Weakly Supervised Monocular 3D Detection with a Single-View Image

Xueying Jiang[1] Sheng Jin[1] Lewei Lu[2] Xiaoqin Zhang[3] Shijian Lu[1*]
[1]S-Lab, Nanyang Technological University [2]Sensetime Research
[3]College of Computer Science and Technology, Zhejiang University of Technology

Abstract

Monocular 3D detection (M3D) aims for precise 3D object localization from a single-view image which usually involves labor-intensive annotation of 3D detection boxes. Weakly supervised M3D has recently been studied to obviate the 3D annotation process by leveraging many existing 2D annotations, but it often requires extra training data such as LiDAR point clouds or multi-view images which greatly degrades its applicability and usability in various applications. We propose SKD-WM3D, a weakly supervised monocular 3D detection framework that exploits depth information to achieve M3D with a single-view image exclusively without any 3D annotations or other training data. One key design in SKD-WM3D is a self-knowledge distillation framework, which transforms image features into 3D-like representations by fusing depth information and effectively mitigates the inherent depth ambiguity in monocular scenarios with little computational overhead in inference. In addition, we design an uncertainty-aware distillation loss and a gradient-targeted transfer modulation strategy which facilitate knowledge acquisition and knowledge transfer, respectively. Extensive experiments show that SKD-WM3D surpasses the state-of-the-art clearly and is even on par with many fully supervised methods.

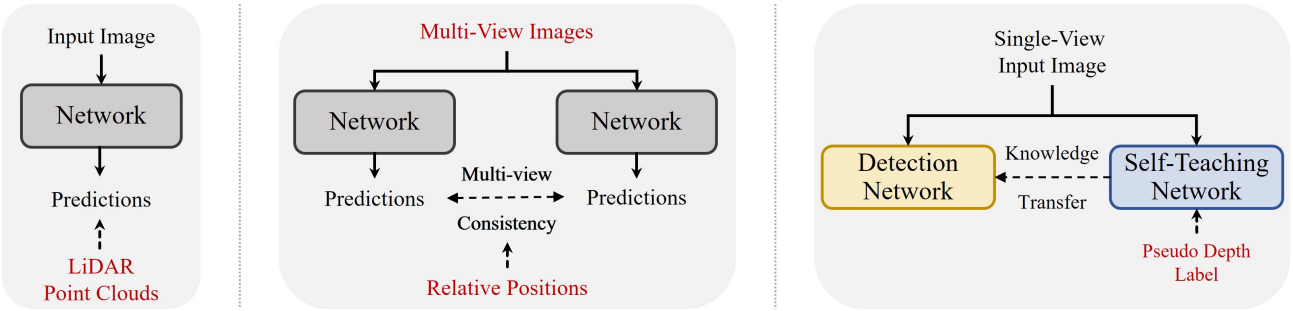
1. Introduction

Monocular 3D detection (M3D) has emerged as one key component in the area of autonomous driving and computer vision. Its primary target is to recognize objects and obtain their 3D localization from single-view images. Thanks to its low deployment cost, M3D [5, 34] has attracted increasing attention in both academic and industrial sectors, achieving very impressive progress in recent years. On the other hand, most existing studies [21, 33, 37, 39] adopt a fully supervised setup which have been facing increasing scalability concern as large-scale 3D boxes are often labor-intensive to collect. Effective M3D training without 3D annotations has become a critical issue while handling M3D problems in various research and practical tasks.

Weakly supervised M3D (WM3D) [34] has recently been explored for learning effective 3D detectors without 3D box annotations, aiming to exploit 2D annotations to make up for the absence of 3D information. For example, WeakM3D [34] exploits LiDAR point clouds to infer 3D information as illustrated in Figure 1(a). However, it requires costly and complicated LiDAR sensors to collect point clouds which limits its applicability and usability greatly. WeakMono3D [42] employs 2D information only by either leveraging multi-view stereo with images from multiple cameras or constructing pseudo-multi-view perspective from sequential video frames as illustrated in Figure 1(b). However, collecting multi-view images is complicated, and resorting to a pseudo multi-view perspective degrades the detection performance clearly. With the advance of single-view depth estimation, WM3D with depth from a single-view image presents a potential solution for compensating the absence of 3D annotations. On the other hand, direct integration of such depth into existing frameworks often necessitates complex network architectures which further incurs significant computational costs. This gives rise to a pertinent question: When not using additional LiDAR point clouds or multi-view image pairs, is it possible to harness the depth from off-the-shelf depth estimators without introducing much computational overhead in inference?

We design SKD-WM3D, a novel weakly supervised monocular 3D object detection method that is exclusively grounded on single-view images. One key design in SKD-WM3D is a self-knowledge distillation framework which consists of a **D**epth-guided **S**elf-teaching **N**etwork (DSN) and a **M**onocular 3D **D**etection **N**etwork (MDN). As illustrated in Figure 1(c), SKD-WM3D utilizes depth information obtained from an off-the-shelf depth estimator [17] to enhance the 3D localization ability of DSN and transfers such ability to MDN via self-knowledge distillation. Such self-distillation design enables MDN to unearth the intrinsic depth information from single-view images independently,

*Corresponding author.



