Beyond the limitation of monocular 3D detector via knowledge distillation

Yiran Yang[1,2,*], Dongshuo Yin[1,2,*], Xuee Rong[1,2], Xian Sun[1,2], Wenhui Diao [1†], Xinming Li[1]
[1]Key Laboratory of Network Information System Technology,
Aerospace Information Research Institute, Chinese Academy of Sciences
[2]School of Electronic, Electrical and Communication Engineering,
University of Chinese Academy of Sciences
{yangyiran19, yindongshuo19, rongxuee19}@mails.ucas.ac.cn
{sunxian, diaowh}@aircas.ac.cn, 13911729321@139.com

Abstract

Knowledge distillation (KD) is a promising approach that facilitates the compact student model to learn dark knowledge from the huge teacher model for better results. Although KD methods are well explored in the 2D detection task, existing approaches are not suitable for 3D monocular detection without considering spatial cues. Motivated by the potential of depth information, we propose a novel distillation framework that validly improves the performance of the student model without extra depth labels. Specifically, we first put forward a perspective-induced feature imitation, which utilizes the perspective principle (the farther the smaller) to facilitate the student to imitate more features of farther objects from the teacher model. Moreover, we construct a depth-guided matrix by the predicted depth gap of teacher and student to facilitate the model to learn more knowledge of farther objects in prediction level distillation. The proposed method is available for advanced monocular detectors with various backbones, which also brings no extra inference time. Extensive experiments on the KITTI and nuScenes benchmarks with diverse settings demonstrate that the proposed method outperforms the state-of-the-art KD methods.

1. Introduction

With the development of autonomous driving, 3D object detection has become a popular field. In general, multi-view-based[39, 20, 15, 14, 26, 27, 21, 42] and LiDAR-based approaches[17, 45] show impressive performance. But considering the massive expensive equipment demand of the above methods, monocular vision methods have attracted more attention in recent years. Moreover, lightweight models [47, 33] are more favored for practical deployments due to the constraints of hardware resources and the need for real-time. But a lightweight 3D monocular detector usually has a weak feature extraction ability and an inaccuracy depth estimation, which results in a not satisfying performance.

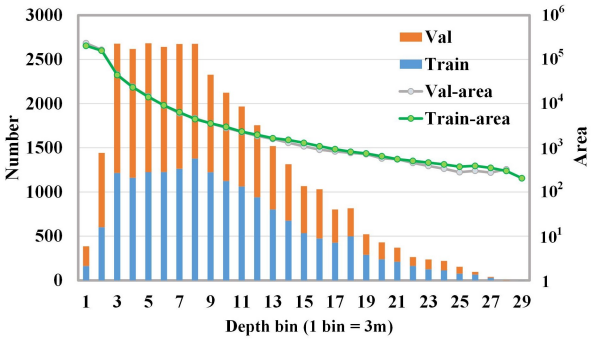


Figure 1: The statistics for object depth and area in KITTI. We set a 3m length in a bin, and the **number** and **area** represent the object amount and average pixel area, respectively. The logarithmic scale is used for area coordinates.

To tackle the issue, knowledge distillation (KD) [12] is proposed to facilitate the compact student model to learn implicit knowledge from the huge teacher model. KD can indeed improve the performance of the student model without extra inference costs, which motivates researchers to propose better distillation techniques in classification [1, 16, 5, 49, 22, 44] and detection [4, 19, 37, 35, 10, 8] task. Moreover, some 3D distillation methods [7, 13] also adopt the LiDAR data as the input for the teacher model to guide the student. But there is a cost associated with the acquisition of LiDAR data. Therefore, an interesting idea naturally arises: can we only use a vision-based strategy to optimize the student model like 2D distillation approaches?

2D distillation methods perform poorly on 3D tasks ow-

*Equal contribution.
†Corresponding author.

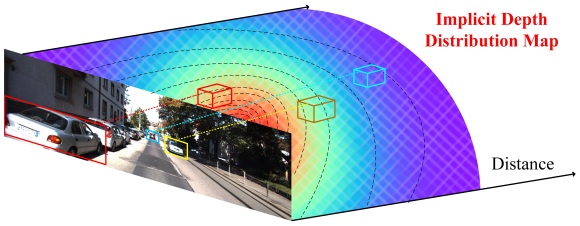


Figure 2: Implicit depth distribution map. An implicit depth distribution can be found in the image, which is the farther in the real world, the smaller in the image.

ing to the lack of spatial clues consideration. Distinguished from the planarity of the 2D task, depth is the key dimension in the 3D task. We believe that the model performance could be improved if depth information is taken into consideration in KD. Firstly, we analyze the depth and area distribution of objects in the KITTI dataset as illustrated in Fig. 1. The results present that distant objects account for an indispensable proportion of all objects and it also demonstrates that farther objects occupy fewer pixels and fewer features, fewer and low-quality features would be a disaster for downstream detection tasks. Therefore, some KD strategies need to be devised for promoting the distant objects' feature quality and detection results. But acquiring the distance in the monocular image usually is an ill-posed problem, and some researchers have taken some prior methods for estimation. In our observation illustrated in Fig. 2, although the image lost the depth dimension, an implicit depth distribution still can be acquired. We discover that distant object appears more frequently at the end of the road, which are close to the center of the camera axis. Moreover, the further the object is in the real world, the smaller the object is in the image. The two above observations motivate us to introduce the implicit depth in KD.

