StereoDistill: Pick the Cream from LiDAR for Distilling Stereo-Based 3D Object Detection

Zhe Liu[1], **Xiaoqing Ye**[2], **Xiao Tan**[2], **Errui Ding**[2], **Xiang Bai**[1]*

[1]Huazhong University of Science and Technology
[2]Baidu Inc., China
zheliu1994@hust.edu.cn, yxq@whu.edu.cn, tanxchong@gmail.com, dingerrui@baidu.com, xbai@hust.edu.cn

Abstract

In this paper, we propose a cross-modal distillation method named StereoDistill to narrow the gap between the stereo and LiDAR-based approaches via distilling the stereo detectors from the superior LiDAR model at the response level, which is usually overlooked in 3D object detection distillation. The key designs of StereoDistill are: the X-component Guided Distillation (XGD) for regression and the Cross-anchor Logit Distillation (CLD) for classification. In XGD, instead of empirically adopting a threshold to select the high-quality teacher predictions as soft targets, we decompose the predicted 3D box into sub-components and retain the corresponding part for distillation if the teacher component pilot is consistent with ground truth to largely boost the number of positive predictions and alleviate the mimicking difficulty of the student model. For CLD, we aggregate the probability distribution of all anchors at the same position to encourage the highest probability anchor rather than individually distill the distribution at the anchor level. Finally, our StereoDistill achieves state-of-the-art results for stereo-based 3D detection on the KITTI test benchmark and extensive experiments on KITTI and Argoverse Dataset validate the effectiveness.

Introduction

3D detectors equipped with LiDAR points (Shi, Wang, and Li 2019; Yang et al. 2020; Deng et al. 2020; Chen et al. 2017b; Huang et al. 2020; Liu et al. 2022) for autonomous driving have presented outperforming performance. However, LiDAR sensors usually have a high cost and sensitivity to weather, which limit their application. Alternatively, stereo cameras are capturing increasing interest thanks to their good trade-off in low cost and accuracy. There is still a huge performance gap between stereo-based and cutting-edge LiDAR-based 3D detection methods due to the inaccurate depth estimation by stereo matching. A question naturally arises: can the LiDAR model help to improve the performance of the stereo model?

Knowledge distillation (KD) (Hinton et al. 2015) might be a promising solution for this question, which guides the student model to mimic the knowledge of the teacher model for performance improvement or model compression. The current KD methods of object detection can be mainly classified into the feature-based and response-based streams, in which the former carry out distillation at the feature level (Zagoruyko and Komodakis 2017; Romero et al. 2014; Huang and Wang 2017; Heo et al. 2019; Ye et al. 2020; Du et al. 2020) for enforcing the consistency of feature representations between the teacher-student pair whereas the latter adopts the confident prediction from the teacher model as soft targets in addition to the hard ground truth supervision (Yuan et al. 2020; Zheng et al. 2022; Dai et al. 2021). However, directly migrating the existing KD methods to LiDAR-to-stereo cross-modal distillation is less effective due to the huge gap between the two extremely different modalities. The pioneering work LIGA (Guo et al. 2021) boosted the performance of stereo-based models by applying fine-grained feature-level distillation under the guidance of LiDAR-based models. However, it found little benefit from the response-based distillation due to the erroneous and noisy predictions of the LiDAR teacher.

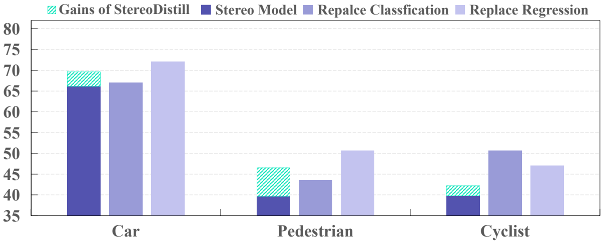

Figure 1: 3D detection performance (3D mAP) on KITTI *validation* set of LIGA (Guo et al. 2021) by replacing the regression and classification results of the stereo model (student) with the teacher LiDAR model SECOND (Yan, Mao, and Li 2018).

On the contrary, we argue that the response-level distillation is promising to shrink the gap in the cross-modal domain (*e.g.*, LiDAR point cloud and binocular images). For illustration, we first obtain the upper bound of the stereo model by replacing its prediction of 3D box regression and classification with the corresponding outputs of the LiDAR model (teacher). As shown in Figure 1, the stereo model with the replaced regression or classification predictions produces impressive results, demonstrating the potential of response-based distillation in the cross-modal domain. How-

*Corresponding Author.

ever, directly applying the vanilla response-level distillation is less effective, either by selecting the high-confident ((Yang et al. 2022)) or high-IoU 3D boxes (box-level) predicted from the LiDAR model as soft targets (Sun et al. 2020c). The reasons are two-fold: 1) unlike dense 2D images, much fewer high-IoU or high-confident boxes can be adopted as soft labels in a 3D scene due to the high sparsity of LiDAR point cloud; 2) the low-quality boxes discarded by one-size-fits-all thresholds contain underlying beneficial components (*e.g.*, center, size, or orientation angle) that have been overlooked.

To tackle the problem, we propose a novel X-component Guided Distillation (XGD) from the response level. The key idea of XGD is to first decompose a 3D box into sub-X-components (X can be center, size, or orientation angle) and retain the beneficial subcomponent as the soft targets if the vector between the teacher's X-component and the student's component is consistent with the vector between the ground truth and the student's, *i.e.*, the two vectors are acute-angled.

Moreover, we find that only one out of all anchors at the same position can be selected as being responsible for a foreground object in most cases due to the fact that there is usually no overlap among objects in real autonomous driving scenarios, which is different in the 2D domain. Motivated by this observation, we propose a simple and effective Cross-anchor Logit Distillation (CLD) for classification distillation in our StereoDistill to distill by aggregating the confidence distribution of all anchors to a unified distribution so as to highlight the highest probability anchor.

To summarize, our key contributions are as follows.

- We validate that the cross-modal knowledge distillation at the response level can boost the performance of stereo-based 3D object detection. The proposed X-component Guided Distillation (XGD) for regression avoids the negative effect of erroneous 3D boxes from the LiDAR model by keeping the beneficial X-component as soft targets under the guidance of acute-angled vectors.
- Given the fact that there is no overlap among objects in autonomous driving scenarios, we introduce the simple yet effective Cross-anchor Logit Distillation (CLD) for classification to aggregate the probability distribution of all anchors at the same position rather than distilling the distribution at anchor level.

