This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Alleviating Foreground Sparsity for Semi-Supervised Monocular 3D Object Detection

Weijia Zhang[1] Dongnan Liu[1] Chao Ma[2] Weidong Cai[1]
[1]University of Sydney [2]Shanghai Jiao Tong University
{wzha0649, dongnan.liu, tom.cai}@sydney.edu.au chaoma@sjtu.edu.cn

Abstract

Monocular 3D object detection (M3OD) is a significant yet inherently challenging task in autonomous driving due to absence of explicit depth cues in a single RGB image. In this paper, we strive to boost currently underperforming monocular 3D object detectors by leveraging an abundance of unlabelled data via semi-supervised learning. Our proposed ODM3D framework entails cross-modal knowledge distillation at various levels to inject LiDAR-domain knowledge into a monocular detector during training. By identifying foreground sparsity as a main culprit behind existing methods' suboptimal training, we exploit the precise localisation information embedded in LiDAR points to enable more foreground-attentive and efficient distillation via the proposed BEV occupancy guidance mask, leading to notably improved knowledge transfer and M3OD performance. Besides, motivated by insights into why existing cross-modal GT-sampling techniques fail on our task at hand, we further design a novel cross-modal object-wise data augmentation strategy for effective RGB-LiDAR joint learning. Our method ranks 1st in both KITTI validation and test benchmarks, significantly surpassing all existing monocular methods, supervised or semi-supervised, on both BEV and 3D detection metrics. Code will be released at https://github.com/arcaninez/odm3d.

1. Introduction

3D object detection represents a fundamental problem for applications in autonomous driving and robotics. Among 3D object detection from scene representations of different modalities such as LiDAR, RADAR, range images, and stereo images, monocular 3D object detection (M3OD) possesses unique advantages for practical applications. M3OD allows for easy, lightweight, and low-cost deployment on a moving platform since it only requires a single RGB camera. By performing passive sensing, cameras are also free from interference that active sensors such as LiDAR and RADAR are susceptible to, which is essential to safe autonomous driving.

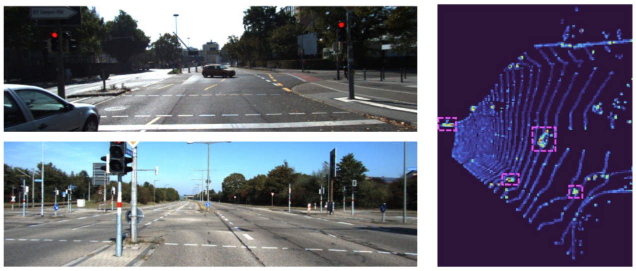

Figure 1. Two types of inefficiency identified in the state-of-the-art CMKD [16]: (a) object sparsity leads to insufficient training signals for the network to learn from; (b) object sparsity leads to foreground (marked with dashed boxes) signals being overwhelmed by background signals in BEV dense distillation.

Despite these benefits, M3OD is arguably the most challenging compared to 3D object detection receiving LiDAR, RADAR, or stereo images as input. On popular 3D object detection benchmarks such as KITTI [13], M3OD methods lag behind their LiDAR-based or stereo counterparts by a daunting margin. This is perhaps not surprising, given that an RGB image does not contain any explicit 3D measurements of a scene. Indeed, inferring 3D attributes from a single 2D image is an ill-posed problem, as pointed out in many previous works [35, 36, 40, 48].

To mitigate this dilemma of 2D-to-3D inference, many existing M3OD approaches resort to incorporating explicit depth estimation for enhanced depth awareness [10, 40, 45, 52], lifting 2D images to 3D "pseudo-LiDAR" representation via off-the-shelf depth estimators [37, 48, 56, 57], or directly utilising matched depth maps in training for RGB-depth feature fusion [18, 58, 64]. Other methods tackle the ill-posed M3OD problem by imposing geometric constraints such as inter-keypoint [4, 23, 29, 33] or inter-object [6, 55] relations to regularise 3D predictions. These have led to consistent and incremental improvements on M3OD.

Parallel to advancements in M3OD, semi-supervised learning (SemiSL) has emerged as a powerful paradigm that enables learning from additional unlabelled data. Motivated by its success, in this paper we advocate exploiting large amounts of unlabelled data to boost M3OD performance. Among preliminary works on semi-supervised M3OD [16, 42, 62], CMKD [16] employs a simple cross-modal knowledge distillation framework to acquire the capability of learning from both images and LiDAR point clouds, labelled and unlabelled, delivering state-of-the-art monocular detection performance.