To better use the observed depth information, we propose a novel distillation framework for 3D monocular vision tasks, which is guided by implicit spatial distribution information. According to the implicit depth distribution map as illustrated in Fig. 2, we put forward a perspective-induced feature imitation module, which focuses more on learning farther object features. Specifically, a perspective matrix is designed according to the pixel distance between the farthest point in the image and other points, which endows more attention on the distant objects. Moreover, we also propose a depth-guided prediction distillation that introduces the depth information, including the depth of ground truth, student, and teacher prediction. We adopt the above depth information to construct the depth-guided matrix, which not only reduces the difference in classification distribution between the teacher and student model but also motivates the student to imitate the depth estimation of the teacher model. Extensive experimental results on the KITTI[9] and nuScenes[3] benchmarks prove that the proposed approach surpasses the previous SOTA methods on different teacher-student pairs with various backbones. Our main contributions can be summarized as follows:

1) We design a perspective matrix that follows the implicit depth distribution map to achieve better feature imitation of the distant object.

2) We devise a depth-guided prediction distillation method, which facilitates the student to learn the distant instance prediction from the teacher model.

3) We propose a unified 3D monocular distillation framework purely based on a visual scheme, which introduces implicit depth information to the knowledge-transferring process. Extensive experiments on KITTI and nuScenes demonstrate that the proposed method outperforms the previous SOTA distillation methods.

2. Related Work

2.1. 3D Monocular Detection

3DOP[6], MLFusion[41], and Deep3DBox [31] usually adopt the depth estimation subnet to assist the detection task. M3D-RPN[2] proposes a single-stage detector that uses a depth-aware convolution for 3D position features learning. MonDIS[34] presents a disentangling loss for better multitask learning. FQNet[24] measures the overlaps between objects and 3D projected proposals to choose the best proposals. RTM3D[18] predicts the nine perspective key points of the 3D bounding box and obtains 3D properties by the geometric relationship of 3D and 2D perspectives. FCOS3D[38] adopts FCOS[36] as baseline and adds new detection heads for 3D attribute predictions. PGD[40] incorporates a probabilistic representation to capture the depth uncertainty for better estimation. MonoEF [50] proposes to capture camera pose to formulate the detector free from extrinsic perturbation. MonoFlex[48] decouples the representations of objects and optimizes them respectively. GUPNet [30] proposes a geometry uncertainty projection net for better depth estimation. AutoShape[29] proposes an approach for incorporating the shape-aware 2D/3D constraints into the 3D detection. MonoCon [25] presents to learn monocular contexts as auxiliary tasks in training.

2.2. Knowledge Distillation

Knowledge distillation has been firstly proposed by Hinton *et al.* [12]. Since then, many approaches are proposed for better knowledge delivery. FitNet [1] introduces intermediate-level hints from the teacher hidden layers to guide the training process of the student. AT [16] designs a attention mechanism to deliver more valuable information. ReviewKD [5] proposes a knowledge review mechanism by the cross-stage connection paths. DKD [49] de-

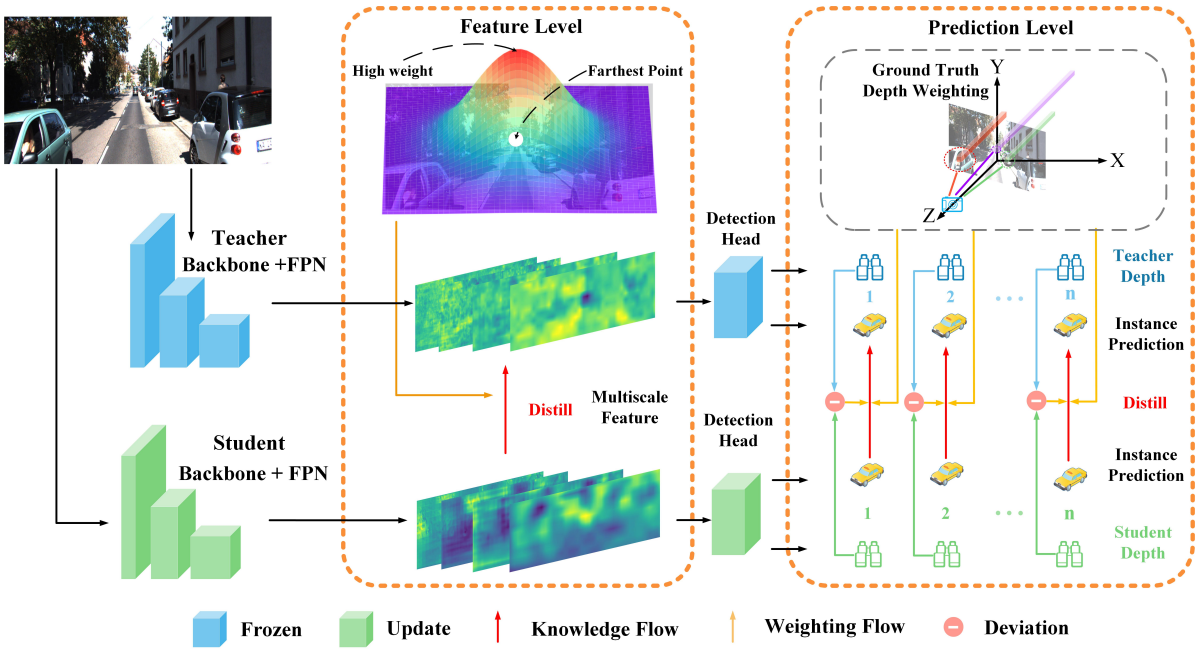


Figure 3: The distillation framework of our 3D monocular detection. Firstly, we adopt the implicit depth distribution map to endow more weight on farther object imitation at the feature level. Then, we introduce the depth information of the ground truth, student, and teacher estimation for assistance at prediction level distillation.

couples the target class knowledge distillation and non-target class knowledge distillation enabling them to play their roles more efficiently. Lin *et al.* [22] propose a one-to-all spatial matching knowledge distillation method considering the feature similarity between teacher and student models. MGD [44] proposes that teachers can improve students' representation by guiding students' feature recovery.