Related Works

Stereo-based 3D Object Detection. The earlier methods (Li, Chen, and Shen 2019; Sun et al. 2020a; Xu et al. 2020) achieve stereo 3D detection based on a strong 2D detector (Ren et al. 2015; He et al. 2017), which does not fully explore the 3D information, leading to suboptimal performance. To introduce more 3D information, (Wang et al. 2019; You et al. 2020; Qian et al. 2020) try to convert the estimated depth maps combined with the corresponding image to pseudo point clouds and then can apply the existing LiDAR-based 3D detectors (Yan, Mao, and Li 2018; Lang et al. 2019) to detect 3D boxes. However, directly applying pseudo point clouds for 3D detection might bring erroneous localization due to the limitation of depth estimation, leading to sub-optimal performance. To tackle this problem, the recent methods DSGN (Chen et al. 2020), CDN (Garg et al. 2020), DSGN++ (Chen et al. 2022) and PLUME (Wang et al. 2021) build cost volume (Flynn et al. 2016) to encode the implicit 3D geometry features instead of the raw pseudo point representations for 3D object detection. In this paper, we select the prominent DSGN as our stereo model and keep the same configuration with LIGA (Guo et al. 2021).

LiDAR-based 3D Object Detection. Due to the plentiful geometric structure information and accurate depth information from LiDAR sensors, LiDAR-based 3D detectors (Shi, Wang, and Li 2019; Yan, Mao, and Li 2018) usually achieve superior performance than the camera-based (Brazil and Liu 2019; Chen et al. 2016; Li et al. 2020; Simonelli et al. 2019; Chen et al. 2020). At present, the mainstream 3D detection methods are divided into two types according to the input data format, including point-based and voxel-based detectors. The point-based methods (Shi, Wang, and Li 2019; Yang et al. 2020) usually apply PointNets (Qi et al. 2017a,b) to deal with this problem of permutation invariance. The voxel-based methods (Yan, Mao, and Li 2018; Zhou and Tuzel 2018; Lang et al. 2019; Liu et al. 2020; Deng et al. 2020) convert the irregular 3D points to the regular voxel grids and employ 2D/3D convolution operation to estimate the final 3D boxes. In this paper, to better align the predictions with the stereo model DSGN (Chen et al. 2020), we choose the popular voxel-based detector SECOND (Yan, Mao, and Li 2018) as the LiDAR model.

Knowledge Distillation. Knowledge distillation (KD) is initially proposed by (Hinton et al. 2015), which can transfer knowledge from a larger network to a small network to promote the performance or achieve model compression for lightweight devices. Recently, (Dai et al. 2021; Yang et al. 2021; Chen et al. 2021; Zhang and Ma 2020) achieve feature-based distillation by focusing on the foreground area or considering a weight matrix for the features. LD (Zheng et al. 2022) implements the difficult problem of localization distillation from the response level by converting the regression of bounding boxes to the probability distribution representation. Besides, Cross-modal feature distillation approaches (Chong et al. 2022; Guo et al. 2021) are gaining popularity as a way to take advantage of the complementarity between different modalities. LIGA (Guo et al. 2021) is the first attempt to explore the fine-grained feature distillation from LiDAR to stereo 3D detector. However, LIGA fails to benefit the stereo model through the response-based distillation due to the erroneous targets from the LiDAR model. In this paper, we propose an X-component Guided Distillation (XGD) to deal with this problem by retaining the beneficial component which is consistent with ground truth.

Method

In this part, we introduce the proposed cross-modal distillation StereoDistill, which consists of the X-component Guided Distillation (XGD) and Cross-anchor Logit Distillation (CLD) at the response level. As shown in Figure 2, we present the pipeline of our StereoDistill, which employs a stereo model, DSGN (Chen et al. 2020) for instance, as the student network and a LiDAR model, SECOND (Yan, Mao,

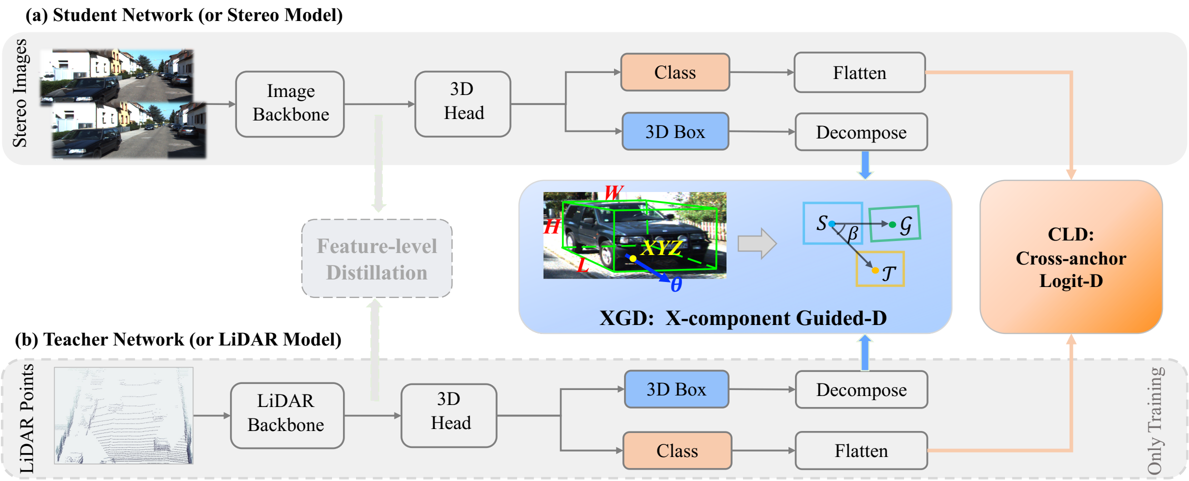
(a) Student Network (or Stereo Model)
Stereo Images
Image Backbone
3D Head
Class
3D Box
Flatten
Decompose
Feature-level Distillation
W
H
L
XYZ
θ
S
β
G
T
XGD: X-component Guided-D
CLD: Cross-anchor Logit-D
(b) Teacher Network (or LiDAR Model)
LiDAR Points
LiDAR Backbone
3D Head
3D Box
Class
Decompose
Flatten
Only Training