Despite its impressive results, upon in-depth investigation we made two insightful observations as to how CMKD lacks efficiency in its training as a result of *foreground sparsity* (illustrated in Fig. 1): (i) Autonomous driving scenes often contain too few or even no objects of interest, leading to insufficient training signals for the network (Fig. 1a). In particular, CMKD utilises large amounts of unlabelled samples from the KITTI Raw dataset [12], in which scenes containing no objects at all are common. (ii) In dense distillation, object sparsity results in foreground signals being overwhelmed by the background noise of a much larger area in bird's-eye view (BEV) (Fig. 1b), which undermines accurate feature extractions for cross-modal learning.

To mitigate (i), we design a novel object-wise cross-modal data augmentation technique to paste additional objects into training scenes. GT-sampling-based [61] cross-modal augmentation [26,51,65] has recently been employed by several multi-modal 3D object detectors [5,24,26]. However, these strategies are limited since their augmented scenes are produced via IoU-based collision tests, which fail to consider the relative depth of objects, as discussed in Sec. 4.3. To alleviate this issue, we propose an occlusion-aware cross-modal GT-sampling strategy to augment the training scenes for enhanced RGB-LiDAR distillation.

For (ii), inspired by foreground-attentive distillation in 2D [53, 63, 67] and 3D object detection [9, 14, 68], our proposed distillation method focuses on regions where objects more likely exist rather than treating all locations indifferently. In the absence of ground-truth labels indicating where objects are, we resort to the underlying point location knowledge embedded in point clouds, which serves as an implicit indicator of where objects and foreground might be. Intuitively, locations containing LiDAR points more likely contain an object or part of an object, and vice versa. Hence, we propose to exploit LiDAR point occupancy as a guidance for distillation in BEV.

Our designs effectively alleviate aforementioned issues caused by foreground sparsity, leading to a top-performing M3OD framework based on cross-modal distillation and semi-supervised learning. Besides, our designs are only involved in training and therefore do not introduce any additional computational or memory overhead at inference.

In summary, our contributions include:

1. We propose occupancy-guided cross-modal distillation for M3OD, utilising the underlying localisation information in LiDAR as guidance for foreground-attentive knowledge transfer.
2. We design CMAug, a new and versatile cross-modal augmentation strategy built upon a novel occlusion-aware collision criterion, which suits both supervised and semi-supervised learning settings.
3. Our ODM3D framework achieves 1st place in KITTI *val* and *test* benchmarks, in term of both AP_{3D} and AP_{BEV}, among all published supervised and semi-supervised monocular methods.

2. Related Work

2.1. Monocular 3D Object Detection

Monocular 3D object detection (M3OD) methods can be primarily categorised into image-only, geometric-pior-assisted, and depth-assisted methods. Image-only methods [1,34,38,47,54] directly regress objects' 3D attributes from an RGB image. Prior-assisted methods introduce complex geometric constraints in forms of keypoint [4, 23, 29, 33], inter-object relational [6, 55], camera extrinsic [69] and temporal [2] regularisation. Depth-assisted methods make explicit use of depth to alleviate inherent depth ambiguity in images. Among them, some [37, 44, 48, 56, 57] leverage off-the-shelf depth estimators (*e.g.* DORN [11]) to convert images into a "Pseudo-LiDAR" representation, on which standard LiDAR-based detectors can be applied; some [10, 40, 45, 52] introduce depth estimation as an auxiliary task to learn depth-aware features for accurate 3D inference through RGB-depth fusion or multi-task learning; others [7, 59] exploit depth in the form of disparity maps. Besides, uncertainty modelling is commonly adopted [4, 6, 35, 38, 66] for more accurate and robust estimation of 3D attributes. More recently, transformer [3, 50] has been utilised for more effective contextual and depth-aware feature aggregation [18, 58, 64, 70, 71]. A few methods [9, 58, 72] also utilise external data during training via knowledge distillation, which are detailed in Sec. 2.2.