Besides, more work devote themselves to object detection distillation. LEOD [4] transfers the knowledge of featurem map and respone. Mimic [19] points out that RoI regions are more valuable for learning including positive and negative samples. FGFI [37] suggests that the student learns the fine-grained feature of the teacher model. TAR [35] proposes a gaussian mask to enhance the object information and suppress those background regions. DeFeat [10] suggests that both the background feature and object feature should be distilled together but assigned different weights. GID[8] proposes to distill the feature, response, and relation knowledge of discriminative instances. Monodistill[7] proposes to distill LiDAR signals to the monocular 3D detectors. CMKD [13] introduces the cross-modal information for better 3D distilaltion. In conclusion, existing vision-based methods are not explored in 3D monocular detection.

3. Methods

3.1. Overview

Knowledge distillation (KD) [12] is a promising approach to improve the performance of student model, which delivers dark knowledge except for the ground truth. Given the student model parameters θ_s and teacher model parameters θ_t. The dataset is D which contains images and labels. A standard KD process is formulated below,

$$KD(\theta_s) = argmin_{\theta_s}\mathcal{L}(I(\theta_s, D), I(\theta_t, D)), \quad (1)$$

where $\mathcal{L}$ is the loss function. The model input data and acquires information I, which contains the intermediate features F and final predictions P. Moreover, the KD transfers the knowledge by the $\mathcal{L}$, which means the student needs to imitate the information of the teacher model.

In our distillation framework, not only the feature maps from the feature level but also some responses at the prediction level are selected as the knowledge to facilitate the student model beyond its limitation. Both teacher and student models are monocular 3D detectors. They adopt the same network structure but use different backbones. The teacher is a well-trained frozen model and the student is trained from scratch during the distillation process. Fig. 3 illustrates the whole process of our distillation framework.

3.2. Perspective-induced Feature Imitation

Feature quality affected by model extraction ability usually decides the performance of the subsequent downstream task. A heavy teacher model has a better feature capture ability than a lightweight student model. Intermediate features distillation is introduced by FitNet [1] firstly, and most recent work [37, 35, 10, 43] have demonstrated the effectiveness. Since Feature Pyramid Network (FPN) [23] can

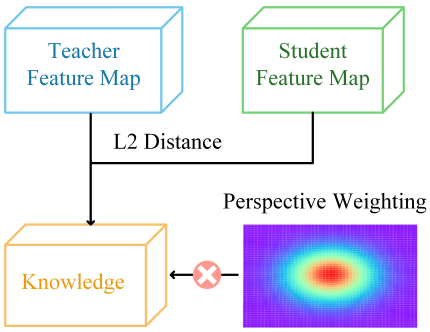


Figure 4: Perspective-induced feature imitation. We calculate the L2 distance of feature maps between teacher and student models and use the perspective matrix $\mathcal{M}$ to weigh the knowledge.

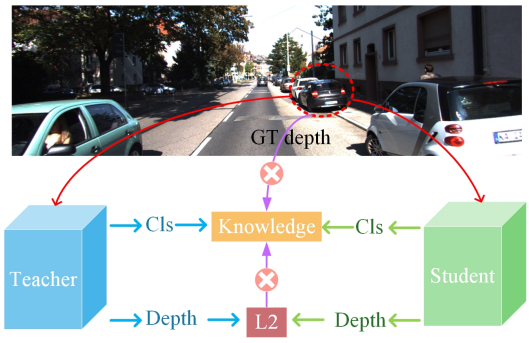


Figure 5: Depth-guided predcition distillation. We adopt ground truth depth "$\rho\left(\tau\left(\mathcal{D}\right)\right)$" and the depth deviation "$\phi\left(\left|\tau\left(P_s^d\right)-\tau\left(P_t^d\right)\right|\right)$" between teacher and student models to weigh the knowledge.

fuse features of different levels to obtain better detection effects, the existing work usually chooses the multi-scale feature maps output by FPN as distillation knowledge. Therefore, we also follow this strategy, and the feature imitations are defined as follows,

$$\begin{aligned}\mathcal{L}_f &= \frac{1}{L}\sum_{l=1}^{L}\frac{1}{N_l}\left\|F_t^l-\varphi(F_s^l)\right\|^2,\\ N_l &= W_l * H_l * C_l,\end{aligned} \tag{2}$$

where L represents the FPN level numbers and $F^l \in R^{W_l * H_l * C_l}$ refers the multiscale feature maps. φ is an adaption layer that ensures the feature between the teacher and student model is alignment. H_l, W_l, C_l denote the height, width, and channel of the l-th level features, respectively.

We want to combine the information of depth with knowledge distillation. But the depth estimation in the monocular image is an ill-posed issue, and some priors may help to get some help. According to the implicit depth distribution map, the farther the object is in the world, the smaller it is in the image. Generally, the input image is downsampled four times during the feature extraction, which means the distant object will have fewer features and results in inaccuracy detection. Therefore, the teacher with a heavy backbone can impart valuable feature knowledge to the student with a lightweight backbone. But classic feature imitations endow equal attention to all features, which is not suitable for 3D distillation affected by depth.

