on GT depth helps provide students with extra regularization to alleviate detection overfitting caused by depth estimation errors. Particularly, our framework is compatible with both object- and pixel-wise GT depth for teacher building up, i.e., not necessarily relying on extra LiDAR support.

Specifically, we conduct knowledge distillation on both intermediate features and network responses and leverage attention-based transformer modules to capture long-range 3D dependencies during student adaptation. Firstly, we design a *3D-aware self-attention module* to encode student intermediate features. Motivated by using PE for depth and 3D embedding (Huang et al. 2022; Liu et al. 2022a,b), to involve stronger hints for 3D reasoning, we integrate semantic features with precise 3D PE from GT depth. Credit to our KD-based framework, our PE in adaptation modules is discarded during inference, thus not bothering our framework's fairness relative to prior monocular detectors. Secondly, for encoder-decoder detecting heads (Carion et al. 2020), we adopt a *target-aware cross-attention module* for response adaptation, which revolves student object queries under the guidance of depth-aware teacher queries.

Compared with MonoDistill (Chong et al. 2022), whose teacher and student models are based on different data modalities, our framework is domain-gap free, as our teachers originate from monocular detectors and hence are easily implementable. Compared with Pseudo-stereo (Chen, Dai, and Ding 2022), which takes an advanced stereo-based network (Guo et al. 2021b) as a teacher, our teachers are easily accessible and effective as well. Besides, architectural differences between stereo and monocular detectors result in potentially spatially unaligned features and responses, while our framework is seamless with guaranteed feature- and output-level consistency.

Down to the aforementioned advantages, we successfully implement our framework on three representative monocular 3D object detection methods, including MonoDETR (Zhang et al. 2022), CaDDN (Reading et al. 2021) and PatchNet (Ma et al. 2020) using the challenging KITTI 3D objection benchmark (Geiger, Lenz, and Urtasun 2012). The performance improvements are shown in Figure 1. Overall, we summarize our contributions as follows:

(i) We propose a seamless KD framework for depth-assisted monocular 3D object detectors. It injects mighty and dynamic depth-positional cues into students to boost their performances. Our teachers are easily accessible and effective with guaranteed feature- and output-level consistency. Our teacher-student framework is domain gap-free and hence easily implementable.

(ii) We propose a 3D-aware self-attention module for the adaptation of multi-level intermediate features by integrating GT depth guided PE with student semantic features, as well as a target-aware cross-attention module for the adaptation of decoding head (Carion et al. 2020) re-

sponses. Our attention-based adaptation modules benefit from capturing long-range 3D dependencies.

(iii) We conduct extensive experiments to show the effectiveness of our framework on three representative baselines using the challenging KITTI 3D object detection benchmark (Geiger, Lenz, and Urtasun 2012). Our model achieves state-of-the-art performance with no extra inference computational cost relative to the baselines.

Related Work

Monocular 3D Object Detection. Monocular 3D object detection takes as input only a single image, aiming at identifying objects of interest and localizing their 3D bounding boxes. To alleviate the ambiguity in 2D-3D projection caused by the lack of accurate 3D measurements, existing approaches either leverage neural networks to design semantic and geometric constraints or adopt external depth information for depth reasoning to facilitate 3D object detection. For methods without depth assistance, Mousavian et al. (Mousavian et al. 2017) solved constraints between 2D bounding boxes and 3D dimensions to recover the 3D location. CenterNet (Zhou, Wang, and Krähenbühl 2019) proposed a centerness-based object detection paradigm that adds extra 3D task heads to lift the 2D detector to 3D space. M3D-RPN (Brazil and Liu 2019) proposed 3D anchors by aligning 2D anchors with 3D statistics. Instead of directly regressing object-wise depth, some methods estimated 2D and 3D heights with uncertainty modeling to recover 3D locations based on geometric priors (Shi et al. 2021; Lu et al. 2021). Several works (Li et al. 2020; Liu et al. 2021; Liu, Xue, and Wu 2022; Chen et al. 2021a) proposed keypoint-based approaches to limit the searching space of geometric constraints. Recently, MonoJSG (Lian, Li, and Chen 2022) designed a cost volume from semantic and geometric cues to model the depth error in a two-stage manner. However, an obvious drawback of the above methods is the decayed performance in 3D localization due to the lack of direct depth guidance.

On the other hand, for methods with depth assistance, some works (Ma et al. 2020; Ding et al. 2020; Shi, Chen, and Kim 2020) accomplished monocular 3D object detection by directly taking estimated depth maps as extra inputs. DD3D (Park et al. 2021) leveraged GT depth maps for depth pretraining. MonoDTR (Huang et al. 2022) generated intermediate pixelwise depth estimations to provide positional encoding to a depth-aware transformer. Other works (Wang et al. 2019b; Weng and Kitani 2019; Ma et al. 2019; Reading et al. 2021) jointly employed the estimated depth maps and intrinsic and extrinsic calibrations to project semantic features from monocular images into 3D space in a point cloud or voxel manner. However, the performance of the above depth-assisted methods is greatly affected by the accuracy of the estimated depth. Unfortunately, it is nontrivial to estimate an accurate depth from only 2D monocular images.