2.2. Knowledge Distillation for M3OD

A popular technique to transfer knowledge from a stronger model to a weaker one, knowledge distillation [15] has been under-explored in the context of M3OD. Among early efforts, SGM3D [72] distills the knowledge of a teacher trained with stereo images to a monocular CaDDN [45] student; MonoDistill [9] and ADD [58] have their teacher and student based on an identical architecture. MonoDistill's teacher directly takes as input LiDAR-projected depth maps and guides a MonoDLE [38] detector. In contrast, ADD's teacher receives the depth maps as extra input, and is shown to boost multiple monocular detec-

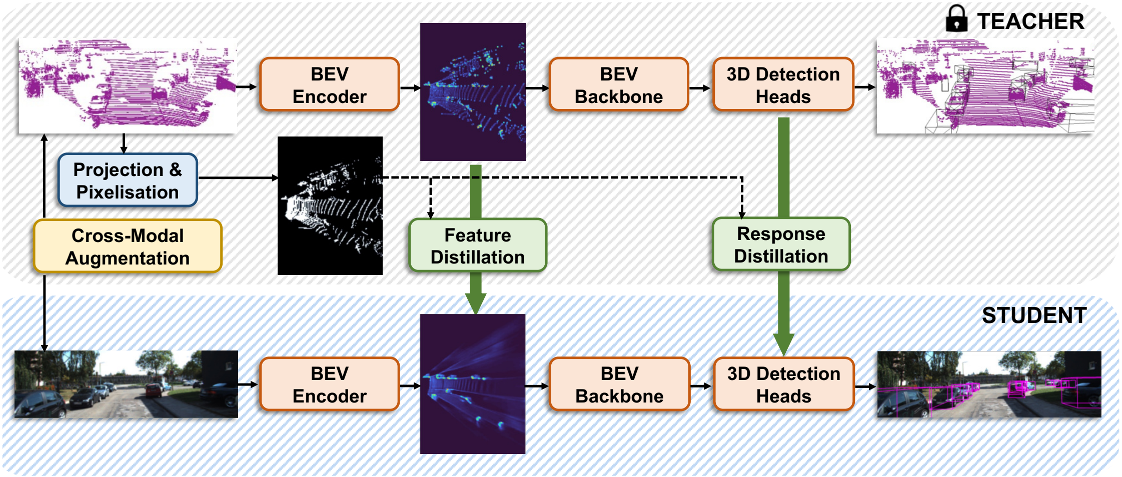

Figure 2. The ODM3D framework. Knowledge distillation is conducted in both feature and prediction spaces in BEV, both guided by a BEV occupancy map derived from ground-truth point clouds. The teacher model is frozen during knowledge distillation, and is discarded at inference. All modules and operations within the grey shaded region are not involved at inference.

tors [36, 45, 64]. Contrary to these approaches, we directly employ a LiDAR-based teacher and distill stronger, more 3D-aware knowledge learnt from raw point clouds.

2.3. Semi-Supervised M3OD

Semi-supervised learning (SemiSL) enables learning from both labelled and unlabelled data. In the context of M3OD, KM3D [22], MVC-MonoDet [31], and Lian *et al.* [30] enforce consistency in terms of object keypoints [22], bounding box predictions [30, 31], or object-level photometry [31] between teacher and student responses. More akin to our method are pseudo-labelling-based approaches [16, 42, 62], which employ a teacher model pre-trained on labelled data to produce predictions (*i.e.* pseudo-labels) for unlabelled data. Instead of directly using the teacher's detection results as in LPCG [42] and Mix-Teaching [62], CMKD [16] lets the student learn the teacher's intermediate features and dense prediction maps via knowledge distillation, achieving state-of-the-art M3OD performance.

2.4. Cross-Modal Data Augmentation

Data augmentation has been a major driver behind the enormous success of deep learning. In LiDAR-based 3D object detection, GT-sampling [61] pastes ground-truth object points into training scenes to diversify and proliferate objects that can be used to train the detector, and is widely adopted by subsequent LiDAR-based detectors [17, 20, 39, 46]. However, extending it to RGB-LiDAR cross-modal learning tasks is less straightforward due to difficulties in maintaining scene-level consistency between augmented RGB and LiDAR data. Recently, several cross-modal augmentation strategies based on GT-sampling have been proposed [8, 26, 51]. They all crop and paste image regions corresponding to pasted object points, and conduct collision tests to avoid severe overlapping in perspective view (PV), with promising results yielded on recent multi-modal 3D object detectors [5, 19, 24, 26]. Yet, to our best knowledge, such strategies have not been explored for cross-modal distillation and semi-supervised learning. In this work, we show that existing strategies lead to augmented scenes extremely challenging if not infeasible for the monocular detector to learn from, and alleviate the issue with our proposed designs.