Figure 2: The pipeline of our proposed StereoDistill method. The student and the teacher model take the stereo images and LiDAR points as inputs, respectively. At the response level, X-component Guided Distillation (XGD) and Cross-anchor Logit Distillation (CLD) are applied to the 3D box regression and classification head, respectively. In XGD, we decompose the 3D box into sub-components, *i.e.*, size (HWL), center (XYZ) and rotation angle (θ) and keep the components as soft targets if the vectorial angle between teacher-student and GT-student pair is acute. In CLD, we flatten the confidence scores of all anchors falling in the same position and convert them to a unified distribution to highlight the most valuable anchor.

and Li 2018) for instance, as the teacher network only for training. Although StereoDistill contains the feature-level and response-level distillations, our main contribution focuses on the response-level distillation since the effectiveness on the feature-level has been illustrated in LIGA (Guo et al. 2021). For the feature level, we mainly revise the feature distillation in LIGA (Guo et al. 2021) by introducing the attention weight of features (Zagoruyko and Komodakis 2017) and the relationship among instance features (Hou et al. 2020) to further improve the performance, which is regarded as our baseline (named Improved LIGA). For more details, please refer to our supplementary materials.

For the response-based distillation, however, the predicted boxes (box-level) from a teacher network inevitably contain false predictions. Therefore, using all predicted boxes directly without any purifying process is likely harmful to the student network and results in a sub-optimal solution (Guo et al. 2021). To resolve this problem, we propose a novel XGD to preserve the beneficial X-component (*e.g.*, center, size and angle) decomposed from a box through the proposed positive component updating algorithm. In addition, we notice that only one out of all anchors at the same position can usually be selected as being responsible for a foreground object in autonomous driving scenarios. Thus, CLD is proposed to highlight the highest probability anchor across all anchors at the same position. Next, we introduce the proposed XGD and CLD in detail.

X-component Guided Distillation. As we all know, the LiDAR model has an inherent advantage in localization since the LiDAR sensor can provide more accurate geometrical information and depth information. However, the final predictions from the teacher model benefit little from training the stereo network (Guo et al. 2021). The main reason is that the erroneous regression of the teacher model may guide the student model to learn in a detrimental direction. Although an available solution is to only keep these high-quality boxes for distillation, it brings two flaws. One is that high-quality boxes are too few, resulting in inefficient distillation. The other is that some low-quality discarded boxes can also provide the estimated beneficial component through further decomposing a 3D box into three components (the center position, the size, and the orientation angle). To be more intuitive, we take the center position as an example and show

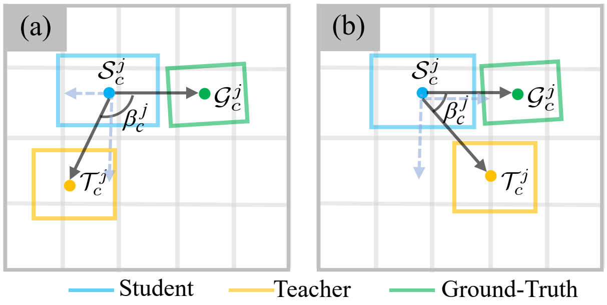

Figure 3: Our X-component Guided Distillation (take the center component as an example to illustrate whether the teacher's prediction is beneficial to the student. Case (a) depicts an obtuse angle between the student-to-GT vector and the student-to-teacher vector, showing that the teacher is inconsistent with the GT. Conversely, in Case (b) we observe an acute angle between the two vectors, validating that it is beneficial to be adopted as soft targets to guide the student to regress towards the direction of GT.

Algorithm 1: positive component updating

Input: Boxes of teacher $\mathcal{B}_t = (\mathcal{T}_c, \mathcal{T}_s, \mathcal{T}_o)$, Boxes of student $\mathcal{B}_s = (\mathcal{S}_c, \mathcal{S}_s, \mathcal{S}_o)$, Boxes of GT $\mathcal{B}_g = (\mathcal{G}_c, \mathcal{G}_s, \mathcal{G}_o)$. The number of assigned positive boxes N_{pos}.

Output: Updated boxes of teacher $\mathcal{B}_{t*}$.

```
Let $\mathcal{T}_{c*}, \mathcal{T}_{s*}, \mathcal{T}_{o*}$ =[], [], []
for $j \in \{1, 2, ..., N_{\text{pos}}\}$ do
    Compute $\cos\beta_c^j$, $\cos\beta_s^j$ and $\cos\beta_o^j$ by the formula (1)
    for $x \in \{c, s, o\}$ do
        if $\cos\beta_x^j > 0$ then
            $\mathcal{T}_{x*}^j \leftarrow \mathcal{T}_x^j$
        else
            $\mathcal{T}_{x*}^j \leftarrow \mathcal{S}_x^j$; # Disable the harmful X-component
        end if
    end for
end for
$\mathcal{B}_{t*}$ = $(\mathcal{T}_{c*}, \mathcal{T}_{s*}, \mathcal{T}_{o*})$
return $\mathcal{B}_{t*}$
```

the harmful and beneficial predicted center position from the teacher model in Figure 3 (a) and (b), respectively.

Motivated by the above observation, we propose a novel component Guided Distillation (XGD) to tackle this problem. XGD first obtains all the boxes predicted from the assigned positive anchors to keep more valuable 3D boxes. Then XGD selects the 'soft boxes' from these predicted boxes at the X-component level rather than the box level. Specifically, for the predicted j^{th} 3D box of the teacher network, we decompose $\mathcal{B}_t^j = (\mathcal{T}_c^j, \mathcal{T}_s^j, \mathcal{T}_o^j)$ into three components, where $\mathcal{T}_c^j = (x_t^j, y_t^j, z_t^j)$, $\mathcal{T}_s^j = (w_t^j, h_t^j, l_t^j)$, and $\mathcal{T}_o^j = \theta_t^j$, where $\mathcal{T}_c^j$ is the center position of the box along X, Y and Z axes, $\mathcal{T}_s^j$ represents the size including the width, height and length of the 3D box and $\mathcal{T}_o^j$ means the orientation angle of the 3D box. Similarly, we can define the predicted box $\mathcal{B}_s^j = (\mathcal{S}_c^j, \mathcal{S}_s^j, \mathcal{S}_o^j)$ from the student network and the corresponding ground-truth (GT) assigned boxes of $\mathcal{B}_g^j = (\mathcal{G}_c^j, \mathcal{G}_s^j, \mathcal{G}_o^j)$. Then, we can judge whether the estimated center $\mathcal{T}_c^j$ from the LiDAR model is beneficial to the stereo model by measuring the cosine value of $\mathcal{T}_c^j - \mathcal{S}_c^j$ and $\mathcal{G}_c^j - \mathcal{S}_c^j$, which can be formulated as:

$$\cos\beta_c^j = \frac{(\mathcal{T}_c^j - \mathcal{S}_c^j)(\mathcal{G}_c^j - \mathcal{S}_c^j)^T}{\|\mathcal{T}_c^j - \mathcal{S}_c^j\|_2 \|\mathcal{G}_c^j - \mathcal{S}_c^j)^T\|_2} \tag{1}$$