Knowledge Distillation. Knowledge distillation (KD) was first developed for model compression (Hinton et al. 2015), which makes student networks learn from both GT labels and soft labels generated from teacher networks.
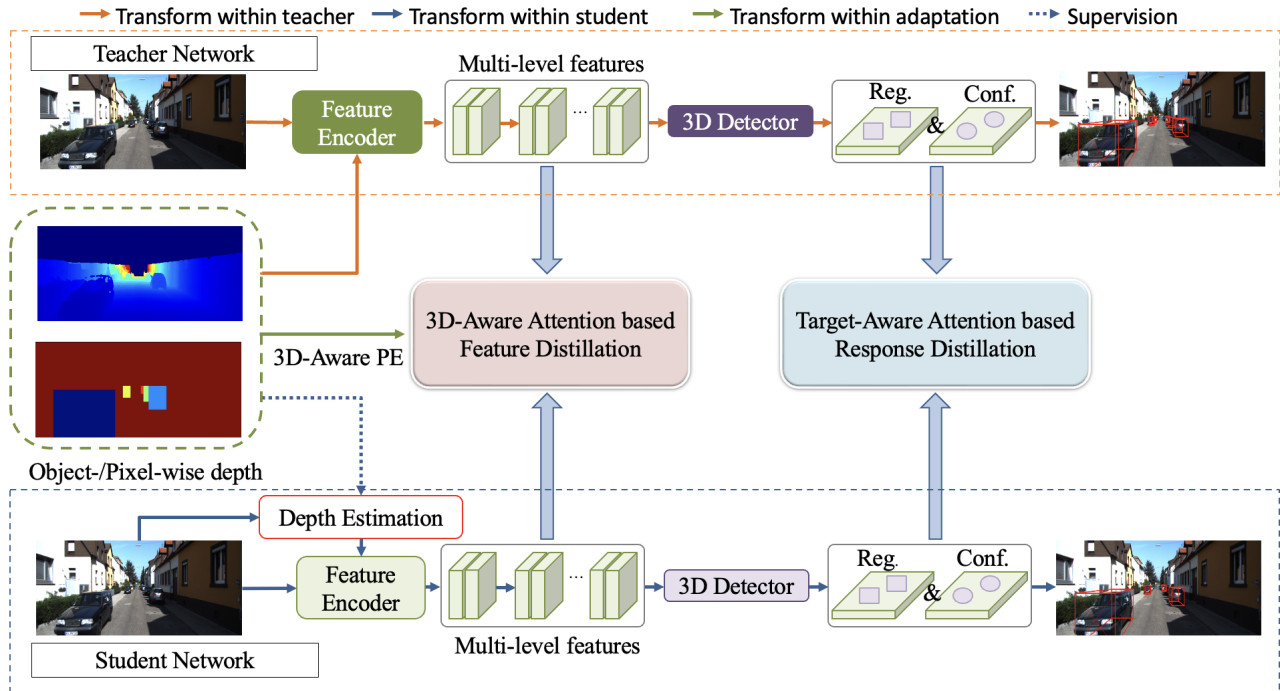
Figure 2: An overview of the proposed attention-based depth distillation framework pipeline. Firstly, we process and send object-/pixel-wise GT depth and monocular images into teacher network, whose architecture is identical to the student, and train it independently. Secondly, we load and freeze teacher parameters, and leverage distillations on features and responses to supervise the student. Finally, we use student network without teacher parameters for inference.

Romero et al. (Romero et al. 2014) showed that student networks can also be guided by features from intermediate layers. Afterwards, many methods successfully adopted knowledge distillation in various tasks, such as image classification (Heo et al. 2019; Tung and Mori 2019; Yim et al. 2017; Fukuda et al. 2017), semantic segmentation (Jiao et al. 2019; Liu et al. 2019) and depth estimation (Chen et al. 2021b; Hu et al. 2021), etc.

In the field of 2D object detection, Chen et al. (Chen et al. 2017) first implemented knowledge distillation by distilling features from neck, classification head and regression head. To tackle the problem of foreground-background balancing, some works (Li, Jin, and Yan 2017; Wang et al. 2019a) either mimic areas from the region proposal or distill the student with fine-grained features from foreground object regions. Guo et al. (Guo et al. 2021a) showed that distillation on regions excluding objects is also important. In 3D object detection, LiGA-Stereo (Guo et al. 2021b) guided the learning of a stereo-based detector by geometric-aware high-level features from a LiDAR-based detector to mitigate the influence of depth estimation error. Recently, MonoDistill (Chong et al. 2022) designed a teacher based on inputs of a projected LiDAR signal to educate a monocular 3D detector with spatial cues. Rather than using a LiDAR-based detector as a teacher, Chen et al. (Chen, Dai, and Ding 2022) proposed distilling a monocular 3D detector from an advanced stereo 3D detector as there is a smaller domain gap in image-to-image than image-to-LiDAR. Unlike previous

works, our network comprises a series of novel distillation techniques on intermediate features and responses to transfer depth knowledge from the teacher network that is trained using GT depth maps to the student network, and greatly improves the 3D detection performance compared with the baselines without knowledge distillation. In particular, our method takes no extra stereo or LiDAR data as input, thus expanding its application scope.