3. Methodology

3.1. Overall Framework

Our proposed "Occupancy-Guided Distillation for Monocular 3D Object Detection" (ODM3D) framework follows a teacher-student paradigm with cross-modality knowledge distillation, as shown in Fig. 2. The teacher is a pre-trained LiDAR-based 3D object detector which produces intermediate BEV features within its pipeline and performs subsequent 3D object detection in the BEV space. The student is a monocular detector which takes as input a single RGB image and also involves intermediate BEV features. It is trained to mimic the teacher's intermediate BEV features at its BEV encoder and dense prediction maps at its detection heads. In this process, the student acquires LiDAR-induced knowledge from the teacher. Throughout the cross-modality training, a BEV occupancy mask obtained by projecting each scene's point cloud (detailed in Sec. 3.2) is used to guide distillation in both feature and prediction domains (Sec. 3.3). Due to object sparsity in training scenes, we design and apply cross-modal data augmentation, pasting ground-truth objects into each training scene to enrich supervisory signals (Sec. 3.4). At inference, the

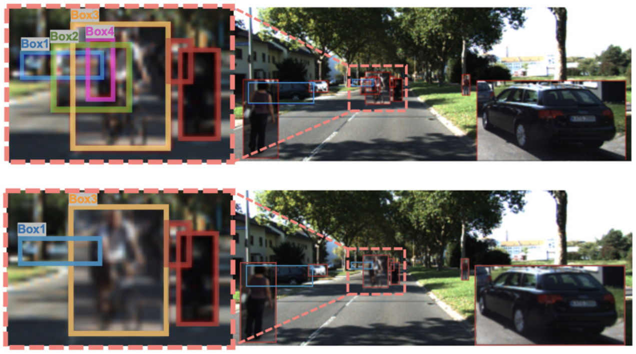

Figure 3. Examples of the same scene augmented with collision tests using IoU and our proposed OAIS thresholds.

LiDAR-based teacher is discarded and only the monocular student is deployed.

3.2. LiDAR-Projected BEV Occupancy Mask

Given raw point cloud $\mathbf{L}$ in a continuous 3D domain, we first project the voxelised $\mathbf{L}$ into BEV to obtain a 3D BEV map $\mathbf{M}' \in \mathbb{R}^{W_{\mathrm{BEV}} \times H_{\mathrm{BEV}} \times D}$ that has the same width W_{BEV} and height H_{BEV} as intermediate BEV features of the teacher and student networks used for feature distillation. Next, we perform "pixelisation" of $\mathbf{M}'$. Specifically, we consider BEV representation $\mathbf{M}'$ an indicator of point occupancy status in the 3D space. Each element in $\mathbf{M}'$ can be regarded as a grid that corresponds to a 3D volume in the voxelised LiDAR space. We let an element in $\mathbf{M}'$ equal to one if its corresponding 3D volume in the LiDAR space contains at least one point, and zero if the 3D volume contains no points. Consequently, a "one" grid represents an active occupancy grid and a "zero" grid represents an empty occupancy grid. Afterwards, we collapse $\mathbf{M}'$ along dimension D to form our 2D BEV occupancy mask $\mathbf{M}_{\mathrm{OCC}} \in \mathbb{R}^{W_{\mathrm{BEV}} \times H_{\mathrm{BEV}}}$. Concretely, an element in $\mathbf{M}_{\mathrm{OCC}}$ is one if there is at least one active grid among all D grids at this location in $\mathbf{M}'$; an element in $\mathbf{M}_{\mathrm{OCC}}$ is zero if all D grids at this location in $\mathbf{M}'$ are empty.

In our experiments, voxelised point cloud $\mathbf{L}$ has a shape of $(W = 1{,}120, H = 1{,}540, D = 40)$, which is determined by our choice of voxelisation resolution and point cloud range, and the intermediate BEV feature has $W_{\mathrm{BEV}} = 140$ and $H_{\mathrm{BEV}} = 188$. Hence, a grid in the proposed occupancy mask corresponds to a total of $8 \times 8 \times 40 = 2{,}560$ voxels, equivalent to a cubic volume of dimension $[0.32m, 0.32m, 4m]$.

3.3. Occupancy-Guided Knowledge Distillation

Occupancy-guided feature distillation. Feature-level distillation is carried out between intermediate BEV features of the teacher and the student, $\mathbf{F}^{\mathrm{Tea}}_{\mathrm{BEV}} \in \mathbb{R}^{W_{\mathrm{BEV}} \times H_{\mathrm{BEV}} \times C_{\mathrm{BEV}}}$ and $\mathbf{F}^{\mathrm{Stu}}_{\mathrm{BEV}} \in \mathbb{R}^{W_{\mathrm{BEV}} \times H_{\mathrm{BEV}} \times C_{\mathrm{BEV}}}$, respectively. The mean square error (MSE) loss is adopted as feature distillation loss $\mathcal{L}_{\mathrm{Feat}}$, on which the proposed 2D BEV occupancy mask $\mathbf{M}_{\mathrm{OCC}}$ is imposed, guiding the distillation to focus on foreground regions while ignoring unimportant and interfering background. In addition, we apply Gaussian smoothing to the BEV occupancy mask which is originally binary. Converting a hard occupancy mask to a soft one effectively mitigates potential misalignment between occupancy maps and feature maps incurred in calibration, projection, or feature extraction. Mathematically, the occupancy-guided feature distillation loss is given by:

$$\mathcal{L}_{\mathrm{FeatKD}} = \|(\mathbf{G}(\sigma) \circledast \mathbf{M}_{\mathrm{OCC}}) \odot \mathcal{L}_{\mathrm{Feat}}(\mathbf{F}^{\mathrm{Stu}}_{\mathrm{BEV}}, \mathbf{F}^{\mathrm{Tea}}_{\mathrm{BEV}})\|^2 \quad (1)$$

where $\mathbf{G}(\sigma)$ is a Gaussian kernel with standard deviation σ determined by the kernel size; $\circledast$ is the convolution operator; $\odot$ denotes the Hadamard product operator; the process of broadcasting $\mathbf{M}_{\mathrm{OCC}}$ to match the channel dimension of $\mathbf{F}^{\mathrm{Tea}}_{\mathrm{BEV}}$ and $\mathbf{F}^{\mathrm{Stu}}_{\mathrm{BEV}}$ is omitted here for brevity.

Occupancy-guided response distillation. Inheriting the spirit of occupancy-guided feature distillation, we further impose occupancy guidance in the response space (*i.e.* dense predictions). This design is made feasible and rational by the fact that predictions of both our teacher and student, along with pre-defined anchors, are made in the BEV space, carrying similar physical connotation to BEV features and BEV occupancy masks. As such, we again apply the BEV occupancy mask on the dense prediction maps generated by both the teacher and student networks.

Nevertheless, considering how anchors and boxes are defined and carried on these dense maps, we argue that it would be more desirable to adopt slackened occupancy guidance rather than stringent, pixel-to-pixel dictation. Essentially, the validity of direct pixel-wise multiplication between a BEV occupancy mask and a BEV feature mask stems from a pixel-to-pixel correspondence between the two representations (even though we have chosen to slacken this correspondence), which however does not hold between a BEV occupancy mask and boxes or anchors defined on prediction maps. It is perfectly normal for the centre of an anchor not to be in the immediate vicinity of points of the ground-truth object to which the anchor is matched. In light of this, we once again opt for Gaussian smoothed occupancy masks, whose benefits will be shown in Sec. 4.3.

Our teacher and student both adopt SSD-style [32] detection heads, comprising a classification head, a localisation head, and a direction head. The occupancy-guided distillation loss for the classification head is given by:

$$\mathcal{L}_{\mathrm{ClsKD}} = \|(\mathbf{G}(\sigma) \circledast \mathbf{M}_{\mathrm{OCC}}) \odot \mathcal{L}_{\mathrm{Cls}}(\mathbf{P}^{\mathrm{Stu}}_{\mathrm{cls}}, \mathbf{P}^{\mathrm{Tea}}_{\mathrm{cls}})\|^2 \quad (2)$$

where $\mathbf{P}^{\mathrm{Stu}}_{\mathrm{cls}}$ and $\mathbf{P}^{\mathrm{Tea}}_{\mathrm{cls}}$ are student's and teacher's classification prediction maps, and other symbols follow Eqn. 1. The localisation and direction distillation losses (*i.e.* $\mathcal{L}_{\mathrm{LocKD}}$ and $\mathcal{L}_{\mathrm{DirKD}}$), based on $\mathcal{L}_{\mathrm{Loc}}$ and $\mathcal{L}_{\mathrm{Dir}}$ respectively, are defined likewise and omitted for brevity. We adopt the quality

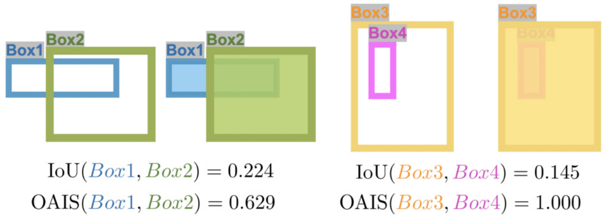
Box1
Box2
Box1
Box2
Box3
Box4
Box3
Box4
IoU(*Box*1, *Box*2) = 0.224
OAIS(*Box*1, *Box*2) = 0.629
IoU(*Box*3, *Box*4) = 0.145
OAIS(*Box*3, *Box*4) = 1.000