Where β_c^j is the angle between the vector of $\mathcal{T}_c^j - \mathcal{S}_c^j$ and $\mathcal{G}_c^j - \mathcal{S}_c^j$. Here, when β_c^j is an acute angle (or $\cos\beta_c^j > 0$), we think the provided center regression $\mathcal{T}_c^j$ from the teacher model can guide the student model regress a more accurate center position. Similarly, we can obtain the β_s^j and β_o^j for the size and angle components following the formulation (1). Then, the final 'soft boxes' $\mathcal{B}_{t*}$ is produced by our positive component updating in Algorithm 1. Finally, we employ 3D IoU loss (Zhou et al. 2019) with rotation as the soft regression term since 3D IoU can comprehensively evaluate the quality of a bounding box. The XGD loss can be computed as:

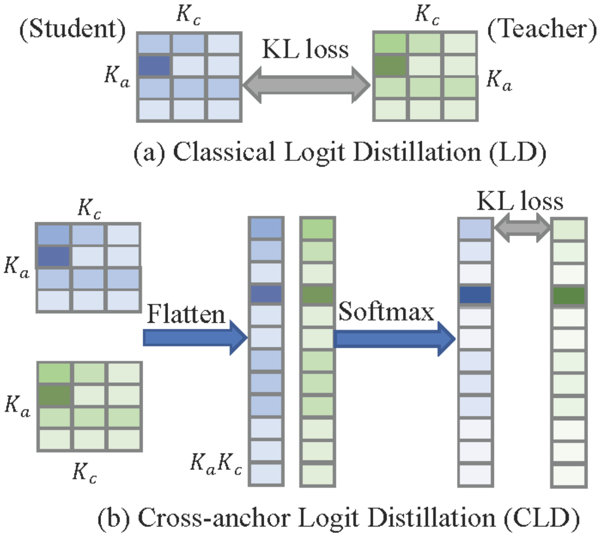

Figure 4: The process of the Classical Logit Distillation and our Cross-anchor Logit Distillation. The confidence scores from the student network and the teacher network are marked in blue and green, respectively. And the darker the color, the higher the confidence.

$$\mathcal{L}_{\text{xgd}} = \sum_{j=1}^{N_{\text{pos}}} (1 - \text{IoU3D}(\mathcal{B}_s^j, \mathcal{B}_{t*}^j)), \tag{2}$$

where N_{pos} is the number of the positive anchors in the stereo model and $\text{IoU3D}(\mathcal{B}_s^j, \mathcal{B}_{t*}^j)$ denotes the 3D IoU between $\mathcal{B}_s^j$ and $\mathcal{B}_{t*}^j$.

Cross-anchor Logit Distillation. Some distillation methods (Chen et al. 2017a; Dai et al. 2021; Sun et al. 2020c) via the classification probability usually bring benefits to final results for the 2D detection task, where these distillations are only carried out for positive boxes. However, our distillation is carried out in all foreground regions since the LiDAR model generates fewer positive 3D samples compared with 2D detection counterparts. Moreover, another distinct characteristic of the 3D detection task against 2D detection lies in the fact that it is rare to find a conflict or overlapping among 3D boxes in autonomous driving scenarios. That is to say, distinct anchors lying in the same position are designed for different objects with different scales and aspect ratios, and hence only one out of these anchors can be selected as being responsible for a foreground object in most cases. However, these classical logit distillation approaches (Chen et al. 2017a; Dai et al. 2021; Sun et al. 2020c) designed for 2D detection tasks treat anchors separately and do not work well in the 3D detection task, shown in Figure. 4 (a). Given that, we propose a Cross-anchor Logit Distillation (CLD) approach to highlight the most representative anchor from all anchors in the same position by converting the confidence distribution of each anchor to a unified distribution, whose process is described in Figure. 4 (b). Specifically,

Modality	Method	Car $\text{AP}_{0.7}$			Pedestrian $\text{AP}_{0.5}$			Cyclist $\text{AP}_{0.5}$		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
LiDAR	MV3D (Chen et al. 2017b)	74.97	63.63	54.00	–	–	-	–	–	–
	SECOND (Yan, Mao, and Li 2018)	83.34	72.55	65.82	–	–	–	–	–	–
	AVOD-FPN (Ku et al. 2018)	83.07	71.76	65.73	50.46	42.27	39.04	63.76	50.55	44.93
Stereo	Stereo R-CNN (Li, Chen, and Shen 2019)	47.58	30.23	23.72	–	–	–	–	–	–
	Pseudo-Lidar (Wang et al. 2019)	54.53	34.05	28.25	–	–	–	–	–	–
	ZoomNet (Xu et al. 2020)	55.98	38.64	30.97	–	–	–	–	–	–
	Pseudo-LiDAR++ (You et al. 2020)	61.11	42.43	36.99	–	–	–	–	–	–
	CDN (Garg et al. 2020)	74.52	54.22	46.36	–	–	–	–	–	–
	SNVC (Li et al. 2022)	78.54	61.34	54.23	–	–	–	–	–	–
	OC-Stereo (Pon et al. 2020)	55.15	37.60	30.25	24.48	17.58	15.60	29.40	16.63	14.72
	YOLOStereo3D (Liu, Wang, and Liu 2021)	65.68	41.25	30.42	28.49	19.75	16.48	–	–	–
	Disp-RCNN (Sun et al. 2020a)	68.21	45.78	37.73	37.12	25.80	22.04	40.05	24.40	21.12
	DSGN (Chen et al. 2020)	73.50	52.18	45.14	20.53	15.55	14.15	27.76	18.17	16.21
	CG-Stereo (Li, Ku, and Waslander 2020)	74.39	53.58	46.50	33.22	24.31	20.95	47.40	30.89	27.23
	LIGA (Guo et al. 2021)	81.39	64.66	57.22	40.46	30.00	27.07	54.44	36.86	32.06
	StereoDistill (Ours)	**81.66**	**66.39**	**57.39**	**44.12**	**32.23**	**28.95**	**63.96**	**44.02**	**39.19**

Table 1: 3D Detection results on the KITTI *test* benchmark. AP_{thr} means the threshold value of 3D IoU between the prediction and the ground truth as *thr*. 'Mod.' is short for Moderate.