(a) LiDAR-based WM3D (b) Multi-view WM3D (c) Single-view WM3D (Ours)

Figure 1. Different paradigms in weakly supervised monocular 3D detection. Our approach in (c) leverages *Pseudo Depth Labels* from a single-view image to achieve weakly supervised monocular 3D detection, requiring no extra training data like LiDAR point clouds or multi-view images as in (a) and (b). It improves usability and applicability greatly. The *Pseudo Depth Labels* are obtained with an off-the-shelf depth estimator [17] without extra training and ground-truth depth labels. Data in red denotes extra data in network training.

bypassing additional modules such as pre-trained depth estimation networks and leading to precise and efficient 3D localization with little computational overhead during inference. On top of DSN and MDN, we design an uncertainty-aware distillation loss to optimize the utilization of the transferred 3D localization knowledge by weighting up more certain knowledge while weighting down less certain knowledge. In addition, we design a gradient-targeted transfer modulation strategy to synchronize the learning paces of DSN and MDN during the process of learning 3D localization knowledge, by prioritizing MDN learning at the initial stage when MDN lags behind DSN and enabling it to provide more feedback to DSN when MDN is better trained at late stages.

Our contribution can be summarized in three aspects. *First*, we design a novel framework that achieves weakly supervised monocular 3D detection by distilling knowledge between a depth-guided self-teaching network and a monocular 3D detection network. Without any extra training data like LiDAR point clouds or multi-view images, the framework exploits depth exclusively from a single image with little computational overhead in inference. *Second*, we design an uncertainty-aware distillation loss and a gradient-targeted transfer modulation strategy which facilitate knowledge acquisition and knowledge transfer, respectively. *Third*, the proposed approach clearly outperforms the state-of-the-art in weakly supervised monocular 3D detection, and its performance is even on par with several fully supervised methods.

2. Related Work

2.1. Monocular 3D Detection

Monocular 3D object detection aims to predict 3D object localization from single-view images. Standard monocular detectors [1, 6, 14, 53, 56] operate solely on single images, without utilizing additional data. However, the inherent depth ambiguity of monocular detection significantly hinders its performance compared to its stereo counterparts. To address this limitation, various approaches seek solutions with the help of extra data, such as LiDAR point clouds [4, 7, 21, 27], video sequences [2], 3D CAD models [5, 24, 31], and depth estimation [10, 35, 45, 48]. Specifically, MonoRUn [4] adopts an uncertainty-aware regional reconstruction network for regressing pixel-associated 3D object coordinates with LiDAR point clouds as extra supervision. MonoDistill [7] introduces an effective distillation-based approach that incorporates spatial information from LiDAR signals into monocular 3D detection. Additionally, pseudo-LiDAR-based methods [45, 48] convert estimated depth maps to simulate the real LiDAR point clouds to utilize the well-designed LiDAR-based 3D detector. During inference, compared with methods using depth estimation, our method eliminates the need for pseudo depth labels and complex network architectures, with little computational overhead. Besides, existing fully supervised methods require large-scale 3D box ground truth, which is labor-intensive to collect and annotate.

2.2. Weakly Supervised 3D Object Detection

Due to the high cost of annotating 3D boxes in the 3D object detection task, various weakly supervised approaches [26, 29, 34, 36] have been proposed. For example, WS3D [29] presents a weakly supervised method for 3D LiDAR object detection, which requires only a limited number of weakly annotated scenes with center-annotated BEV maps. VS3D [36] introduces a cross-model knowledge distillation strategy to transfer the knowledge from the RGB domain to the point cloud domain, using LiDAR point clouds as weak supervision. Recent research on weakly supervised 3D object detection has turned to explore the monocular setting. For example, WeakM3D [34] generates 2D boxes to select