Methodology

As shown in Figure 2, we design our attention-based depth distillation framework with three main components: (i) a depth-assisted monocular Student Network that takes GT depth for supervision, (ii) an aligned effective (see Table 3) Teacher Network with identical architecture as the student that takes GT depth as extra input, and (iii) distillation branches between them. We adopt knowledge distillation on both intermediate features and network responses, where we propose *3D-aware self-attention* and *target-aware cross-attention* for feature adaptation, respectively.

General Depth-Assisted Model

First, we briefly review the general pipeline for depth-assisted monocular 3D object detectors. Here, we denote I_M as the monocular image input. To accomplish object classification and 3D bounding box regression over all possible targets, the depth-assisted models estimate a corresponding

depth map $\hat{D}_M$ from I_M and integrate it with monocular semantic features. Formally, the model can be formulated as:

$$\hat{D}_M = G_{Dep}(I_M),$$
$$F_{Cls}, F_{Box} = G_D(G_E(T(I_M, \hat{D}_M))), \quad (1)$$

where G_{Dep} denotes a monocular depth estimation network that either predicts depth distributions or continuous depth values; T denotes a data representation transformation that may project monocular features into 3D space or sustain them in a 2D manner; G_E denotes a feature encoding architecture to fuse semantic and depth features; G_D is a decoding head for generating target predictions from high-level features; F_{Cls} and F_{Box} are detection responses for classification and regression, respectively.

Note that the accuracy of $\hat{D}_M$ directly affects the final object prediction performance, and different methods usually utilize different designs of G_{Dep}. Specifically, some works (Ma et al. 2019; Ding et al. 2020; Ma et al. 2020) directly take estimations from pretrained monocular depth estimators as $\hat{D}_M$. CaDDN (Reading et al. 2021) jointly learns a depth distribution estimator and a 3D detector, and MonoDETR (Zhang et al. 2022) learns depth-aware feature maps by using auxiliary depth supervision. However, there is still a significant gap between the predicted and ground-truth depths. Therefore, inevitable errors in depth priors result in easily affected high-level features, thus leading to sub-optimal predictions. To mitigate this problem, we propose a KD framework by using a teacher model equipped with precise 3D measurements to teach a student model for reliable depth estimation, thus facilitating accurate 3D object detection. Below, we shall discuss our considerations when designing such a KD-based framework.

Knowledge Distillation Considerations

For these depth-assisted detectors, we believe that they can show stronger localization ability if we could inject them with **mighty** and **dynamic** depth-positional cues. Hence, intuitively, leveraging the KD mechanism with a depth-aware teacher is expected to be beneficial. Before introducing our KD framework, we first discuss the relevant design considerations that we have taken.

(i) Considering the difficulties and costs in 3D data collection, especially from erroneous multi-sensor alignment, we cannot introduce extra data requirements when designing teacher-student models. For example, we should avoid using expensive LiDAR assistance when training the teacher model.

(ii) Second, although introducing the dark knowledge from different domains benefits KD learning, it causes greater hardship for efficient implementation. Thus, a desired framework needs to consider data distribution consistency between teacher and student data usages.

(iii) Third, to develop a seamless KD framework, we need to ensure both feature- and output-level consistency in the teacher-student model. Otherwise, it takes extra costs and complexity to deal with spatially-unaligned intermediate features or responses.
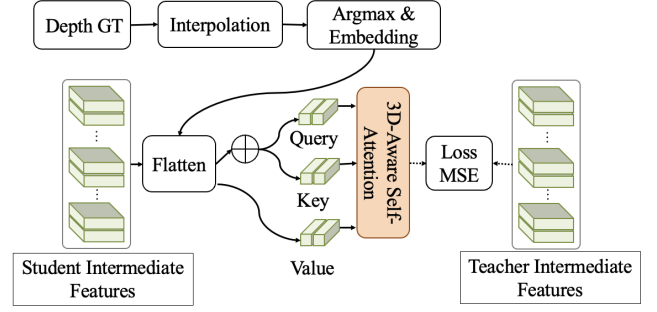

Figure 3: The visualization of our KD on multi-level intermediate features with 3D-aware positional encoding for feature adaptation.

(iv) Besides, it is usual to take teacher detectors with advanced designs for performance boosting. However, a complex architecture is not flexible, which limits the generalization of such KD methods. Thus, we should use easily accessible architectures for teacher design.

(v) To make effective feature approximation, it is important to adopt adaptation modules on student outcomes (Wang et al. 2019a) before taking distillation losses with teachers. We believe that leveraging transformer feature adaptation modules based on attention mechanisms will contribute to 3D information reasoning as they could consider long-range 3D dependencies.