Although it is hard to directly acquire an accurate depth estimation from an image, we still can distinguish relative distance. As illustrated in Fig. 3, we observe that distant objects appear more often at the end of the road, which usually is in the center of the camera. Therefore, the white point in the image is farthest from the observer. With it as the center, the other points gradually approach the observer. Based on this observation, we propose a perspective-induced feature imitation that imposes more attention on features of farther objects in distillation. Assuming the farthest point p^* is (x^*, y^*) in the image, we adopt the 2D-Gauss function to transform the pixel distances to perspective matrix $\mathcal{M}$ below,

$$\mathcal{M} = A * e^{-\frac{(x/w-x^*/w)^2}{\sigma_x^2}-\frac{(y/h-y^*/h)^2}{\sigma_y^2}}, \tag{3}$$

where $x \in [0, w], y \in [0, h]$. σ_x^2 and σ_y^2 are the decay factor along the two directions. A refers to the amplitude. We set the farthest point in the center of an image which satisfies most situations. But for occlusion by close front car, we will decrease the occlusion region weight of the depth distillation matrix in practice. Then, multiscale $\mathcal{M}$ are acquired by operating in the multiscale features as Eq. 3.

Finally, we employ the multiscale perspective matrix $\mathcal{M}$ in the feature imitation as follows,

$$\begin{aligned}\mathcal{L}_f &= \frac{1}{L}\sum_{l=1}^{L}\frac{\mathcal{M}_l}{N_l}\left\|F_t^l-\varphi(F_s^l)\right\|^2,\\ N_l &= W_l * H_l * C_l,\end{aligned} \tag{4}$$

where F_t^l and F_s^l represents the multiscale features from the teacher and studnet model, respectively. Fig. 4 presents the workflow of $\mathcal{M}$ in the feature imitation.

3.3. Depth-guided Prediction Distillation

Except for the intermediate features, the predictions are also valuable for distillation. Hinton *et al.* [12] propose the category probabilities of the teacher predictions can be viewed as soft labels to help the student to learn the relationship between different classes. For example, cyclists and pedestrians have some similarities rather than with cars. In our framework, we mainly distill the classification relationship between the student and teacher models at the prediction level.

In addition, considering that depth estimation usually has a huge influence on the accuracy of 3D location in monocular detection, we propose to use depth estimation to guide the distillation at the prediction level. Firstly, the distillation strategy can reduce the difference in classification between the teacher and student model. Secondly, better depth estimation knowledge will be transferred to the student. To be specific, given a set of the positive sample K_{obj} in the image, we can get the depth predictions P_s^d and P_t^d of the student and teacher, respectively. Besides, $\mathcal{D}_{gt}$ is the corresponding depth ground truth for each sample. The depth-guided matrix is defined below,

$$\mathcal{D} = \phi\left(\left|\tau\left(P_s^d\right) - \tau\left(P_t^d\right)\right|\right) * \rho\left(\tau\left(\mathcal{D}_{gt}\right)\right), \tag{5}$$

where τ is the normalization function, ϕ and ρ are mapping functions for weight control. In experiments, we set $\phi(x) = 10x + 1$. And we test two types $\rho(x) = exp(x)$ and $\rho(x) = x + 1$ in our ablation study. The matrix $\mathcal{D}$ can be splited into two parts. $\phi\left(\left|\tau\left(P_s^d\right) - \tau\left(P_t^d\right)\right|\right)$ facilitates the student to learn the depth estimation from the teacher model. $\rho\left(\tau\left(\mathcal{D}\right)\right)$ is designed for enhancing the weight of the distant object. We want the student to learn the teacher's more precise depth estimation and pay more attention to the knowledge of farther objects. Therefore, we have multiplied them in Eq. 5. In conclusion, the $\mathcal{D}$ adopts depth information to guide prediction level distillation, which contributes to improving the performance of detectors.

In the detection task, samples are usually separated into two groups, which contain positive samples (objects) and negative samples (background). But most previous work usually distills the knowledge of the positive samples to avoid the influence of background noise. However, DeFeat [10] demonstrates that the decoupled distillation is more valid than only distilling the positive samples. It endows the positive and negative samples with different weights in knowledge transfer to make full use of underlying information. Therefore, the depth-guided prediction distillation is defined below,

$$\begin{aligned}\mathcal{L}_p = &\frac{\alpha_{obj}}{K_{obj}} \sum_{i=1}^{K_{obj}} \mathcal{D}_i M_i L_{KL}(P_t^i, P_s^i) \\ &+ \frac{\alpha_{bg}}{K_{bg}} \sum_{i=1}^{K_{bg}} (1 - M_i) L_{KL}(P_t^i, P_s^i),\end{aligned} \tag{6}$$

where K_{obj}, and K_{bg} represent the number of positive sample and negative sample, respectively. α_{obj} and α_{bg} are coefficients to keep the scale balance. $M_i \in \{0, 1\}$ is the binary label of i-th sample with respect to ground-truth object. P_t and P_s are the classification prediction of the teacher and student, respectively. Fig. 5 presents the workflow of $\mathcal{D}$ in the prediction distillation.

3.4. Overall loss

The overall loss is defined as follows,

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_1 * \mathcal{L}_f + \lambda_2 * \mathcal{L}_p, \tag{7}$$

where the $\mathcal{L}_{task}$ is the specific detection task loss such as classification, location loss, and others. λ_1 and λ_2 are distillation weight for the balance of each loss in the same scale.

4. Experiment

4.1. Datasets and Metrics

4.1.1 KITTI

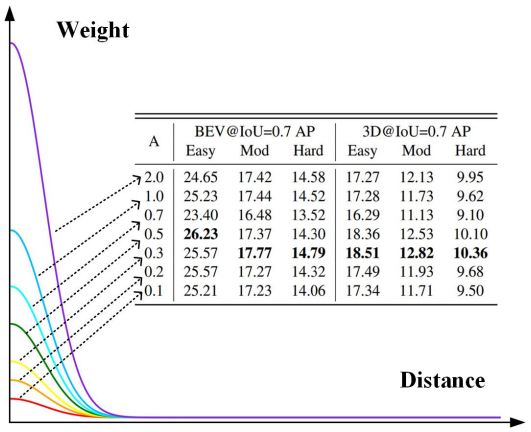
We evaluate our method on commonly used KITTI[9] 3D object detection benchmark. We follow the previous work [6] to split the training images into two groups, 3712 images are selected as $train$ sets, and 3769 images are as val sets. We evaluate the results by 3D detection (3D IoU) and Bird's-Eye-View IoU (BEV IoU) with three levels of difficulty: easy, moderate, and hard. Both metrics are evaluated by $AP|_{R40}$ at the 0.7 IoU threshold. We jointly train the detector with three classes such as car, pedestrian, and cyclist. Considering the difficulty of detecting small objects including pedestrians and cyclists with limited samples, we only report the car detection results.