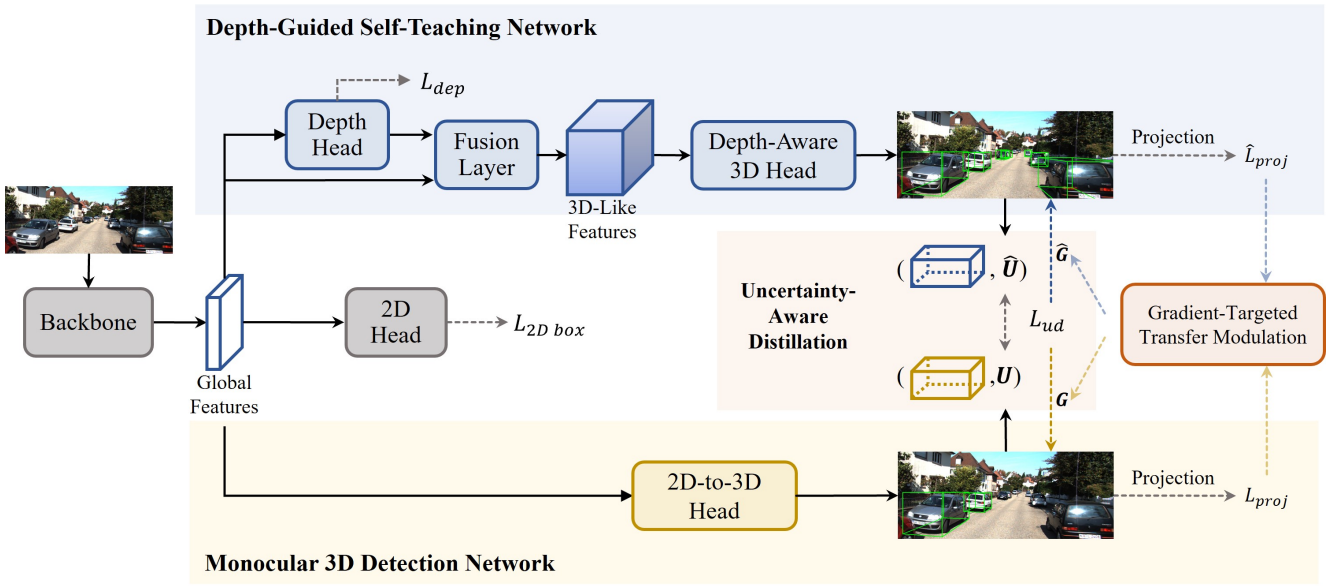


Figure 2. The framework of the proposed self-knowledge distillation network. The framework consists of a depth-guided self-teaching network and a monocular 3D detection network. The depth-guided self-teaching network acquires comprehensive 3D localization knowledge by leveraging depth information and transfers its learned expertise to the monocular 3D detection network via soft label distillation to enhance its performance. We design an uncertainty-aware distillation loss and a gradient-targeted transfer modulation strategy to facilitate the knowledge transfer between the two networks effectively. During inference, the monocular 3D detection network extracts intrinsic depth information from single-view images independently with little computational overhead.

RoI LiDAR point clouds as weak supervision and then predicts 3D boxes that closely align with the selected RoI LiDAR point clouds. More recently, WeakMono3D [42] eliminates the need for LiDAR, offering both multi-view and single-view yet multi-frame versions. While the former acquires stereo image inputs from multiple cameras, the latter constructs a pseudo-multi-view perspective using multiple video frames. The multi-frame version exhibits inferior 3D scene comprehension compared with the multi-view approach due to its smaller inter-frame disparity, leading to degraded performance. Instead of requiring extra training data like LiDAR point clouds or multi-view images, we tackle the challenge of weakly supervised monocular 3D detection by leveraging a single-view image exclusively.

2.3. Self-Knowledge Distillation

Knowledge distillation [8, 9, 11, 16, 18, 23, 32, 38, 43, 54] aims to transfer knowledge from a pre-trained teacher network to a student network for improving its performance. Self-knowledge distillation [30, 41, 47], distinct from traditional knowledge distillation, leverages the information within the student network to facilitate its learning without the pre-trained teacher network. Specifically, data augmentation approach [15, 46, 50] transfers knowledge through different distortions of the same training data. However, they are susceptible to inappropriate augmentations, such as improper instance rotation or distortion, potentially introducing noise that hampers network learning. Another typical approach exploits auxiliary networks [52, 57]. For example, DKS [40] introduces auxiliary supervision branches and pairwise knowledge alignments, while FRSKD [19] adds a new branch supervised by the original features and utilizes both soft-label and feature-map distillation. Our work is the first that introduces self-knowledge distillation with auxiliary networks for weakly supervised monocular 3D detection. It effectively exploits depth information from single-view images with little computational overhead during inference.

3. Methodology

This section presents the proposed SKD-WM3D. First, the problem definition and overview are presented in Sec. 3.1. Then detailed designs of SKD-WM3D are introduced, including the self-knowledge distillation framework in Sec. 3.2, the uncertainty-aware distillation loss in Sec. 3.3 and the gradient-targeted transfer modulation strategy in Sec. 3.4. Finally, loss functions are presented in Sec. 3.5.

3.1. Problem Definition and Overview

Weakly supervised monocular 3D detection takes an RGB image and the corresponding 2D bounding boxes as supervision, aiming to classify objects and determine their

bounding boxes in 3D space without involving any 3D annotations in training. The prediction of each object is composed of the object category C, a 2D bounding box B^{2D}, and a 3D bounding box B^{3D}. Specifically, the 3D box B^{3D} can be further decomposed to the object 3D location (x_{3D}, y_{3D}, z_{3D}), the object dimension with height, width and length (h_{3D}, w_{3D}, l_{3D}), as well as orientation θ.

We design a self-knowledge distillation framework to tackle the challenge of weakly supervised monocular 3D detection from a single-view image. As Figure 2 shows, the framework consists of two subnetworks including a *Depth-Guided Self-Teaching Network* and a *Monocular 3D Detection Network*. In the *Depth-Guided Self-Teaching Network*, the global features F_G extracted by the backbone are fed into a *Depth Head* to obtain depth features. Next, the global features F_G and the extracted depth features are fed into a *Fusion Layer* to obtain 3D-like features F_{3D}. The 3D-like features of each object are then obtained via RoIAlign, and further fed to a *Depth-Aware 3D Head* to predict 3D box $\widehat{B}_p^{3D}$ and uncertainty $\widehat{U}$. In the *Monocular 3D Detection Network*, We first use RoIAlign to generate object-level features from the global features F_G, and then feed them to a *2D-to-3D Head* to predict 3D box $\widehat{B}_p^{2D}$ and uncertainty U. The 3D boxes predicted by both networks are further projected into 2D boxes. Moreover, we design an uncertainty-aware distillation loss L_{ud} to obtain low-uncertainty knowledge, and a gradient-targeted transfer modulation strategy to synchronize the learning paces between the two networks by controlling gradients $\widehat{G}$ and G of L_{ud}.

3.2. Self-Knowledge Distillation Framework

The self-knowledge distillation framework enhances the 3D localization ability of the depth-guided self-teaching network by utilizing depth information from an off-the-shelf depth estimator and then transfers the ability to the monocular 3D detection network via self-knowledge distillation.

Depth-Guided Self-Teaching Network. To equip the self-teaching network with 3D localization ability, we propose to learn from global features F_G and depth information from an off-the-shelf depth estimator to acquire comprehensive 3D knowledge. The depth information is exploited via two major designs. Firstly, we introduce a depth head $\mathcal{D}$ that extracts depth features F_D as follows:

$$F_D = \mathcal{D}(F_G), \tag{1}$$

The depth features F_D are exploited to generate depth maps D_p, where the depth map generation is supervised by the pseudo ground truth of the depth map D_{gt} that is predicted by an off-the-shelf depth estimator by using the focal loss [22] as depth loss L_{dep}. Hence, the depth features can be acquired by the depth-guided self-teaching network effectively.