3D-Aware Attention Based Feature Distillation

Our feature-level distillation pipeline is shown in Figure 3. We define different levels of student and teacher intermediate features obtained from convolutional neural network (CNN) backbones as $\{F^k\}_{k \in (1,...,n)}$ and $\{T^k\}_{k \in (1,...,n)}$ respectively, where n is the total number of intermediate feature levels. Besides, we denote the depth GT as $\mathcal{D}$. It is important to consider the information imbalance between foreground and background regions when adopting KD on detection tasks. We integrate features from these two types of regions by introducing binary masks M based on 2D bounding boxes B^k:

$$M^k = \{M_{i,j}^k | M_{i,j}^k = \mathbb{I}\left[(i,j) \in B^k\right]\}, \quad (2)$$

where i, j varies over the width and length of F^k, and $M_{i,j}^k$ equals one if position i, j of feature map F^k belongs to the foreground but zero otherwise.

3D-Aware Position Encoding. Inspired by the leverage of positional encoding (PE) for depth and 3D embedding (Huang et al. 2022; Liu et al. 2022a,b) with estimated depth and discretized meshgrid coordinates, credit to our KD-based framework, we are able to introduce semantic features with more accurate 3D positional encoding under the guidance of depth GT. Specifically we generate 3D-aware positional encodings F_{emb} under the guidance of $\mathcal{D}$, and introduce F_{emb} to self-attention operations on student features. We thus adopt a multi-layer perceptron (MLP) and a bilinear interpolation on $\mathcal{D}$ to produce encoded depth feature F_d,

which has C_d channels and match the size of F^k. Afterwards, we solve its argmax value at the channel dimension, denoted as $G_{arg}(\cdot)$, to obtain a map of the depth positional index. We send positional index maps to a position embedding $G_{pe}(\cdot)$ with dictionary size C_d and embedding dimension C_{dim} to produce the 3D embedding maps F_{emb}. Hence we send F_{emd} to another MLP, flatten output embedding maps into vectors, and point-wisely add them to flattened F^k to generate semantic queries and keys with 3D-aware positional encoding. In function, we have:

$$F_{emb} = G_{pe}(G_{arg}(G_A(\mathcal{D}))), \qquad (3)$$

where G_A represents implementing an MLP and a bilinear interpolation. To ease the computational burden, we build up our MLP with convolution layers of kernel size 1.

Afterwards, queries and keys interact with semantic values in a self-attention manner to generate 3D-aware semantic features F_{3d}^k. Taking in our foreground masks, we have the distillation loss for intermediate features as :

$$\mathcal{L}_{feat} = \sum_{k=1}^{n} \alpha_I \|M^k(F_{3d}^k - T^k)\|^2 + \beta_I \|(1-M^k)(F_{3d}^k - T^k)\|^2, \qquad (4)$$

where α_I and β_I are hyperparameters for feature-level foreground and background balancing.

Target-Aware Attention Based Response Distillation

Our output-level distillation pipeline is shown in Figure 4. For detectors with an encoder-decoder transformer structure (Carion et al. 2020) for object detection, we define student and teacher decoders' intermediate object queries as $\{F_v^k\}_{k \in (1,...,m)}$ and $\{T_v^k\}_{k \in (1,...,m)}$, where m is the number of decoders' repeated transformer blocks. To achieve proper foreground and background balancing, we adopt the efficient Hungarian algorithm (Stewart, Andriluka, and Ng 2016) between teacher predictions and hard labels to generate a foreground query mask M_f for all object query levels. We leverage a cross-attention design as our feature adaptation module to guide student queries to learn meaningful 3D reasoning from the teacher. Specifically, we adopt MLPs to produce q_v^k from F_v^k, and produce k_v^k, v_v^k from T_v^k. Afterwards, we integrate these vectors in a cross-attention manner with q_v^k as the query, k_v^k as the key, and v_v^k as the value to generate adapted student object queries F_a^k. We have our distillation loss as:

$$\mathcal{L}_{ed} = \sum_{k=1}^{m} \alpha_v \|M_f(F_a^k - T_v^k)\|^2 + \beta_v \|(1-M_f)(F_a^k - T_v^k)\|^2, \qquad (5)$$

where α_v and β_v are hyperparameters for response-level foreground and background balancing.

End-to-End Network Training

We denote inherited losses from depth-assisted students as $\mathcal{L}_{cls}$, $\mathcal{L}_{reg}$ and $\mathcal{L}_{depth}$ for object classification, 3D bounding box regression and depth regression, respectively. $\mathcal{L}_{depth} =$
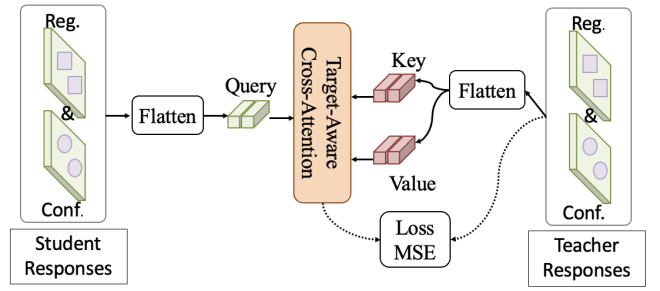

Figure 4: The visualization of our KD on responses with target-aware cross-attention for feature adaptation.