Method	Car $\text{AP}_{0.5}$		Pedestrian $\text{AP}_{0.25}$	
	3D	BEV	3D	BEV
SECOND (2018) (***T***)	80.67	87.06	48.96	49.14
DSGN (2020) (***S***)	33.68	42.76	8.58	8.95
LIGA (2021)	34.37	45.47	10.76	11.01
Ours	**37.55**	**46.99**	**13.70**	**14.04**

Table 2: Car and Pedestrian detection results on Argoverse *validation* set with the evaluation metric of 11 recall positions. ***T*** and ***S*** denote the teacher and the student.

we first reshape the output confidence map of the teacher network $P_t \in \mathcal{R}^{N_{\text{fore}} \times K_c}$ as $P'_t \in \mathcal{R}^{M_{\text{fore}} \times (K_c K_a)}$, where $N_{\text{fore}} = M_{\text{fore}} K_a$. Here, M_{fore}, K_a and K_c represent the number of all foreground positions, the pre-defined anchors for each position and the categories on the 3D object detection task, respectively. Then, the softmax function is applied to further normalize the flattened confidence scores P'_t along the dimension of $K_c K_a$ and obtain the unified confidence distribution P^*_t across all anchors at the same position:

$$P^*_t = \text{softmax}(P'_t) \tag{3}$$

Similarly, we can get the confidence distribution P^*_s for the student network. Finally, the CLD loss can be computed by KL divergence:

$$\mathcal{L}_{\text{cld}} = \text{KL}(P^*_t, P^*_s), \tag{4}$$

Total Loss Function. We train the stereo model in an end-to-end manner, and the total loss function is as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ori}} + \mathcal{L}_{\text{xgd}} + \mathcal{L}_{\text{cld}}, \tag{5}$$

where $\mathcal{L}_{\text{ori}}$ denotes the loss function except the feature distillation loss in the LIGA (Guo et al. 2021). For training the LiDAR model, we adopt the same loss function with SECOND (Yan, Mao, and Li 2018).

Experiments

Experimental Datasets and Evaluation Metrics

KITTI. The KITTI dataset (Geiger, Lenz, and Urtasun 2012) includes 7,481 training and 7,518 testing stereo image pairs with the corresponding LiDAR point clouds. We further split the training data into *training* set with 3712 samples and a *validation* set with 3769 samples following (Chen et al. 2020; Qi et al. 2018; Shi, Wang, and Li 2019). The evaluation metric (Simonelli et al. 2019) adopts the mean Average Precision (mAP) with 40 recall positions. If not specified, the metric of all results in the following tables uses 40 recall positions. In this paper, we evaluate our method on the *validation* set and the *test* benchmark for three categories of Cars, Pedestrians and Cyclists under three difficulty levels (*e.g.*, Easy, Moderate, and Hard).

Argoverse. The Argoverse dataset (Chang et al. 2019) contains 3D detection and tracking annotations from 113 scenes. Different from the Waymo (Sun et al. 2020b) and Nuscnes (Caesar et al. 2020) datasets, Argoverse provides stereo image pairs, which can be adopted to verify the generality of our method. For convenience, we convert the Argoverse dataset to the format of KITTI following (Wang et al. 2020) and obtain a *training* set with 13122 samples and a *validation* set with 5015 samples. We adopt the same evaluation metric with KITTI.

Implementation Details

For the stereo model DSDN (Chen et al. 2020) and the LiDAR model SECOND (Yan, Mao, and Li 2018), we use the same network structure with LIGA (Guo et al. 2021) for fair comparisons. The stereo model is trained on 4 NVIDIA V100 GPUs with a batch size of 4 and is optimized by Adaptive Momentum Estimation (Adam) (Kingma and Ba 2014) with the initial learning rate, weight decay, and momentum factor set to 0.003, 0.01, and 0.9, respectively. Random horizontal flipping is adopted for data augmentation. For both

#	XGD	CLD	Car AP$_{3D}$			Pedestrian AP$_{3D}$			Cyclist AP$_{3D}$		
			Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
SECOND* (2018) (*teacher*)			90.72	81.66	78.78	69.45	62.39	56.49	84.97	64.06	60.21
DSGN† (2020) (*student*)			83.27	64.21	58.61	40.45	34.33	29.07	54.76	32.91	30.04
LIGA* (2021) (*paper*)			86.84	67.71	62.02	45.54	37.80	32.09	60.00	37.31	34.25
LIGA† (2021) (*reproduced*)			84.32	67.14	61.93	47.16	38.97	34.09	63.98	38.49	36.01
Improved LIGA† (2021) (*baseline*)			86.62	67.03	61.94	47.77	40.11	35.19	65.02	40.90	37.81
I	✓	–	86.78	67.65	62.43	52.89	45.37	39.40	66.25	41.38	38.27
II	–	✓	86.67	67.57	62.19	47.81	40.69	35.84	67.77	41.65	38.62
III	✓	✓	**87.57**	**69.75**	**62.92**	**55.19**	**46.76**	**40.42**	**69.43**	**42.31**	**39.10**
Improvement over *baseline*			**+0.95**	**+2.72**	**+0.98**	**+7.42**	**+6.65**	**+5.23**	**+4.41**	**+1.41**	**+1.29**

Table 3: Ablation studies for our proposed XGD and CLD on the KITTI *validation* set. $\star$ and $\dagger$ in this table denote the results reported in the paper and our reproduced results. 'Mod.' is short for Moderate.