Figure 4. A schematic comparison of IoU and the proposed OAIS.

focal loss (QFL) [25], smooth L1 loss, and cross-entropy loss for $\mathcal{L}_{\text{Cls}}$, $\mathcal{L}_{\text{Loc}}$, and $\mathcal{L}_{\text{Dir}}$, respectively. The occupancy-guided response distillation loss $\mathcal{L}_{\text{RespKD}}$ is a weighted sum of distillation losses of all three detection heads. Finally, the overall distillation loss to train the teacher-student framework is a weighted sum of $\mathcal{L}_{\text{FeatKD}}$ and $\mathcal{L}_{\text{RespKD}}$. Note that while a concurrent work [21] also exploits LiDAR-projected masks as guidance, we stress that our occupancy guidance generalises to both feature and response domains and is used in a different setting (*i.e.* cross-modal distillation and semi-supervised learning) and task (*i.e.* M3OD).

3.4. Cross-Modal Data Augmentation

IoU is flawed. Existing cross-modal GT-sampling strategies for 3D object detection [8, 26, 65] universally adopt Intersection of Union (IoU) as a measure of the extent of overlap between pairs of objects in PV to avoid severe occlusion which harms training. However, we argue that IoU is a suboptimal criterion that often fails to indicate severe or even complete occlusion. In Fig. 3, an existing car in Box 1 is largely occluded by the pasted car in Box 2, leaving very limited visual cues for the monocular detector to learn from. The pasted pedestrian in Box 4 is almost entirely occluded by the car in Box 2 and indeed entirely occluded by another pasted pedestrian in Box 3. These severe occlusion cases successfully passed collision tests with an IoU threshold of 0.5. They cause objects having very limited or zero pixels to remain in the augmented scene, and only serve to mislead and harm the monocular detector's learning.

An occlusion-aware criterion. The root of IoU's malfunction lies in its inability to reason about relative depth of boxes. In simple words, IoU measures the extent of overlap of boxes on a 2D plane, but is unaware of which box is being occluded by which in 3D space. We avoid this shortcoming of IoU by introducing a novel Occlusion-Aware Intersection Score (OAIS), which instead calculates the intersection *over the area of the 2D box that has a larger depth value*. Mathematically:

$$\text{OAIS}(B1, B2) = \frac{\text{Area}(B1 \cap B2)}{\text{Area}(\text{Max}_{\text{D}}(B1, B2)))} \tag{3}$$

where $B1$ and $B2$ are 2D boxes projected from 3D boxes with respective depth values; $\text{Max}_{\text{D}}(\cdot)$ is an operator that selects the box with a larger depth value (a random selection if equal depth values). For the case $B1 \cap B2 \neq \varnothing$, this translates into intersection *over area of the box being occluded*. In practice, we take the depth of the ground-truth 3D bounding box from which a 2D box is projected as the 2D box's depth.

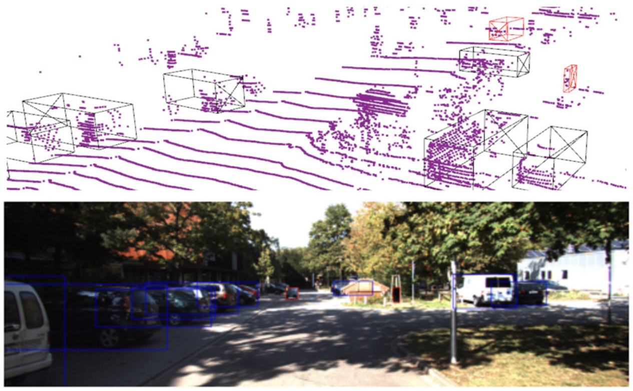

Figure 5. A pasted object (marked with red boxes) can result in an adequately distinguishable cluster of points in LiDAR but a tiny patch in the image.

As shown in Fig. 4, when using IoU as the collision metric, significantly occluded Box 1 has a low IoU of 0.224 with Box 2. Box 4 has an even lower IoU of 0.145 with Box 3 despite being fully occluded by the latter. In comparison, OAIS between Box 1 and Box 2 is 0.629 which provides an intuitive measure of how much of Box 1 has been occluded. Box 3 and Box 4 yield a maximal OAIS of 1.0, implying that "whichever box being occluded has itself 100% occluded". Under a nominal collision threshold of 0.5, these two severe occlusion cases are kept when using IoU, but would have been avoided with the proposed OAIS.