Remark 1. We generate depth pseudo labels using an off-the-shelf depth estimator [17] with frozen weights, eliminating the need for additional training and ground-truth depth labels. Adopting an off-the-shelf depth estimator incurs negligible costs as compared with prior studies that require either point clouds [34] or multi-view images [42].

Secondly, we obtain 3D-like features F_{G3D} by integrating the depth features F_D that provide information along the depth dimension, as well as the global features F_G that capture knowledge about the 2D image plane. Specifically, we design a fusion layer that fuses the depth features F_D with the global features F_G to derive the F_{G3D} as follows:

$$F_{G3D} = FFN(CA(SA(F_D), F_G)), \tag{2}$$

where the FFN is the feed-forward network, and CA, SA denote $CrossAttention$, $SelfAttention$, respectively. The structures of $CrossAttention$ and $SelfAttention$ employ the standard transformer architecture [44]. The obtained 3D comprehension improves the network's ability to precisely locate objects, effectively mitigating depth ambiguity arising from single-view image input.

Monocular 3D Detection Network. The monocular 3D detection network acquires the 3D localization knowledge from the depth-guided self-teaching network. By distilling soft labels generated by the depth-guided self-teaching network, the monocular 3D detection network can extract intrinsic depth information from images independently during inference. This kills the need for additional complex modules such as pre-trained depth estimation networks or depth fusion modules, facilitating the inference with little computational overhead.

3.3. Uncertainty-Aware Distillation Loss

During the knowledge distillation process, uncertain knowledge could affect the network training negatively if all transferred knowledge is treated equally. To benefit more from certain knowledge and weaken the effect of uncertain knowledge, we design an uncertainty-aware distillation loss between the 3D boxes that are predicted by the two networks in the self-knowledge distillation framework. The uncertainty-aware distillation loss exploits the prediction uncertainty to modulate the distillation loss magnitude as follows:

$$L_{ud} = \frac{L_d}{min((\widehat{U}+U)/2, \alpha)} + \left\| min(\frac{\widehat{U}+U}{2}, \alpha) \right\|^2, \tag{3}$$

where $\widehat{U}$ and U denote the uncertainty of the 3D boxes that are predicted by the two networks, respectively. Here

we assume the 3D box predictions have the Laplace distribution, and we adopt the standard deviations as the uncertainties, inspired by [20, 25, 33]. $\left\|min(\frac{\widehat{U}+U}{2}, \alpha)\right\|^2$ is the L2 regularization, and α is a fixed value set to 0.1. L_d denotes the basic distillations loss, and we employ the commonly used SmoothL1 [13] loss to enforce the consistency between the 3D boxes predicted by the two networks. The SmoothL1 loss leaves a soft margin when computing the difference between the two 3D boxes:

$$L_d = \begin{cases} 0.5 \times (\widehat{B}_p^{3D} - B_p^{3D})^2, & \text{if } |\widehat{B}_p^{3D} - B_p^{3D}| < 1.0 \\ |\widehat{B}_p^{3D} - B_p^{3D}| - 0.5, & \text{otherwise} \end{cases}, \tag{4}$$

where $\widehat{B}_p^{3D}$ and B_p^{3D} are the predicted 3D boxes from the depth-guided self-teaching network and the monocular 3D detection network, respectively.

Remark 2. The proposed uncertainty-aware distillation loss L_{ud} integrates average uncertainty $\frac{\widehat{U}+U}{2}$ as regularization and a weighted component for the basic distillation loss L_d, allowing adaptive learning adjustments based on the knowledge's uncertainty level. Specifically, when dealing with uncertain knowledge, a smaller weight is assigned to the basic distillation loss L_d to mitigate potential adverse effects on network learning. Consequently, the network prioritizes optimizing uncertainty reduction in such scenarios. When dealing with certain knowledge, the network emphasizes optimizing the basic distillation loss L_d due to its higher weight. Notably, the basic distillation loss L_d simply considers box consistency, while integrating uncertainty is beneficial for enhancing the knowledge distillation process.

3.4. Transfer Modulation Strategy

The depth-guided self-teaching network, which leverages depth information to predict 3D boxes, transfers its learned 3D knowledge to the monocular 3D detection network. The asynchronous learning paces of the two networks pose potential challenges to effective 3D knowledge transfer.

We design a gradient-targeted transfer modulation strategy to synchronize the learning pace of the depth-guided self-teaching network and the monocular 3D detection network. We modulate the knowledge transfer dynamically, by controlling the gradients from the uncertainty-aware distillation loss L_{ud}. Specifically, we adapt the gradients based on the 2D projection performance of each network, assigning smaller backward gradients for the good-performing network and higher backward gradients for the bad-performing network. The gradient-targeted transfer modulation strategy is formulated as follows:

$$\widehat{G}' = \frac{2 \times \widehat{L}_{proj}}{\widehat{L}_{proj} + L_{proj}} \times \widehat{G}, G' = \frac{2 \times L_{proj}}{\widehat{L}_{proj} + L_{proj}} \times G, \tag{5}$$

Where $\widehat{G}$ and G are the original gradients of the two networks, $\widehat{G}'$ and G' are the modified gradients, $\widehat{L}_{proj}$ and L_{proj} are projection losses, computed between the projected 2D boxes from 3D box predictions and 2D box annotations.