Methods	Car AP$_{3D}$			Pedestrian AP$_{3D}$			Cyclist AP$_{3D}$		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
PointPillars (2019) (*teacher*)	88.89	78.47	75.38	62.84	56.01	51.87	82.58	62.06	58.37
DSGN (2020) (*student*)	83.27	64.21	58.61	40.45	34.33	29.07	54.76	32.91	30.04
LIGA (2021)	83.46	63.40	58.29	41.68	35.76	30.38	62.75	37.28	34.25
XGD + CLD	84.74	65.49	60.13	45.98	40.18	34.75	66.67	41.00	37.95
Improvement	**+1.28**	**+2.09**	**+1.84**	**+4.30**	**+4.42**	**+4.37**	**+3.92**	**+3.72**	**+3.70**
Improved LIGA (2021)	83.94	64.27	59.00	42.37	36.84	31.54	62.84	37.71	35.12
XGD + CLD	**85.24**	**67.62**	**60.72**	**47.81**	**40.69**	**34.78**	**67.10**	**41.19**	**38.13**
Improvement	**+1.30**	**+3.35**	**+1.72**	**+5.44**	**+3.85**	**+3.24**	**+4.26**	**+3.48**	**+3.01**

Table 4: Generality of our StereoDistill (XGD + CLD) on the KITTI *validation* set. We select the popular LiDAR model PointPillars (Lang et al. 2019) as the teacher model. 'Mod.' is short for Moderate.

the KITTI dataset and the Argoverse dataset, we employ the range of the detection area to [-30, 30], [-1, 3], [2, 59.6] meters along the X (right), Y (down), Z (forward) axis in the camera coordinate. The voxel size of the LiDAR model is (0.2, 0.2, 0.2) meters and the volume size of the stereo model is (0.05, 0.1, 0.05) meters. All experiments are conducted on a single model for multiple categories. For more details, please refer to our supplementary materials.

Comparisons with State-of-the-art Methods

Evaluation on KITTI. In Table 1, we present quantitative comparison with the leading stereo-based 3D detectors and several popular LiDAR-based 3D detectors on the KITTI *test* benchmark. Our method outperforms the SOTA model LIGA (Guo et al. 2021) with 1.73%, 2.23% and 7.16% 3D mAP on Cars, Pedestrian and Cyclists at the moderate difficulty level, without introducing any extra cost during inference. Our StereoDistill even surpasses the LiDAR-based method MV3D (Chen et al. 2017b) with 3D mAP of 2.76% on Cars. These superior results demonstrate the effectiveness of our StereoDistill. For visualization, please refer to our supplementary materials.

Evaluation on Argoverse. To further verify the generality of our proposed method, we conduct experiments on the Argoverse dataset. For fair comparisons, we adopt the same network of student and teacher models with LIGA (Guo et al. 2021) and also re-implement LIGA under the same setting on the Argoverse dataset. In Table 2, we present the results with the 3D IoU thresholds of 0.5 and 0.25 for both the BEV and 3D detection on moderate Cars and Pedestrians. Our method exceeds LIGA with 3D mAPs of 3.18% and 2.94% and BEV mAPs of 1.52% and 3.03% on Cars and Pedestrians, which validates the generality of our method.

Ablation Studies

Ablation Studies on StereoDistill. In this part, we verify the effectiveness of the proposed compositions, including the XGD for regression, CLD for classification, and their combinations in StereoDistill. The baseline model is our improved LIGA (Guo et al. 2021) by further enhancing the feature distillation (refer to our supplementary materials). In Table 3, by comparing (*I*), (*III*) with the baseline model, the proposed XGD and CLD bring consistent improvements over the baseline on all difficulty levels for three categories, which demonstrates their effectiveness. Note that the proposed XGD greatly boosts the detection performance on small objects (*e.g.*, Pedestrians), which requires more accurate regression. It illustrates that the X-component guided distillation can indeed transfer superior location awareness from the LiDAR model to the stereo model so as to obtain better performance. Integrating these two ingredients, StereoDistill in (*III*) outperforms the baseline with remarkable margins of 2.72%, 6.65% and 1.41% 3D mAP on the moderate Cars, Pedestrians and Cyclists, respectively.

Generality of StereoDistill. To verify the generality of StereoDistill, we replace the common teacher network SEC-

#	Methods	Cars	Pedestrians	Cyclists
I	Baseline	67.57	40.69	41.65
II	XGD-Center	68.22	46.33	42.19
III	XGD-Size	68.02	42.46	40.71
IV	XGD-Angle	67.84	44.11	41.18
V	**XGD**	**69.75**	**46.76**	**42.31**
VI	High-quality boxes	67.72	43.65	42.11
VII	**Positive anchors (Ours)**	**69.75**	**46.76**	**42.31**

Table 5: Ablation studies for XGD. The results are evaluated with 3D mAP on the moderate difficulty level for Cars, Pedestrians and Cyclists, respectively. XGD-* means the manner of only adopting * for computing XGD loss.

#	Methods	Cars	Pedestrians	Cyclists
I	Baseline	67.65	45.37	41.38
II	Positive Anchors	67.66	45.71	**42.32**
III	**All Foregrounds (Ours)**	**69.75**	**46.76**	42.31
IV	Classical (Chen et al. 2017a)	67.74	46.18	40.69
V	**CLD (Ours)**	**69.75**	**46.76**	**42.31**

Table 6: Ablation studies for CLD. The results are evaluated with 3D mAP on the moderate difficulty level.

Methods	Cars	Pedestrians	Cyclists	Mean
DSGN (LIGA wo FD)	63.32	34.23	30.26	42.60
+ FD	67.14	38.97	38.49	48.20
Improvement	+3.82	+4.74	+8.23	+5.60
+ our RD	67.42	45.28	37.66	50.12
Improvement	+4.10	+11.05	+7.4	+7.52
StereoDistill	69.75	46.76	42.31	52.94
Improvement	**+6.43**	**+12.53**	**+12.05**	**+10.34**

Table 7: Comparisons for the feature-based and response-based distillation. The results are evaluated with 3D mAP on moderate Cars, Pedestrians and Cyclists. FD and RD are short for feature-based distillation in LIGA (Guo et al. 2021) and the response-based distillation in StereoDistill.

OND (Yan, Mao, and Li 2018) with the other popular 3D detector Pointpillars (Lang et al. 2019). In Table 4, we provide two baseline settings: the original LIGA (Guo et al. 2021) in *Line 4* and the improved LIGA in *Line 7*. Not surprisingly, our StereoDistill yields obvious performance gains on all difficulty levels for three categories, which further the superiority and generality of our proposed XGD and CLD.