4.1.2 nuScenes

The nuScenes [3] dataset is a large-scale autonomous driving benchmark, which includes multi-modal data. 1000 driving scenarios are in total, which is officially split into train/val/test sets with 700/150/150 scenes. nuScenes detection score (NDS) and mean average precision (mAP) are the main 3D detection evaluation metric. mAP is calculated by averaging over the distance thresholds of 0.5m, 1m, 2m, and 4m across 10 classes. NDS is the weighted combination of mAP, mATE, mASE, mAOE, mAVE, and mAAE.

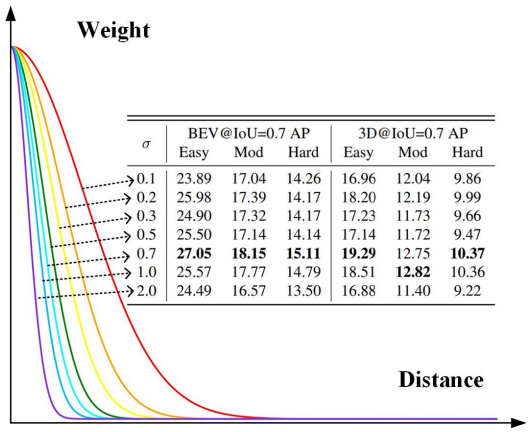
4.2. Implementation Details

We adopt the PGD[40] as the detector baseline for the teacher and student models. For the KITTI dataset, we train the model on a single Nvidia 3090 GPU for 48 epochs with a batch size of 9. The initial learning rate is 0.001 and decayed by 0.1 on the 32nd and 44th epochs, respectively. The weight decay and momentum are set to 0.0001 and 0.9, respectively. SGD is set as the optimizer. All images are resized to the same size of 375 $\times$ 1242 on the $train$ and $test$ stages. For the nuScenes dataset, we train the model on four NVIDIA V100 GPUs for 12 epochs with a batch size of 4. The initial learning rate is 0.004 and decayed by 0.1 on the 8th and 11th epochs, respectively. All images are resized to the same size of 900 $\times$ 1600 on the $train$ and $test$ stages.



A	BEV@IoU=0.7 AP			3D@IoU=0.7 AP		
	Easy	Mod	Hard	Easy	Mod	Hard
2.0	24.65	17.42	14.58	17.27	12.13	9.95
1.0	25.23	17.44	14.52	17.28	11.73	9.62
0.7	23.40	16.48	13.52	16.29	11.13	9.10
0.5	**26.23**	17.37	14.30	18.36	12.53	10.10
0.3	25.57	**17.77**	**14.79**	**18.51**	**12.82**	**10.36**
0.2	25.57	17.27	14.32	17.49	11.93	9.68
0.1	25.21	17.23	14.06	17.34	11.71	9.50

Figure 6: The impact of A. The curve represents the distillation weight attenuation according to the distance to the farthest point.



σ	BEV@IoU=0.7 AP			3D@IoU=0.7 AP		
	Easy	Mod	Hard	Easy	Mod	Hard
0.1	23.89	17.04	14.26	16.96	12.04	9.86
0.2	25.98	17.39	14.17	18.20	12.19	9.99
0.3	24.90	17.32	14.17	17.23	11.73	9.66
0.5	25.50	17.14	14.14	17.14	11.72	9.47
0.7	**27.05**	**18.15**	**15.11**	**19.29**	12.75	**10.37**
1.0	25.57	17.77	14.79	18.51	**12.82**	10.36
2.0	24.49	16.57	13.50	16.88	11.40	9.22

Figure 7: The impact of σ. Note that $\sigma_x = \sigma_y$.

4.3. Ablation Study

In the ablation study, we adopt ResNet101 as the backbone for the teacher model, and ResNet18 is used for the student model. Besides, to create a lightweight student model, we also reduce the channel of detection heads to half, which indeed helps decrease the FLOPs and parameters of the student model.

4.3.1 Impact of perspective-induced feature imitation

In this part, we evaluate the effectiveness of perspective-induced feature imitation. As discussed in Section 3.2, we introduce the perspective matrix to endow the far object with more weight to imitate the feature of the teacher model. The perspective matrix is involved with an important transformation that maps the distance to weight by the 2D-Gauss function. To be specific, A refers to the amplitude which affects weight, and σ_x, σ_y are the variance that influences distribution.

We first fix $\sigma_x = \sigma_y = 1$ and change A to test the amplitude impact. As shown in Fig. 6, although the easy index of BEV IoU gets the maximum when $A = 0.5$, we want to get better results in the more index. Hence, $A = 0.3$ is the optimum parameter we need, which has five better results. Besides, we fixed the $A = 0.3$ and alter σ_x, σ_y. In our ablation study, we set $\sigma_x = \sigma_y$ for simplicity. It can be seen in Fig. 7, $\sigma_x = \sigma_y = 0.7$ has more better results compared to others, which improves **6.59%, 3.97%, 4.02%, 6.73%, 3.98%, 3.69%** performance.