Filtering objects by PV size. We further observed experimentally that excessively tiny patches pasted into images can harm the training of the monocular detector. Traditionally, GT-sampling [61] filters off objects with very few LiDAR points (*e.g.* < 5 points). Yet, we observed that while a faraway object may very well contain a dozen of points, clearly distinguishable in the LiDAR scene, it can occupy a rather limited number of pixels in the image (*e.g.* pasted "Pedestrian" in Fig. 5 takes up 30×13 pixels - 0.083% of the entire PV space) due to perspective, which is further exacerbated when occluded by other pasted objects. Therefore, we design an extra filter to prevent objects excessively small in PV from being pasted into the image. As shown in Sec. 4.3, this simple design leads to a further $0.3 AP_{3D}$ increase on moderate cars.

Pseudo-labels for collision tests. It is noteworthy that collision tests demand knowledge on the location of objects existing in a scene. This has been conveniently obtained from the ground-truth annotations of labelled data in prior works [5, 8, 24, 26, 65]. Our semi-supervised setting, however, involves large amounts of unlabelled scenes. To ac-

Method	Venue	Extra Data	*Test* AP_{3D}@IoU=0.7			*Test* AP_{BEV}@IoU=0.7			*Val* AP_{3D}@IoU=0.7		
			Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
CaDDN [45]	CVPR'21	LiDAR	19.17	13.41	11.46	27.94	18.91	17.19	23.57	16.31	13.84
MonoFlex [66]	CVPR'21	-	19.94	13.89	12.07	28.23	19.75	16.89	23.64	17.51	14.83
GUPNet [35]	ICCV'21	-	20.11	14.20	11.77	30.29	21.19	18.20	22.76	16.46	13.72
MonoDTR [18]	CVPR'22	LiDAR	21.99	15.39	12.73	28.59	20.38	17.14	24.52	18.57	15.51
MonoDistill [9]	ICLR'22	LiDAR	22.97	16.03	13.60	31.87	22.59	19.72	24.31	18.47	15.76
MonoJSG [29]	CVPR'22	-	24.69	16.14	13.64	32.59	21.26	18.18	26.40	18.30	15.40
DID-M3D [43]	ECCV'22	-	24.40	16.29	13.75	32.95	22.76	19.83	22.98	16.12	14.03
DD3D [40]	ICCV'21	Depth	23.22	16.34	14.20	30.98	22.56	20.03	-	-	-
MonoDETR [64]	ICCV'23	-	25.00	16.47	13.58	33.60	22.11	18.60	28.84	20.61	16.38
ADD [58]	AAAI'23	LiDAR	25.61	16.81	13.79	35.20	23.58	20.08	30.71	21.94	18.42
MonoNeRD [60]	ICCV'23	LiDAR	22.75	17.13	15.63	31.13	23.46	20.97	-	-	-
MonoDDE [27]	CVPR'22	-	24.93	17.14	15.10	33.58	23.46	20.37	26.66	19.75	16.72
MonoATT [71]	CVPR'23	-	24.72	17.37	15.00	36.87	24.42	21.88	29.56	22.47	18.65
DD3Dv2 [41]	ICRA'23	LiDAR	26.36	17.61	15.32	35.70	24.67	21.73	-	-	-
MoGDE [70]	NeurIPS'22	-	27.07	17.88	15.66	38.38	25.60	**22.91**	-	-	-
3DSeMo* [28]	arXiv'23	LiDAR	23.55	15.25	13.24	30.99	21.78	18.64	27.35	20.87	17.66
LPCG* [42]	ECCV'22	LiDAR	25.56	17.80	15.38	35.96	24.81	21.86	31.15	23.42	20.60
Mix-Teaching* [62]	CSVT'23	LiDAR	26.89	18.54	15.79	35.74	24.23	20.80	29.74	22.27	19.04
CMKD* [16]	ECCV'22	LiDAR	28.55	18.69	16.77	38.98	25.82	22.80	30.20	21.50	19.40
ODM3D* (Ours)	-	LiDAR	**29.75**	**19.09**	**16.93**	**39.41**	**26.02**	22.76	**35.09**	**23.84**	**20.57**
Improvements	-	-	*+1.20*	*+0.40*	*+0.16*	*+0.43*	*+0.20*	*-0.15*	*+4.89*	*+2.34*	*+1.17*

Table 1. $AP_{3D}|_{R_{40}}$ and $AP_{BEV}|_{R_{40}}$ results of "Car" objects on KITTI *test* and *val* sets. * denotes semi-supervised methods. "*Improvements*" indicates absolute AP improvements compared to a CMKD baseline. Best results within each sub-category are marked in **bold**.

quire the rough location of unannotated objects, we apply the pre-trained teacher for inference on unlabelled scenes and utilise the generated pseudo-labels for collision tests in CMAug. During the teacher's inference, we adopt a lower confidence threshold to discourage false negative detections, since a missed detection may cause intermingled existing and pasted objects in both images and point clouds.