Analysis of XGD. In Table 5, we conduct extensive ablation studies to analyze the effectiveness of XGD. The baseline (*I*) is our StereoDistill without XGD loss. Then, we decompose the regression of 3D boxes into three components including the center position, the size, and the orientation angle to analyze the effect of each component on XGD. We observe that XGD with only the center position (*II*) exhibits the most competitive performance of the three (*II*, *III*, *IV*), which illustrates the positive guidance of the center position is the crucial component to help the student model to acquire more beneficial localization information. Finally, combined with these three components, XGD (*V*) exceeds the baseline (*I*) with 3D mAP of 2.08%, 6.07% and 0.66% on moderate Cars, Pedestrians, and Cyclists, demonstrating the superiority of XGD. Moreover, compared with retaining high-quality boxes in (*VI*) whose confidence scores are greater than 0.3 (a proper threshold), our manner of adopting all positive anchors in (*VII*) has obvious gains on Cars and Pedestrians. In the real scene, these two categories usually occupy a much larger number than Cyclists, which means that there may be more false positives on Cars and Pedestrians. This demonstrates that our XGD provides a reasonable workaround to deal with some low-quality boxes by retaining the beneficial X-component but discarding the harmful X-component decomposed from a 3D box.

Effectiveness of CLD. In Table 6, we first present the results of distilling the confidence distribution from two alternative regions. The way based on the foreground (*III*) exceeds the approach by only considering the positions from positive anchors (*II*) on average, which illustrates the importance of introducing the useful foreground positions beyond the positions of the positive anchors. Furthermore, we provide a classical logit distillation (*IV*) (Sun et al. 2020c) as a comparison, which individually treats the confidence distribution of each anchor from each position. It can be observed that our CLD boosts the performance with mAP of 2.01%, 0.58% and 1.62% on Cars, Pedestrian and Cyclists, clearly demonstrating the effectiveness of underlining the confidence distribution for the best competitive anchor from all anchors in a position.

Comparison of Different Distillation. In Table 7, we individually present the comparisons for adopting the feature-based in LIGA (Guo et al. 2021) or response-based distillation in StereoDistill based on the stereo model DSGN (Chen et al. 2020). First, these two distillations can consistently boost performance over the baseline DSGN (Chen et al. 2020). Then, the proposed response-based distillation of our XGD and CLD in StereoDistill even outperforms the feature-based distillation in LIGA (Guo et al. 2021) with 3D mAP of 1.92% on average under the same setting, which further demonstrates the effectiveness of the proposed XGD and CLD. Moreover, combined with the response-based distillation and the feature-based distillation, our StereoDistill produces superior performance over the baseline model DSGN with a 3D mAP of 10.34% on average. This reveals that superior feature representations often lead to better responses, which in turn can further facilitate feature learning.

Conclusions

This paper presents an effective cross-modal distillation approach termed StereoDistill from the response levels for the stereo 3D detection task. The extension ablation studies demonstrate the superiority of our proposed X-component Guided Distillation (XGD) for regression and Cross-anchor Logit Distillation (CLD) for classification. In the future, we wish StereoDistill can be applied to more 3D detectors to improve their performance.

Acknowledgements

This work was supported by the National Science Fund of China for Distinguished Young Scholars (No. 62225603).

References

Brazil, G.; and Liu, X. 2019. M3d-rpn: Monocular 3d region proposal network for object detection. In *CVPR*, 9287–9296.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 11621–11631.

Chang, M.-F.; Lambert, J. W.; Sangkloy, P.; Singh, J.; Bak, S.; Hartnett, A.; Wang, D.; Carr, P.; Lucey, S.; Ramanan, D.; and Hays, J. 2019. Argoverse: 3D Tracking and Forecasting with Rich Maps. In *CVPR*.

Chen, G.; Choi, W.; Yu, X.; Han, T.; and Chandraker, M. 2017a. Learning efficient object detection models with knowledge distillation. *NeurIPS*, 30.

Chen, P.; Liu, S.; Zhao, H.; and Jia, J. 2021. Distilling knowledge via knowledge review. In *CVPR*, 5008–5017.

Chen, X.; Kundu, K.; Zhang, Z.; Ma, H.; Fidler, S.; and Urtasun, R. 2016. Monocular 3d object detection for autonomous driving. In *CVPR*, 2147–2156.

Chen, X.; Ma, H.; Wan, J.; Li, B.; and Xia, T. 2017b. Multiview 3D Object Detection Network for Autonomous Driving. In *CVPR*.

Chen, Y.; Huang, S.; Liu, S.; Yu, B.; and Jia, J. 2022. DSGN++: Exploiting Visual-Spatial Relation for Stereo-Based 3D Detectors. *TPAMI*.

Chen, Y.; Liu, S.; Shen, X.; and Jia, J. 2020. Dsgn: Deep stereo geometry network for 3d object detection. In *CVPR*, 12536–12545.

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*.

Dai, X.; Jiang, Z.; Wu, Z.; Bao, Y.; Wang, Z.; Liu, S.; and Zhou, E. 2021. General instance distillation for object detection. In *CVPR*, 7842–7851.

Deng, J.; Shi, S.; Li, P.; Zhou, W.; Zhang, Y.; and Li, H. 2020. Voxel r-cnn: Towards high performance voxel-based 3d object detection. *arXiv preprint arXiv:2012.15712*, 1(2): 4.

Du, L.; Ye, X.; Tan, X.; Feng, J.; Xu, Z.; Ding, E.; and Wen, S. 2020. Associate-3Ddet: Perceptual-to-conceptual association for 3D point cloud object detection. In *CVPR*, 13329–13338.

Flynn, J.; Neulander, I.; Philbin, J.; and Snavely, N. 2016. Deepstereo: Learning to predict new views from the world's imagery. In *CVPR*, 5515–5524.

Garg, D.; Wang, Y.; Hariharan, B.; Campbell, M.; Weinberger, K. Q.; and Chao, W.-L. 2020. Wasserstein distances for stereo disparity estimation. *NeurIPS*, 33: 22517–22529.

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

Guo, X.; Shi, S.; Wang, X.; and Li, H. 2021. LIGA-Stereo: Learning LiDAR Geometry Aware Representations for Stereo-based 3D Detector. In *ICCV*, 3153–3163.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*, 2961–2969.

Heo, B.; Lee, M.; Yun, S.; and Choi, J. Y. 2019. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *AAAI*, volume 33, 3779–3787.