The gradient-targeted transfer modulation strategy prioritizes training the monocular 3D detection network when its learning lags behind the depth-guided self-teaching network at the early training stage. As the monocular 3D detection network learns and improves gradually, it is enabled to provide more feedback progressively to the depth-guided self-teaching network.

3.5. Loss Functions

The overall objective consists of three losses including L_{ud}, L_{dep} and L_{base}. L_{ud} is the uncertainty-aware distillation loss as defined in Sec. 3.3. L_{dep} is the depth loss for supervising the predicted depth map. L_{base} includes losses for supervising 2D boxes prediction by 2D heads and the 3D box predictions, which has been adopted in prior CenterNet [55] and WeakMono3D [42]. We set the weight for each loss item to 1.0, and the overall loss function can be formulated as follows:

$$L = L_{ud} + L_{dep} + L_{base}. \tag{6}$$

4. Experiments

4.1. Datasets

We conduct experiments over the KITTI 3D dataset [12] and the nuScenes dataset [3] that have been widely adopted for benchmarking of 3D object detection methods. The KITTI 3D dataset consists of 7,481 images for training and 7,518 images for testing. The labels of the train set are publicly available and the labels of the test set are stored on a test server for evaluation. For ablation studies, we follow [5] which divides the 7,481 training samples into a new train set with 3,712 images and a validation set with 3,769 images. The nuScenes dataset comprises 1,000 video scenes, including RGB images captured by 6 surround-view cameras. The dataset is split into a training set (700 scenes), a validation set (150 scenes), and a test set (150 scenes). Following [34], the performance on the validation set is reported.

4.2. Evaluation Protocols

For the KITTI 3D dataset, following [39], we adopt the evaluation metric $\text{AP}|_{R_{40}}$ which is the average of the AP of 40 recall points. We report the average precision on bird's eye view and 3D object detection as $\text{AP}_{BEV}|_{R_{40}}$ and $\text{AP}_{3D}|_{R_{40}}$. In addition, as most weakly supervised 3D object detection methods apply IoU threshold of 0.7 for the test set and 0.5 for the validation set, we adopt the same thresholds for fair benchmarking. We adopt four metrics

Method	Backbone	Supervision	$AP_{BEV}/AP_{3D}(IoU=0.7)\|_{R_{40}}$		
			Easy	Moderate	Hard
WeakM3D [34]	ResNet-50	Weak	11.82/5.03	5.66/2.26	4.08/1.63
WeakMono3D [42]	DLA-34		12.31/6.98	8.80/4.85	7.81/4.45
SKD-WM3D (Ours)	DLA-34		**15.71/8.95**	**10.15/5.54**	**8.08/4.53**

Table 1. Comparison on the performance of the Car category on KITTI *test* set. For all results, we use $AP|_{R_{40}}$ metrics with IoU threshold equal to 0.7. The best results are in **bold**.

Method	Backbone	Supervision	$AP_{BEV}/AP_{3D}(IoU=0.5)\|_{R_{40}}$		
			Easy	Moderate	Hard
CenterNet [55]	DLA-34	Full	34.36/20.00	27.91/17.50	24.65/15.57
MonoGRNet [35]	VGG-16		52.13/47.59	35.99/32.28	28.72/25.50
M3D-RPN [1]	DenseNet-121		53.35/48.53	39.60/35.94	31.76/28.59
MonoPair [6]	DLA-34		61.06/55.38	47.63/42.39	41.92/37.99
MonoDLE [28]	DLA-34		60.73/55.41	46.87/43.42	41.89/37.81
GUPNet [25]	DLA-34		61.78/57.62	47.06/42.33	40.88/37.59
Kinematic [2]	DenseNet-121		61.79/55.44	44.68/39.47	34.56/31.26
MonoDistill [7]	DLA-34		71.45/65.69	53.11/49.35	46.94/43.49
MonoDETR [53]*	ResNet-50		72.34/68.05	51.97/48.42	46.94/43.48
VS3D [36]	VGG-16	Weak	31.59/22.62	20.59/14.43	16.28/10.91
Autolabels [51]	ResNeXt101		50.51/38.31	30.97/19.90	23.72/14.83
WeakM3D [34]	ResNet-50		**58.20**/50.16	38.02/29.94	30.17/23.11
WeakMono3D [42]	DLA-34		54.32/49.37	42.83/39.01	40.07/36.34
SKD-WM3D (Ours)	DLA-34		55.47/**50.21**	**44.35/41.57**	**41.86/36.92**

Table 2. Comparison on the performance of the Car category on KITTI *val* set. For all results, we use $AP|_{R_{40}}$ metric with IoU threshold equal to 0.5. * denotes this performance is reproduced from the official code. The best results of weakly supervised 3D object detection approaches are in **bold**.

Method	AP↑	ATE↓	ASE↓	AAE↓
WeakM3D [34]	0.214	0.814	0.234	0.682
SKD-WM3D (Ours)	**0.242**	**0.795**	**0.231**	**0.659**

Table 3. Comparison on the performance of the Car category on nuScenes *val* set. The best results are in **bold**.

for the evaluation on the nuScenes dataset, namely, AP (Average Precision), ATE (Average Translation Error), ASE (Average Scale Error), and AAE (Average Attribute Error). Following [34], AVE (Average Velocity Error) and AOE (Average Orientation Error) are not reported due to the lack of supervision for velocity and movement direction in the weakly supervised approach.