0 if baseline models utilize estimations from pretrained depth estimators. Our total loss function is:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{depth} + \alpha\mathcal{L}_{feat} + \beta\mathcal{L}_{ed}, \qquad (6)$$

where α and β are hyperparameters for loss balancing.

Experiments

Dataset. We evaluate our framework on the challenging KITTI (Geiger, Lenz, and Urtasun 2012) 3D object detection benchmark. Following 3DOP (Chen et al. 2015), we separate the dataset into 3,712 training samples, 3,769 validation samples, and 7,518 test samples. Ablation study was evaluated on the validation split with models trained on the training split. There are three object classes of interest in the KITTI dataset: car, pedestrian and cyclist. Each class is divided into three difficulty levels (easy, moderate and hard) based on occlusion, truncation and size.

Evaluation Metric. Following prior works (Guo et al. 2021b; Reading et al. 2021; Huang et al. 2022), we adopt two vital evaluation metrics, BEV Average Precision (AP_{BEV}) and 3D Average Precision (AP_{3D}), to analyze our framework. 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, respectively. Specifically, the car category has an IoU threshold of 0.7 while the pedestrian and cyclist categories have an IoU threshold of 0.5.

Baseline Models. To show the effectiveness of our KD framework, we implement it on three recent depth-assisted monocular detectors, MonoDETR (Chong et al. 2022), CaDDN (Reading et al. 2021) and PatchNet (Ma et al. 2020). Particularly, MonoDETR (Chong et al. 2022) leverages an encoder-decoder transformer structure (Carion et al. 2020) for object detection.

Implementation Details. Our framework is trained on NVIDIA V100 GPUs, with $\alpha_I = 1.0$, $\beta_I = 0.1$ for feature-level balancing, $\alpha_v = 1.0$, $\beta_v = 0.5$ for response-level balancing and $\alpha = 1.0$, $\beta = 1.0$ for final loss balancing. We set $\beta = 0$ if a baseline model does not adopt the transformer based encoder-decoder detecting head (Carion et al. 2020). Settings for optimizers, batch sizes and numbers of GPUs

Method	AP_{BEV}@IoU=0.7(Car test)			AP_{BEV}@IoU=0.7(Car val)			AP_{3D}@IoU=0.7(Car val)		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
MonoDETR (Zhang et al. 2022)	33.60	22.11	18.60	37.86	26.95	22.80	28.84	20.61	16.38
MonoDETR+**ADD**	35.20	23.58	20.08	40.38	29.07	25.05	30.71	21.94	18.42
Improvement	+1.60	+1.47	+1.48	+2.52	+2.12	+2.25	+1.87	+1.33	+2.04
CaDDN (Reading et al. 2021)	27.94	18.91	17.19	30.28	21.53	18.96	23.57	16.31	13.84
CaDDN+**ADD**	30.11	20.80	18.04	34.14	23.49	21.24	25.30	16.64	14.90
Improvement	+2.17	+1.89	+0.85	+3.86	+1.96	+2.28	+1.73	+0.33	+1.06
PatchNet (Ma et al. 2020)	22.97	16.86	14.97	41.49	23.60	19.93	31.60	16.80	13.80
PatchNet+**ADD**	28.15	17.38	15.06	42.15	24.75	20.26	32.21	16.92	13.87
Improvement	+5.18	+0.52	+0.09	+0.66	+1.15	+0.33	+0.61	+0.12	+0.07

Table 1: Quantitative comparisons of the Car category on the KITTI validation and testing splits. The results are evaluated using 40 recall positions. Our improvements relative to baseline models are listed with +.