4.3.2 Impact of depth-guided prediction distillation

In this part, we explore the effectiveness of depth-guided prediction distillation. Previous approaches usually use the KL divergence to measure the classification distribution between teacher and student models. And they minimize the KL divergence for reducing the differences in classification prediction. The L2 constraint is also used in classification distillation in recent research. As illustrated in Table 1, we first evaluate the impact of L2 constraint and KL divergence for basic classification distillation. The L2 constraint is more stable for more indexes, so we adopt L2 for subsequent ablation study. Then we add the depth attention module which facilitates the student learning depth estimation from the teacher model. Besides, we set $\phi(x) = 10x + 1$ for a scale balance. And the $\tau(x) = x/max(\mathcal{D})$ is the function for depth normalization. The $max(\mathcal{D})$ refers to the max depth of ground truth in each mini-batch. The results (third row) present depth attention module validly improves the performance of detectors compared to basic classification distillation. Finally, we test the effectiveness of the gt-depth-weighted module. In our analysis, we use the ground truth depth to weigh each object, the farther object will have more weight in distillation. We adopt a monotonically increasing mapping function ρ to achieve this goal. Two types of mapping functions are tested, including exponential function ($\rho(x) = exp(x)$) and linear function ($\rho(x) = x + 1$). Tabel 1 shows the accuracy of the exponential function surpasses the linear function. And combined with the depth attention module, the performance has achieved a new peak (improved by **5.36%, 3.97%, 4.06%, 4.96%, 3.48%, 3.28%**). Table 2 shows the specific accuracy improvement through two modules.

4.3.3 Analysis for differnet T-S pairs distillation

To verify the generalization of the proposed methods, we also conduct experiments with different teacher and student pairs. As shown in Table 3, we adopt ResNet101 [11] and RegNet-3.2GF [32] as the backbone for teacher model. In contrast, ResNet18[11], RegNet-800MF [32], MobileNetV2 [33], and ShuffleNet [47] are used as the student backbone. In some teacher-student pairs, although they have heterogeneous network structures, the distil-

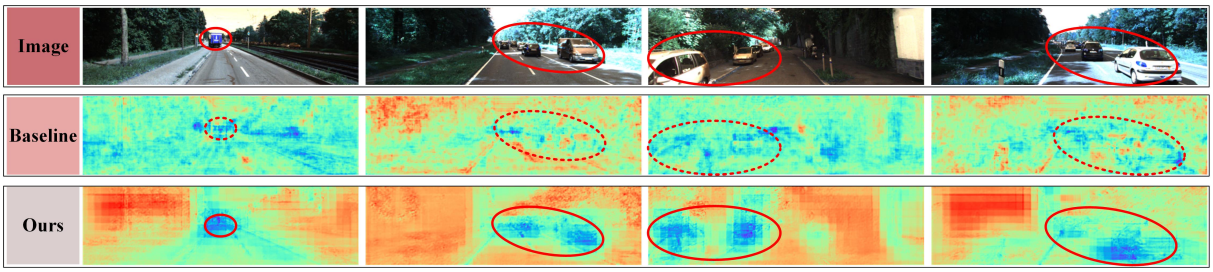


Figure 8: Visualization of FPN features. Our method can extract better features of objects.

L	D	W	BEV@IoU=0.7 AP			3D@IoU=0.7 AP		
			Easy	Mod	Hard	Easy	Mod	Hard
KL	-	-	23.67	16.07	13.25	16.52	11.18	8.92
L2	-	-	24.62	16.92	14.03	15.90	11.26	9.36
L2	$\surd$	-	25.23	17.79	14.42	17.01	11.80	9.60
L2	-	linear	24.49	16.77	13.86	16.12	11.45	9.36
L2	$\surd$	linear	25.56	17.53	14.56	17.14	11.76	9.57
L2	-	exp	25.26	17.66	14.84	17.23	12.22	**10.20**
L2	$\surd$	exp	**25.82**	**18.15**	**15.15**	**17.52**	**12.25**	9.96

Table 1: Ablation study for depth-guided prediction distillation. L refers the loss function which transfesr classification knowledge. D represents depth attention module ($\phi\left(\left|\tau\left(P_s^d\right)-\tau\left(P_t^d\right)\right|\right)$). W denotes the gt-depth-weighted module ($\rho\left(\tau\left(\mathcal{D}\right)\right)$).

F	P	BEV@IoU=0.7 AP			3D@IoU=0.7 AP		
		Easy	Mod	Hard	Easy	Mod	Hard
Teacher		25.48	18.12	15.02	18.59	13.22	10.73
Student		20.46	14.18	11.09	12.56	8.77	6.68
$\surd$	-	27.05	18.15	15.11	19.29	12.75	10.37
-	$\surd$	25.82	18.15	15.15	17.52	12.25	9.96
$\surd$	$\surd$	27.98	18.94	15.87	19.91	13.24	10.91

Table 2: Effectiveness of each module. F refers feature imitation and P represents the prediction distillation.

lation validly helps them promote performance that can not achieve only by themselves. Results demonstrate our method is universal for the distillation framework and stable for accuracy improvement. Moreover, some cases in Table 3 present that the student can obtain better performance than the teacher. We think the student model learns ground truth and extra soft knowledge from the teacher, which may help the student break the limitation of model capacity and have a better convergence.

We also test another detector architecture SMOKE [28] and results are presented in Table 4. Teacher and student adopt the DLA60 [46] and DLA34 [46] as the backbone, respectively. Table 4 proves the stability of our method for different detector architectures.