4.3. Implementation Details

We conduct experiments on 2 NVIDIA V100 GPUs with batch size of 16, and train the framework with 150 epochs. We use the Adam optimizer with the initial learning rate $1e^{-5}$, which is gradually increased to $1e^{-3}$ for the first 5 epochs and decayed with rate 0.1 at the 90 and 120 epochs. We employ DLA-34 [49] as the detector's backbone. The pseudo ground truth of the depth map is generated with an off-the-shelf depth estimator [17] without using the ground truth of depth label.

4.4. Comparison with State-of-the-Art Methods

We compare our method with several state-of-the-art weakly supervised monocular 3D detection methods on the KITTI test set. As Table 1 shows, our method achieves superior detection performance across all metrics. This superior performance is largely attributed to our designed self-knowledge distillation framework that extracts and exploits intrinsic depth information from a single-view image effectively. It should be highlighted that our method employs a single-view image exclusively without involving additional training data such as LiDAR point clouds [34] or multi-view image pairs [42].

Table 2 shows the benchmarking on the KITTI validation set. Specifically, we compare our method against both state-of-the-art weakly supervised monocular 3D de-

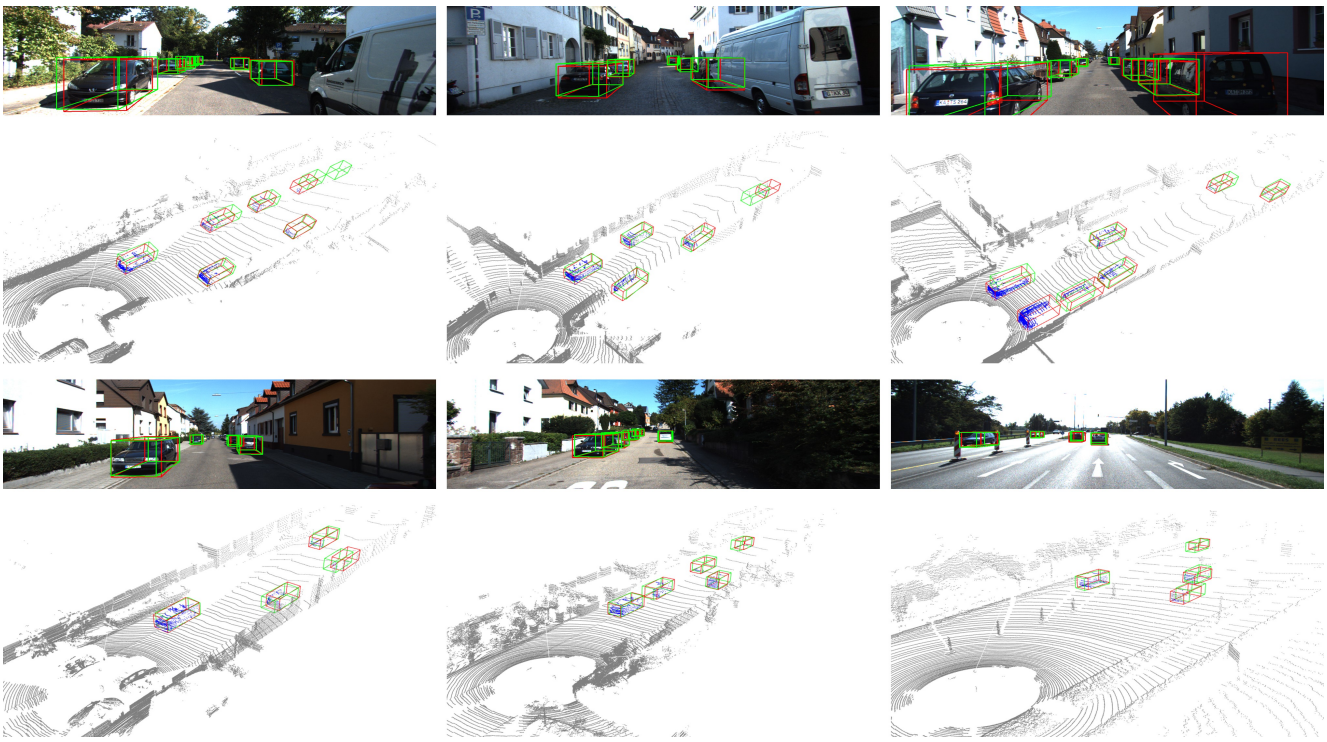


Figure 3. Qualitative illustration on KITTI *val* set. Red boxes denote ground-truth annotations and Green boxes denote our predictions. The ground truth of LiDAR point clouds is utilized for visualization purposes only. Best viewed with zoom-in.

Figure 4. Qualitative illustration of object detection and the corresponding detection uncertainties on KITTI *val* set. Red boxes denote ground-truth annotations and Green boxes denote our predictions. The detection accuracy is closely correlated with the detection uncertainty. Best viewed with zoom-in.

tection methods and fully supervised methods. It can be seen that our method achieves superior performance compared to WeakM3D [34] and WeakMono3D [42] across most metrics, even without using LiDAR point clouds or multi-view images. Notably, our method significantly outperforms WeakM3D [34] in the Moderate and Hard categories, primarily due to the sparse nature of distant LiDAR point clouds adversely affecting its performance. Additionally, its performance is even on par with several fully supervised methods [1, 35, 55].

Table 3 shows the results on the nuScenes validation set. It can be seen that our proposed method outperforms WeakM3D [34] across all four evaluation metrics, validating the effectiveness of our approach.