Methods	Reference	Extra Data	AP_{BEV}(Car test)			AP_{3D}(Car test)		
			Easy	Mod.	Hard	Easy	Mod.	Hard
PatchNet (Ma et al. 2020)	ECCV 2020	LiDAR	22.97	16.86	14.97	15.68	11.12	10.17
D4LCN (Ding et al. 2020)	CVPR 2020	LiDAR	22.51	16.02	12.55	16.65	11.72	9.51
DDMP-3D (Wang et al. 2021)	CVPR 2021	LiDAR	28.08	17.89	13.44	19.71	12.78	9.80
CaDDN (Reading et al. 2021)	CVPR 2021	LiDAR	27.94	18.91	17.19	19.17	13.41	11.46
MonoDTR (Huang et al. 2022)	CVPR 2022	LiDAR	28.59	20.38	17.14	21.99	15.39	12.73
Kinematic3D (Brazil et al. 2020)	ECCV 2020	Temporal	26.69	17.52	13.10	19.07	12.72	9.17
MonoDLE (Ma et al. 2021)	CVPR 2021	None	24.79	18.89	16.00	17.23	12.26	10.29
MonoRUn (Chen et al. 2021a)	CVPR 2021	None	27.94	17.34	15.24	19.65	12.30	10.58
GrooMed-NMS (Kumar, Brazil, and Liu 2021)	CVPR 2021	None	26.19	18.27	14.05	18.10	12.32	9.65
MonoRCNN (Shi et al. 2021)	ICCV 2021	None	25.48	18.11	14.10	18.36	12.65	10.03
MonoEF (Zhou et al. 2021)	CVPR 2021	None	29.03	19.70	17.26	21.29	13.87	11.71
MonoFlex (Zhang, Lu, and Zhou 2021)	CVPR 2021	None	28.23	19.75	16.89	19.94	12.89	12.07
MonoJSG (Lian, Li, and Chen 2022)	CVPR 2022	None	32.59	21.26	18.18	24.69	16.14	13.64
MonoCon (Liu, Xue, and Wu 2022)	AAAI 2022	None	31.12	22.10	19.00	22.50	16.46	**13.95**
Ours	-	None	**35.20**	**23.58**	**20.08**	**25.61**	**16.81**	13.79
	-	*v.s. LiDAR*	+6.61	+3.2	+2.89	+3.62	+1.42	+1.06
Improvement	-	*v.s. Temporal*	+8.51	+6.06	+2.89	+6.54	+4.09	+4.62
	-	*v.s. None*	+2.61	+1.48	+1.08	+0.92	+0.35	-0.16

Table 2: Quantitative comparisons of the Car category on the KITTI test split. The best results are listed in **bold**.

Method	Depth	AP_{BEV}(Car val)			AP_{3D}(Car val)		
		Easy	Mod.	Hard	Easy	Mod.	Hard
MonoDETR	-	37.86	26.95	22.80	28.84	20.61	16.38
	✓	51.99	36.44	30.33	41.43	28.15	22.84
CaDDN	-	30.28	21.53	18.96	23.57	16.31	13.84
	✓	86.65	70.92	66.16	72.77	53.16	48.97
PatchNet	-	41.43	24.66	20.13	31.78	16.60	13.61
	✓	88.05	72.06	60.64	78.55	55.13	46.29

Table 3: Quantitative comparisons between our teachers and baselines on the KITTI validation split. 'Depth' refers to using the corresponding depth GT for training and inference.

used follow those for baseline models (Ma et al. 2020; Reading et al. 2021; Zhang et al. 2022). We design our framework teachers based on corresponding depth GT used in students. Specifically, we use object-wise depth for MonoDETR (Zhang et al. 2022) teacher, and pixel-wise depth for CaDDN (Reading et al. 2021) and PatchNet (Ma et al.

2020) teachers. We make the usage of depth GT following baselines, and report our teacher performance in Table 3.

Evaluation of Our Framework

Comparisons with Baselines. To evaluate the effectiveness of our framework, we implemented it on three depth-assisted baselines and re-trained each network model on KITTI 3D object detection dataset. The detection results of the Car category are reported in Table 1. For easy observation, the percentages of AP improvements are shown in red. Clearly, equipped with our attention-based depth distillation framework, the performance of all three baselines is further improved significantly on both validation and test samples, showing that our framework can successfully transfer precise depth knowledge from teacher model to student.

While performance gains on PatchNet (Ma et al. 2020) are not as obvious as those for MonoDETR (Zhang et al. 2022) and CaDDN (Reading et al. 2021) on the validation split. According to Simonelli et al. (Simonelli et al. 2021), the main reason for this is the contamination of objection detec-

	#	F	+SA	R	+CA	AP_{3D}@IoU=0.7(Car val)			AP_{BEV}@IoU=0.7(Car val)		
						Easy	Mod.	Hard	Easy	Mod.	Hard
Teacher(with object-wise depth GT)						41.43	28.15	22.84	51.99	36.44	30.33
MonoDETR*(Zhang et al. 2022)						27.63	19.93	16.70	36.64	26.32	22.35
Feature-level	(a)	✓	-	-	-	28.92	19.86	17.21	38.08	26.71	22.84
	(b)	✓	✓	-	-	29.50	20.90	17.51	38.62	27.03	23.06
Object-level	(c)	-	-	✓	-	29.00	19.97	16.58	37.93	26.66	22.66
	(d)	-	-	✓	✓	29.41	21.18	17.77	38.69	27.27	23.32
Feature-level & Object-level	(e)	✓	-	✓	-	27.40	19.73	16.37	36.94	26.68	21.89
	(f)	✓	✓	✓	-	28.51	20.63	17.28	38.02	26.81	22.96
	(g)	✓	-	✓	✓	29.20	20.69	17.25	37.95	26.63	22.73
	(h)	✓	✓	✓	✓	30.71	21.94	18.42	40.38	29.07	25.05
Improvement						+3.08	+2.01	+1.72	+3.74	+2.75	+2.70

Table 4: Ablation studies on the KITTI validation split. F denotes distillation on multi-level intermediate features, and $+SA$ denotes adopting our 3D-aware self-attention module for feature-level adaptation. R denotes distillation on network responses, and $+CA$ denotes adopting our target-aware cross-attention module for response-level adaptation. *The results for MonoDETR (Huang et al. 2022) are obtained by training the officially publicized code with our computational environment.