Method	BEV@IoU=0.7 AP			3D@IoU=0.7 AP		
	Easy	Mod	Hard	Easy	Mod	Hard
T:ResNet101[11]	25.48	18.12	15.02	18.59	13.22	10.73
ResNet18*[11]	20.46	14.18	11.09	12.56	8.77	6.68
+Ours	**27.98**	**18.94**	**15.87**	**19.91**	**13.24**	**10.91**
ResNet18[11]	23.35	16.49	14.21	16.34	11.61	9.72
+Ours	**28.11**	**19.49**	**16.14**	**20.44**	**14.12**	**11.59**
MobileNetV2[33]	15.49	10.99	9.05	9.65	7.09	5.85
+Ours	**17.26**	**12.24**	**10.06**	**10.55**	**7.96**	**6.47**
ShuffleNet[47]	19.78	12.93	10.43	12.20	8.33	6.71
+Ours	**20.66**	**14.37**	**10.90**	**13.77**	**9.01**	**7.20**
T:RegNet-3.2GF[32]	25.61	18.08	14.08	18.04	12.75	10.36
RegNet-800MF*[32]	20.24	14.54	11.70	14.03	10.07	8.06
+Ours	**23.08**	**15.94**	**12.89**	**16.45**	**11.40**	**9.16**
RegNet-800MF[32]	22.27	15.54	12.93	15.08	10.61	6.37
+Ours	**24.71**	**17.06**	**14.06**	**16.83**	**11.48**	**9.30**
MobileNetV2[33]	15.49	10.99	9.05	9.65	7.09	5.85
+Ours	**19.06**	**12.81**	**10.62**	**12.07**	**8.49**	**6.91**
ShuffleNet[47]	19.78	12.93	10.43	12.20	8.33	6.71
+Ours	**21.44**	**13.83**	**10.97**	**13.13**	**8.60**	**6.91**

Table 3: KITTI-3D *val* set evaluation on Car with different teacher-student distillation pairs. $*$ refers to the model that is reduced the channels of detection heads to half.

Method	BEV@IoU=0.7 AP			3D@IoU=0.7 AP		
	Easy	Mod	Hard	Easy	Mod	Hard
T: DLA60	16.49	13.73	13.45	12.45	10.82	10.81
S: DLA34	14.87	10.37	9.04	8.93	6.01	5.45
+Ours	**16.37**	**11.79**	**10.02**	**10.62**	**7.79**	**6.46**

Table 4: Smoke architecture distillation

4.3.4 Qualitative analysis

As illustrated in Fig. 8, we present some visualization results of the FPN output features between the baseline and distillation model. We observe that the distillation model focuses more on objects compared to the baseline rather than the background.

Method	BEV@IoU=0.7 AP			3D@IoU=0.7 AP		
	Easy	Mod	Hard	Easy	Mod	Hard
T:Res101	25.48	18.12	15.02	18.59	13.22	10.73
S:Res18*	20.46	14.18	11.09	12.56	8.77	6.68
+KD[12]	22.95	16.73	13.90	15.16	10.94	8.89
+FitNet[1]	25.72	18.14	15.17	18.32	13.06	10.79
+AT[16]	26.69	17.89	14.85	18.47	12.79	10.32
+FGFI[37]	25.09	17.19	14.34	17.60	11.84	9.74
+TAR[35]	26.84	18.11	15.04	19.37	12.92	10.42
+GID[8]	26.39	17.64	14.21	18.52	12.32	9.98
+DeFeat[10]	25.22	17.11	14.09	17.01	11.87	9.60
+FGD[43]	26.69	**19.08**	**15.99**	18.48	13.19	10.69
+Ours	**27.98**	18.94	15.87	**19.91**	**13.24**	**10.91**
S:Res18	23.25	16.49	14.21	16.34	11.61	9.72
+KD[12]	26.25	17.83	14.60	18.19	12.33	10.10
+FitNet[1]	26.39	19.03	15.87	18.02	13.12	11.06
+AT[16]	27.19	18.88	15.83	19.99	13.79	11.22
+FGFI[37]	26.35	18.00	15.06	18.05	12.33	10.05
+TAR[35]	27.44	19.36	**16.15**	19.83	13.67	11.28
+GID[8]	26.57	18.35	15.22	18.41	12.83	10.38
+DeFeat[10]	25.88	18.16	15.05	17.32	12.23	10.08
+FGD[43]	27.99	19.30	15.96	19.85	13.75	11.35
+Ours	**28.11**	**19.49**	16.14	**20.44**	**14.12**	**11.59**
T:Reg3.2	25.61	18.08	14.08	18.04	12.75	10.36
S:Reg800*	20.24	14.54	11.70	14.03	10.07	8.06
+KD[12]	20.82	14.84	12.32	14.17	10.18	8.31
+FitNet[1]	21.92	15.29	12.09	14.96	10.26	8.61
+AT[16]	22.29	15.57	12.55	15.65	11.02	8.79
+FGFI[37]	22.28	15.54	12.68	15.62	11.09	9.02
+TAR[35]	21.88	15.26	12.22	15.14	10.64	8.40
+GID[8]	21.17	14.46	11.56	14.58	9.89	7.71
+DeFeat[10]	21.58	14.33	11.39	14.43	9.55	7.51
+FGD[43]	21.52	14.95	11.95	14.49	10.05	8.06
+Ours	**23.08**	**15.94**	**12.89**	**16.45**	**11.40**	**9.16**

Table 5: KITTI-3D val set evaluation on Car with SOTA distillation approaches. $*$ refers to the model that is reduced the channels of detection heads to half.

4.3.5 Analysis for parameters and FLOPs of models

We calculate the parameters and FLOPs of student and teacher models. The teacher model has more parameters (403.09MB) and FLOPs (54.71G) than the student (106.64MB, 14.31G). But our proposed distillation framework helps the performance of the student model after distillation beyond the teacher model without the overhead.