Qualitative Results Figure 3 shows qualitative results with both 2D RGB images and 3D point clouds. In simple scenarios, our model achieves great prediction accuracy, which is largely attributed to the proposed self-knowledge distillation framework as well as the uncertainty-aware distillation loss and the gradient-targeted transfer modulation strategy, all working together to facilitate comprehensive 3D information extraction effectively. However, for heavily occluded or distant objects, the accuracy of orientation and depth estimation tends to drop, which is common for monocular 3D detection due to its ill-posed nature. In addition, Figure 4 shows the visualization of predicted bounding boxes and their corresponding uncertainties. It can be observed that the prediction accuracy of bounding boxes has a close correlation with the prediction uncertainty.

Index	MDN	DSN	$AP_{BEV}/AP_{3D}(IoU=0.5)|_{R_{40}}$		
			Easy	Moderate	Hard
1	✓		0.00/0.00	0.00/0.00	0.00/0.00
2		✓	45.23/40.96	34.27/31.02	30.17/26.27
3	✓	✓	**55.47/50.21**	**44.35/41.57**	**41.86/36.92**

Table 4. Ablation study of the proposed self-knowledge distillation framework. The best results are in **bold**. MDN denotes the Monocular 3D Detection Network, while DSN denotes the Depth-Guided Self-Teaching Network.

Index	L_{ud}	TMS	$AP_{BEV}/AP_{3D}(IoU=0.5)|_{R_{40}}$		
			Easy	Moderate	Hard
1			49.95/44.61	38.24/35.74	37.28/34.82
2	✓		53.16/48.13	41.85/39.02	40.14/35.70
3		✓	52.35/46.30	41.45/38.91	39.73/35.44
4	✓	✓	**55.47/50.21**	**44.35/41.57**	**41.86/36.92**

Table 5. Ablation study of the proposed uncertainty-aware distillation loss and the gradient-targeted transfer modulation strategy. The best results are in **bold**. L_{ud} denotes the Uncertainty-Aware Distillation Loss, while TMS denotes the gradient-targeted transfer modulation strategy.

Method	WeakM3D [34]	MonoDistill [7]*	SKD-WM3D (Ours)*
FPS	13.9	25.0	30.3

Table 6. Comparison on inference speed of M3D methods. * denotes this method utilizes dense depth maps.

4.5. Ablation Study

We conduct extensive ablation studies on the KITTI validation dataset to evaluate our designs. Specifically, we evaluated the efficacy of the two individual networks in the proposed self-knowledge distillation framework. Additionally, we examine the effect of the proposed uncertainty-aware distillation loss and the gradient-targeted transfer modulation strategy. Lastly, we evaluated the efficiency of our monocular 3D detection framework.

Self-Knowledge Distillation Framework. We train two models to assess the contributions of the two networks in our proposed self-knowledge distillation framework. As Table 4 shows, training the monocular 3D detection network alone produces few meaningful detection results as the absence of depth information leads to ambiguous object localization along the depth dimension. As a comparison, training the depth-guided self-teaching network alone can produce reasonable detection results thanks to the depth map pseudo labels. In addition, training both subnetworks concurrently produces the best 3D detection, validating the effectiveness of extracting 3D information from a single image. We can also see that including the self-knowledge distillation on top of the depth-guided self-teaching network greatly improves the detection by reducing the adverse effects of uncertain knowledge and enabling communication between the two subnetworks during training.

Uncertainty-Aware Distillation Loss and Gradient-Targeted Transfer Modulation Strategy. Table 5 shows the ablation study of the proposed uncertainty-aware distillation loss and the gradient-targeted transfer modulation strategy. It can be observed that the baseline does not perform well due to the adverse effect of uncertain knowledge and the asynchronous learning paces of the two subnetworks. On top of the baseline, including either the uncertainty-aware distillation loss or the gradient-targeted transfer modulation strategy improves the detection performance significantly, underscoring the importance of attaining high-certainty knowledge and synchronizing the learning paces of the two networks. In addition, combining the two designs achieves the best performance, highlighting their complementary nature and collaborative roles in knowledge acquisition and knowledge transfer.

Inference speed comparison. Table 6 compares the inference speed on the KITTI validation set. Our method demonstrates superior efficiency thanks to our designed self-knowledge distillation framework, without utilizing complex network architectures during inference.

5. Conclusion

In this paper, we point out that previous weakly supervised monocular 3D detection methods either require additional LiDAR point clouds or paired images from multiple viewpoints or temporal sequences. To overcome these constraints, we propose a weakly supervised monocular 3D object detection approach that is exclusively grounded on single-view image inputs. Central to our approach is a self-knowledge distillation framework, which effectively harnesses the depth information within a single-view image with little computational overhead during inference. We further introduce an uncertainty-aware distillation loss and a gradient-targeted transfer modulation strategy, facilitating knowledge acquisition and knowledge transfer, respectively. Finally, extensive experiments demonstrate the effectiveness of our method. Moving forward, we plan to further generalize our work to diverse and challenging scenarios, such as occlusions, varying lighting, and weather conditions, thereby enhancing its practical applicability.