Settings	AP_{3D}@IoU=0.7(Car val)		
	Easy	Mod.	Hard
Baseline	27.63	19.93	16.70
$+SA$ w/o PE	28.45	20.37	16.30
$+SA$ w/ estimated PE	27.99	20.19	17.01
$+SA$ w/ GT PE	29.50	20.90	17.51
$\alpha_I = 1.00, \beta_I = 0.00$	28.37	20.32	17.12
$\alpha_I = 1.00, \beta_I = 0.05$	28.66	20.53	16.52
$\alpha_I = 1.00, \beta_I = 0.10$	29.50	20.90	17.51
$\alpha_I = 1.00, \beta_I = 0.20$	28.51	20.45	17.24
$\alpha_v = 1.00, \beta_v = 0.10$	28.55	20.04	16.65
$\alpha_v = 1.00, \beta_v = 0.25$	29.27	20.18	17.54
$\alpha_v = 1.00, \beta_v = 0.50$	29.41	21.18	17.77
$\alpha_v = 1.00, \beta_v = 1.00$	28.85	20.65	17.32

Table 5: Ablation studies for feature and response distillation modules on the KITTI validation split. 'estimated PE' denotes using the student's depth estimation sub-network to generate positional encoding. 'GT PE' denotes using object-wise depth GT to generate positional encoding.

tion validation split (Chen et al. 2015) by depth estimation training split. It leads to that partial depth maps (1226/3769) used for detection validation, obtained from DORN estimation (Fu et al. 2018), are as competitive as GT depth from direct LiDAR projection, and thus retrains the effectiveness of guiding monocular student with precise 3D measurements.

Comparisons with State-of-the-Art Methods. In Table 2, we present the experimental results of our framework and other state-of-the-art methods on the KITTI test split. Our framework achieves the best performance of the car class on three BEV-level metrics. Compared with the second-best approaches, our framework outperforms them with the ratio of 8.01%, 6.70% and 5.68% on the easy, moderate and hard levels, respectively. Furthermore, our **ADD** framework does not introduce any extra data or computational cost in the inference stage, hence is industrially implementable.

Ablation Study

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

Feature and Response Distillation. We evaluate the effectiveness of our 3D-aware self-attention on feature-level KD and target-aware cross-attention on output-level KD in Table 4. The first row Teacher(with object-wise depth GT) is the performance of our monocular teacher based on object-wise depth GT. The second row is our reproduce of MonoDETR(Zhang et al. 2022), but with slightly decayed performance compared with the official declaration due to environmental differences. Our framework enhances the student performance by large margins on all six metrics.

Settings for PE and Loss Weights. From Table 5 we observe that precise 3D measurements from GT PE can provide the most effective spatial guidance. It also illustrates that limiting the background loss weight to one tenth of the foreground achieves the best performance. Afterwards, considering response-level foreground-background balancing, Table 5 shows that the optimal choice is setting the background loss weight as half of the foreground.

Conclusion

In this work, we present **ADD**, an attention-based depth distillation framework for monocular 3D object detectors. We propose a *3D-aware self-attention module* for intermediate feature distillation and a *target-aware cross-attention module* for response distillation. Our framework does not necessarily rely on LiDAR or stereo assistance. In practice, our framework provides students with extra regularization to alleviate detection overfitting caused by depth estimation errors. At a high level, our framework injects students with mighty and dynamic depth-positional cues for 3D reasoning. We successfully implement our framework on three representative baselines using the KITTI 3D object detection benchmark, and achieve state-of-the-art performance without extra computational cost during inference.

References

Brazil, G.; and Liu, X. 2019. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9287–9296.

Brazil, G.; Pons-Moll, G.; Liu, X.; and Schiele, B. 2020. Kinematic 3d object detection in monocular video. In *European Conference on Computer Vision*, 135–152. Springer.

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

Chen, G.; Choi, W.; Yu, X.; Han, T.; and Chandraker, M. 2017. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30.

Chen, H.; Huang, Y.; Tian, W.; Gao, Z.; and Xiong, L. 2021a. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10379–10388.

Chen, X.; Kundu, K.; Zhu, Y.; Berneshawi, A. G.; Ma, H.; Fidler, S.; and Urtasun, R. 2015. 3d object proposals for accurate object class detection. *Advances in neural information processing systems*, 28.

Chen, Y.; Liu, S.; Shen, X.; and Jia, J. 2020. Dsgn: Deep stereo geometry network for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12536–12545.

Chen, Y.-N.; Dai, H.; and Ding, Y. 2022. Pseudo-Stereo for Monocular 3D Object Detection in Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 887–897.

Chen, Z.; Ye, X.; Yang, W.; Xu, Z.; Tan, X.; Zou, Z.; Ding, E.; Zhang, X.; and Huang, L. 2021b. Revealing the Reciprocal Relations Between Self-Supervised Stereo and Monocular Depth Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15529–15538.