4.4. Comparision with State-of-the-art Methods

Comparisons with other state-of-the-art distillation approaches are shown in Table 5 and Table 6. For the KITTI dataset, we conduct experiments on three teacher-student pairs by two network structures which include ResNet [11] and RegNet [32]. As for the nuScenes dataset, we test the ResNet structures, and the results are presented in Table 6. As illustrated in Table 5, our approaches surpass state-of-the-art methods when deploying on diverse network structures (Average improvement for ResNet distillation pairs: 6.19%, 3.88%, 3.59%, 5.73%, 3.49%, 3.05, Average improvement for RegNet distillation pairs: 2.84%, 1.40%, 1.19%, 2.42%, 1.33%, 1.10%). Moreover, our methods also achieve state-of-the-art performance on a large nuScenes dataset and we show the results in Table 6.

Method	mAP	NDS	Method	mAP	NDS
T:Res101	31.7	39.3	**T:Res101**	31.7	39.3
S:Res18	22.9	31.6	**S:Res50**	27.5	35.8
+KD[12]	23.5	32.4	+KD[12]	28.7	37.2
+FitNet[1]	23.7	32.8	+FitNet[1]	28.4	36.9
+AT[16]	23.2	32.1	+AT[16]	28.9	37.8
+FGFI[37]	24.3	33.5	+FGFI[37]	29.3	37.9
+TAR[35]	24.4	33.7	+TAR[35]	29.4	38.1
+GID[8]	24.0	32.9	+GID[8]	29.0	27.3
+DeFeat[10]	24.7	33.9	+DeFeat[10]	29.2	38.2
+FGD[43]	24.9	34.2	+FGD[43]	29.5	38.0
+Ours	**25.1**	**34.6**	+Ours	**30.2**	**38.4**

Table 6: nuScenes val set evaluation with SOTA distillation approaches.

Previous distillation methods based on the visual scheme lack the consideration for spatial cues and they tend to explore better features and response imitation regions. However, depth information is crucial in the 3D task. We introduce two depth modules in KD for learning assistance. On the one hand, we use the depth distillation matrix, which helps farther objects imitate more feature knowledge from the teacher model. On the other hand, we facilitate the student to learn depth estimation knowledge in instance-level prediction with a weight according to distance.

5. Conclusion

In this paper, we consider spatial cues and propose a vision-based distillation framework for the 3D monocular detector. Specifically, we discover that the image contains some implicit depth distributions that are meaningful to guide the distillation. Based on the observation, we propose two modules that facilitate the student learning of the feature and prediction knowledge from the teacher model by spatial cues. Firstly, perspective-induced feature imitation validly enhances the study of features for far objects. Secondly, the depth-guided prediction distillation contributes to the difference reduction of classification distribution and depth estimation between the teacher and the student model. Massive experiments on KITTI and nuScenes demonstrate the effectiveness of our method and the generalization for diverse backbones. It is worth mentioning that the performance of the compact student optimized by distillation is beyond the huge teacher model, which reveals the great potential of knowledge distillation.

References

[1] Romero Adriana, Ballas Nicolas, K Samira Ebrahimi, Chassang Antoine, Gatta Carlo, and B Yoshua. Fitnets: Hints for thin deep nets. *Proc. ICLR*, 2, 2015.

[2] 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.

[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.

[4] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems*, pages 742–751, 2017.

[5] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2021.

[6] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems*, pages 424–432. Citeseer, 2015.

[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. *arXiv preprint arXiv:2201.10830*, 2022.

[8] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2021.

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[10] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2154–2164, 2021.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *stat*, 1050:9, 2015.

[13] Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 87–104. Springer, 2022.

[14] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022.

[15] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.

[16] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *Proceedings of the International Conference on Learning Representations*, 2017.

[17] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019.

[18] Peixuan Li, Huaici Zhao, Pengfei Liu, and Feidao Cao. Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 644–660. Springer, 2020.

[19] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6356–6364, 2017.

[20] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022.

[21] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 1–18. Springer, 2022.

[22] Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, and Gang Wang. Knowledge distillation via the target-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10915–10924, 2022.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[24] Lijie Liu, Jiwen Lu, Chunjing Xu, Qi Tian, and Jie Zhou. Deep fitting degree scoring network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1057–1066, 2019.

[25] Xianpeng Liu, Nan Xue, and Tianfu Wu. Learning auxiliary monocular contexts helps monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1810–1818, 2022.

[26] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d

object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 531–548. Springer, 2022.

[27] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022.

[28] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 996–997, 2020.

[29] 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.

[30] 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.

[31] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017.

[32] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020.

[33] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[34] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1991–1999, 2019.

[35] Ruoyu Sun, Fuhui Tang, Xiaopeng Zhang, Hongkai Xiong, and Qi Tian. Distilling object detectors with task adaptive regularization. *arXiv preprint arXiv:2006.13108*, 2020.

[36] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9627–9636, 2019.

[37] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4933–4942, 2019.

[38] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2021.

[39] Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. *arXiv preprint arXiv:2110.06922*, 2021.

[40] Wang, Tai and Zhu, Xinge and Pang, Jiangmiao and Lin, Dahua. Probabilistic and Geometric Depth: Detecting objects in perspective. In *Conference on Robot Learning (CoRL) 2021*, 2021.

[41] Bin Xu and Zhenzhong Chen. Multi-level fusion based 3d object detection from monocular images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2345–2353, 2018.

[42] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. *arXiv preprint arXiv:2211.10439*, 2022.

[43] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4643–4652, 2022.

[44] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. *arXiv preprint arXiv:2205.01529*, 2022.

[45] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[46] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018.

[47] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018.

[48] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3298, 2021.

[49] 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.

[50] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monocular 3d object detection: An extrinsic parameter free approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7556–7566, 2021.