Acknowledgement

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9287–9296, 2019. 2, 6, 7

[2] Garrick Brazil, Gerard Pons-Moll, Xiaoming Liu, and Bernt Schiele. Kinematic 3d object detection in monocular video. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 135–152. Springer, 2020. 2, 6

[3] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 5

[4] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10379–10388, 2021. 2

[5] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2147–2156, 2016. 1, 2, 5

[6] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. Monopair: Monocular 3d object detection using pairwise spatial relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12093–12102, 2020. 2, 6

[7] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, and Wanli Ouyang. Monodistill: Learning spatial features for monocular 3d object detection. *International Conference on Learning Representations*, 2022. 2, 6, 8

[8] Inseop Chung, SeongUk Park, Jangho Kim, and Nojun Kwak. Feature-map-level online adversarial knowledge distillation. In *International Conference on Machine Learning*, pages 2006–2015. PMLR, 2020. 3

[9] Kaiwen Cui, Yingchen Yu, Fangneng Zhan, Shengcai Liao, Shijian Lu, and Eric P Xing. Kd-dlgan: Data limited image generation via knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3872–3882, 2023. 3

[10] Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning depth-guided convolutions for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1000–1001, 2020. 2

[11] Huan-ang Gao, Beiwen Tian, Pengfei Li, Hao Zhao, and Guyue Zhou. Dqs3d: Densely-matched quantization-aware semi-supervised 3d detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 5

[13] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1440–1448, 2015. 5

[14] Tong He and Stefano Soatto. Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8409–8416, 2019. 2

[15] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1921–1930, 2019. 3

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[17] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Penet: Towards precise and efficient image guided depth completion. In *International Conference on Robotics and Automation*, pages 13656–13662. IEEE, 2021. 1, 2, 4, 6

[18] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35:33716–33727, 2022. 3

[19] Mingi Ji, Seungjae Shin, Seunghyun Hwang, Gibeom Park, and Il-Chul Moon. Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10664–10673, 2021. 3

[20] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017. 5

[21] Jason Ku, Alex D Pon, and Steven L Waslander. Monocular 3d object detection leveraging accurate proposals and shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11867–11876, 2019. 1, 2

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2980–2988, 2017. 4

[23] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7096–7104, 2019. 3

[24] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. Autoshape: Real-time shape-aware monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15641–15650, 2021. 2

[25] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3111–3121, 2021. 5, 6

[26] Zhipeng Luo, Zhongang Cai, Changqing Zhou, Gongjie Zhang, Haiyu Zhao, Shuai Yi, Shijian Lu, Hongsheng Li, Shanghang Zhang, and Ziwei Liu. Unsupervised domain adaptive 3d detection with multi-level consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8866–8875, 2021. 2

[27] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, and Xin Fan. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6851–6860, 2019. 2

[28] Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, and Wanli Ouyang. Delving into localization errors for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4721–4730, 2021. 6

[29] Qinghao Meng, Wenguan Wang, Tianfei Zhou, Jianbing Shen, Luc Van Gool, and Dengxin Dai. Weakly supervised 3d object detection from lidar point cloud. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 515–531. Springer, 2020. 2

[30] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in Neural Information Processing Systems*, 32, 2019. 3

[31] J Krishna Murthy, GV Sai Krishna, Falak Chhaya, and K Madhava Krishna. Reconstructing vehicles from a single image: Shape priors for road scene understanding. In *International Conference on Robotics and Automation*, pages 724–731. IEEE, 2017. 2

[32] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 3

[33] Liang Peng, Xiaopei Wu, Zheng Yang, Haifeng Liu, and Deng Cai. Did-m3d: Decoupling instance depth for monocular 3d object detection. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 71–88. Springer, 2022. 1, 5

[34] Liang Peng, Senbo Yan, Boxi Wu, Zheng Yang, Xiaofei He, and Deng Cai. Weakm3d: Towards weakly supervised monocular 3d object detection. *International Conference on Learning Representations*, 2022. 1, 2, 4, 5, 6, 7, 8

[35] Zengyi Qin, Jinglu Wang, and Yan Lu. Monogrnet: A geometric reasoning network for monocular 3d object localization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 2, 6, 7

[36] Zengyi Qin, Jinglu Wang, and Yan Lu. Weakly supervised 3d object detection from point clouds. In *Proceedings of the ACM International Conference on Multimedia*, pages 4144–4152, 2020. 2, 6

[37] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564, 2021. 1

[38] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 3

[39] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel Lopez Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection: From single to multi-class recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1219–1231, 2020. 1, 5

[40] Dawei Sun, Anbang Yao, Aojun Zhou, and Hao Zhao. Deeply-supervised knowledge synergy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6997–7006, 2019. 3

[41] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 3

[42] Runzhou Tao, Wencheng Han, Zhongying Qiu, Cheng-zhong Xu, and Jianbing Shen. Weakly supervised monocular 3d object detection using multi-view projection and direction consistency. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 3, 4, 5, 6, 7

[43] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. 3

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 4

[45] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. 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*, pages 8445–8453, 2019. 2

[46] Ting-Bing Xu and Cheng-Lin Liu. Data-distortion guided self-distillation for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5565–5572, 2019. 3

[47] Zhendong Yang, Ailing Zeng, Zhe Li, Tianke Zhang, Chun Yuan, and Yu Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[48] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. *International Conference on Learning Representations*, 2020. 2

[49] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2403–2412, 2018. 6

[50] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 13876–13885, 2020. 3

[51] Sergey Zakharov, Wadim Kehl, Arjun Bhargava, and Adrien Gaidon. Autolabeling 3d objects with differentiable rendering of sdf shape priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12224–12233, 2020. 6

[52] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019. 3

[53] Renrui Zhang, Han Qiu, Tai Wang, Xuanzhuo Xu, Ziyu Guo, Yu Qiao, Peng Gao, and Hongsheng Li. Monodetr: Depth-aware transformer for monocular 3d object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2, 6

[54] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11953–11962, 2022. 3

[55] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 5, 6, 7

[56] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monocular 3d object detection: An extrinsic parameter free approach. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[57] Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. *Advances in Neural Information Processing Systems*, 31, 2018. 3