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. *arXiv preprint arXiv:2201.10830*.

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

Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; and Tao, D. 2018. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2002–2011.

Fukuda, T.; Suzuki, M.; Kurata, G.; Thomas, S.; Cui, J.; and Ramabhadran, B. 2017. Efficient Knowledge Distillation from an Ensemble of Teachers. In *Interspeech*, 3697–3701.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

Guo, J.; Han, K.; Wang, Y.; Wu, H.; Chen, X.; Xu, C.; and Xu, C. 2021a. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2154–2164.

Guo, X.; Shi, S.; Wang, X.; and Li, H. 2021b. Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3153–3163.

He, C.; Zeng, H.; Huang, J.; Hua, X.-S.; and Zhang, L. 2020. Structure aware single-stage 3d object detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11873–11882.

Heo, B.; Kim, J.; Yun, S.; Park, H.; Kwak, N.; and Choi, J. Y. 2019. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1921–1930.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Hu, J.; Fan, C.; Jiang, H.; Guo, X.; Gao, Y.; Lu, X.; and Lam, T. L. 2021. Boosting light-weight depth estimation via knowledge distillation. *arXiv preprint arXiv:2105.06143*.

Huang, K.-C.; Wu, T.-H.; Su, H.-T.; and Hsu, W. H. 2022. MonoDTR: Monocular 3D Object Detection with Depth-Aware Transformer. In *CVPR*.

Jiao, J.; Wei, Y.; Jie, Z.; Shi, H.; Lau, R. W.; and Huang, T. S. 2019. Geometry-aware distillation for indoor semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2869–2878.

Kumar, A.; Brazil, G.; and Liu, X. 2021. Groomed-nms: Grouped mathematically differentiable nms for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8973–8983.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12697–12705.

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

Li, P.; Zhao, H.; Liu, P.; and Cao, F. 2020. Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving. In *European Conference on Computer Vision*, 644–660. Springer.

Li, Q.; Jin, S.; and Yan, J. 2017. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 6356–6364.

Lian, Q.; Li, P.; and Chen, X. 2022. MonoJSG: Joint Semantic and Geometric Cost Volume for Monocular 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1070–1079.

Liu, X.; Xue, N.; and Wu, T. 2022. Learning auxiliary monocular contexts helps monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1810–1818.

Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Luo, Z.; and Wang, J. 2019. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2604–2613.

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

Liu, Y.; Yan, J.; Jia, F.; Li, S.; Gao, Q.; Wang, T.; Zhang, X.; and Sun, J. 2022b. PETRv2: A Unified Framework for 3D Perception from Multi-Camera Images. *arXiv preprint arXiv:2206.01256*.

Liu, Z.; Zhou, D.; Lu, F.; Fang, J.; and Zhang, L. 2021. Autoshape: Real-time shape-aware monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15641–15650.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3111–3121.

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

Ma, X.; Wang, Z.; Li, H.; Zhang, P.; Ouyang, W.; and Fan, X. 2019. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6851–6860.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4721–4730.

Mousavian, A.; Anguelov, D.; Flynn, J.; and Kosecka, J. 2017. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 7074–7082.

Park, D.; Ambrus, R.; Guizilini, V.; Li, J.; and Gaidon, A. 2021. Is Pseudo-Lidar needed for Monocular 3D Object detection? In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

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

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; and Li, H. 2020. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10529–10538.

Shi, S.; Wang, X.; and Li, H. 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 770–779.

Shi, X.; Chen, Z.; and Kim, T.-K. 2020. Distance-normalized unified representation for monocular 3d object detection. In *European Conference on Computer Vision*, 91–107. Springer.

Shi, X.; Ye, Q.; Chen, X.; Chen, C.; Chen, Z.; and Kim, T.-K. 2021. Geometry-based distance decomposition for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15172–15181.

Simonelli, A.; Bulo, S. R.; Porzi, L.; Kontschieder, P.; and Ricci, E. 2021. Are we missing confidence in pseudo-lidar methods for monocular 3d object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3225–3233.

Stewart, R.; Andriluka, M.; and Ng, A. Y. 2016. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2325–2333.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10548–10557.

Tung, F.; and Mori, G. 2019. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1365–1374.

Wang, L.; Du, L.; Ye, X.; Fu, Y.; Guo, G.; Xue, X.; Feng, J.; and Zhang, L. 2021. Depth-conditioned dynamic message propagation for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 454–463.

Wang, T.; Yuan, L.; Zhang, X.; and Feng, J. 2019a. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4933–4942.

Wang, Y.; Chao, W.-L.; Garg, D.; Hariharan, B.; Campbell, M.; and Weinberger, K. Q. 2019b. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8445–8453.

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

Yim, J.; Joo, D.; Bae, J.; and Kim, J. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4133–4141.

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

Zhang, Y.; Lu, J.; and Zhou, J. 2021. Objects are Different: Flexible Monocular 3D Object Detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3288–3297.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7556–7566.