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

Hou, Y.; Ma, Z.; Liu, C.; Hui, T.-W.; and Loy, C. C. 2020. Inter-region affinity distillation for road marking segmentation. In *CV*, 12486–12495.

Huang, T.; Liu, Z.; Chen, X.; and Bai, X. 2020. Epnet: Enhancing point features with image semantics for 3d object detection. In *ECCV*, 35–52. Springer.

Huang, Z.; and Wang, N. 2017. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; and Waslander, S. 2018. Joint 3D Proposal Generation and Object Detection from View Aggregation. In *IROS*.

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 *CVPR*, 12697–12705.

Li, C.; Ku, J.; and Waslander, S. L. 2020. Confidence Guided Stereo 3D Object Detection with Split Depth Estimation. *IROS*.

Li, P.; Chen, X.; and Shen, S. 2019. Stereo r-cnn based 3d object detection for autonomous driving. In *CVPR*, 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 *ECCV*, 644–660. Springer.

Li, S.; Liu, Z.; Shen, Z.; and Cheng, K.-T. 2022. Stereo Neural Vernier Caliper. In *AAAI*, volume 36.

Liu, Y.; Wang, L.; and Liu, M. 2021. YOLOStereo3D: A Step Back to 2D for Efficient Stereo 3D Detection. In *ICRA*. IEEE.

Liu, Z.; Huang, T.; Li, B.; Chen, X.; Wang, X.; and Bai, X. 2022. EPNet++: Cascade bi-directional fusion for multimodal 3D object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, Z.; Zhao, X.; Huang, T.; Hu, R.; Zhou, Y.; and Bai, X. 2020. TANet: Robust 3D Object Detection from Point Clouds with Triple Attention. In *AAAI*.

Pon, A. D.; Ku, J.; Li, C.; and Waslander, S. L. 2020. Object-Centric Stereo Matching for 3D Object Detection. In *ICRA*, 8383–8389. IEEE.

Qi, C. R.; Liu, W.; Wu, C.; Su, H.; and Guibas, L. J. 2018. Frustum pointnets for 3d object detection from rgb-d data. In *CVPR*.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 30.

Qian, R.; Garg, D.; Wang, Y.; You, Y.; Belongie, S.; Hariharan, B.; Campbell, M.; Weinberger, K. Q.; and Chao, W.-L. 2020. End-to-end pseudo-lidar for image-based 3d object detection. In *CVPR*, 5881–5890.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28.

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.; Wang, X.; and Li, H. 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*, 770–779.

Simonelli, A.; Bulo, S. R.; Porzi, L.; López-Antequera, M.; and Kontschieder, P. 2019. Disentangling monocular 3d object detection. In *ICCV*, 1991–1999.

Sun, J.; Chen, L.; Xie, Y.; Zhang, S.; Jiang, Q.; Zhou, X.; and Bao, H. 2020a. Disp R-CNN: Stereo 3D Object Detection via Shape Prior Guided Instance Disparity Estimation. In *CVPR*.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020b. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2446–2454.

Sun, R.; Tang, F.; Zhang, X.; Xiong, H.; and Tian, Q. 2020c. Distilling object detectors with task adaptive regularization. *arXiv preprint arXiv:2006.13108*.

Wang, Y.; Chao, W.-L.; Garg, D.; Hariharan, B.; Campbell, M.; and Weinberger, K. Q. 2019. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, 8445–8453.

Wang, Y.; Chen, X.; You, Y.; Erran, L.; Hariharan, B.; Campbell, M.; Weinberger, K. Q.; and Chao, W.-L. 2020. Train in germany, test in the usa: Making 3d object detectors generalize. In *CVPR*, 11713–11723.

Wang, Y.; Yang, B.; Hu, R.; Liang, M.; and Urtasun, R. 2021. PLUME: Efficient 3D Object Detection from Stereo Images. *arXiv preprint arXiv:2101.06594*.

Xu, Z.; Zhang, W.; Ye, X.; Tan, X.; Yang, W.; Wen, S.; Ding, E.; Meng, A.; and Huang, L. 2020. Zoomnet: Part-aware adaptive zooming neural network for 3d object detection. In *AAAI*, 12557–12564.

Yan, Y.; Mao, Y.; and Li, B. 2018. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10): 3337.

Yang, J.; Shi, S.; Ding, R.; Wang, Z.; and Qi, X. 2022. Towards Efficient 3D Object Detection with Knowledge Distillation. *arXiv preprint arXiv:2205.15156*.

Yang, Z.; Li, Z.; Jiang, X.; Gong, Y.; Yuan, Z.; Zhao, D.; and Yuan, C. 2021. Focal and Global Knowledge Distillation for Detectors. *arXiv preprint arXiv:2111.11837*.

Yang, Z.; Sun, Y.; Liu, S.; and Jia, J. 2020. 3dssd: Point-based 3d single stage object detector. In *CVPR*, 11040–11048.

Ye, X.; Du, L.; Shi, Y.; Li, Y.; Tan, X.; Feng, J.; Ding, E.; and Wen, S. 2020. Monocular 3d object detection via feature domain adaptation. In *ECCV*, 17–34. Springer.

You, Y.; Wang, Y.; Chao, W.-L.; Garg, D.; Pleiss, G.; Hariharan, B.; Campbell, M.; and Weinberger, K. Q. 2020. Pseudo-LiDAR++: Accurate Depth for 3D Object Detection in Autonomous Driving. In *ICLR*.

Yuan, L.; Tay, F. E.; Li, G.; Wang, T.; and Feng, J. 2020. Revisiting knowledge distillation via label smoothing regularization. In *CVPR*, 3903–3911.

Zagoruyko, S.; and Komodakis, N. 2017. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *ICLR*.

Zhang, L.; and Ma, K. 2020. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *ICLR*.

Zheng, Z.; Ye, R.; Wang, P.; Ren, D.; Zuo, W.; Hou, Q.; and Cheng, M.-M. 2022. Localization Distillation for Dense Object Detection. In *CVPR*.

Zhou, D.; Fang, J.; Song, X.; Guan, C.; Yin, J.; Dai, Y.; and Yang, R. 2019. Iou loss for 2d/3d object detection. In *3DV*, 85–94. IEEE.

Zhou, Y.; and Tuzel, O. 2018. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In *CVPR*.