CMAug workflow. Finally, we outline the procedures of our proposed CMAug strategy. Prior to cross-modal distillation, we first build a database of all objects in labelled training samples, similar to [61]. Next, we generate pseudo-labels for all unlabelled training samples and store them as their labels. Afterwards, cross-modal distillation training starts and we randomly select an arbitrary number of objects from the object database to paste into each training scene. Objects whose projected box in PV is less than a pre-defined size are discarded. Remaining object points are pasted into the scene's point cloud, and object patches into the image, using 3D and projected 2D bounding boxes, respectively. Each time a new group of objects is sampled, point cloud collision tests are conducted in BEV using IoU, and image patch collision tests in PV using OAIS, between each pair of existing objects and to-be-pasted objects as well as among all to-be-pasted objects. Objects that fail any collision tests will be discarded. Eventually, all kept objects are pasted into the scene in a far-to-near order. In Sec. 4.3, we show that CMAug also generalises beyond cross-modal distillation and M3OD.

4. Experiments

4.1. Implementation Details

Datasets. We validate our framework on the KITTI 3D [13] dataset, which consists of 7,481 training and 7,518 test images with corresponding point clouds. For local evaluation, we follow the convention to divide training images into a training subset of 3,712 images and a validation set of 3,769 images, dubbed KITTI *train* and *val*, respectively. The best model determined by KITTI *val* is evaluated on the test set, denoted as KITTI *test*. Objects in KITTI 3D are annotated into three difficulty levels: "Easy", "Moderate", and "Hard", with Average Precision (AP) as the official evaluation metric. KITTI 3D is an annotated subset of the KITTI Raw dataset [12], which further comprises around 42k unannotated images and corresponding point clouds in sequence form, which are exploited under our semi-supervised learning framework. For validation, we follow [16, 30] and use the `eigen-clean` subset [48] of KITTI Raw to avoid data leakage due to scenes overlapping with KITTI *val*.

Network details. We choose the state-of-the-art CMKD [16] as our baseline and implement our framework based

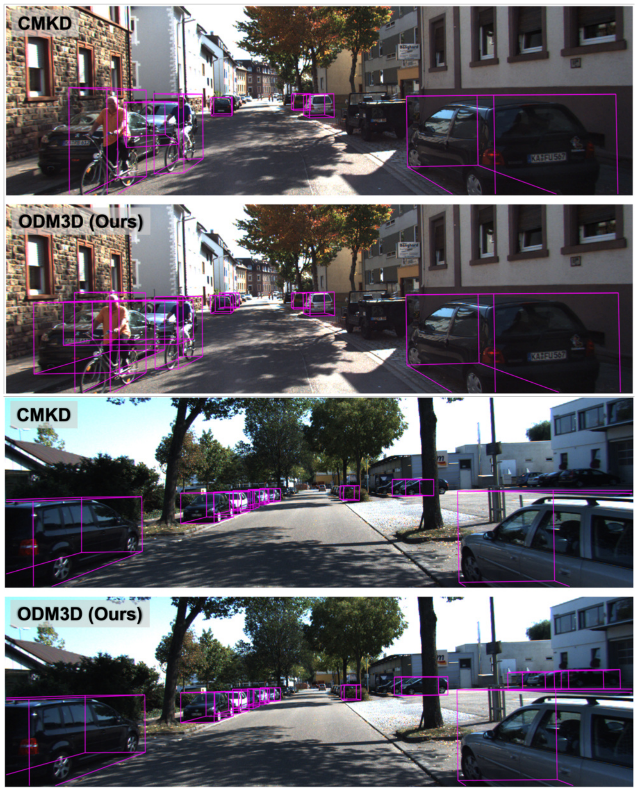

Figure 6. Qualitative comparison of detection results by our method and CMKD [16].

on the OpenPCDet [49] codebase. We use SECOND [61] as our LiDAR-based teacher and CaDDN [45] our monocular student. Our CaDDN student follows the same settings as in [45] and [16], except that depth maps are not utilised for supervising categorical depth estimation since our training scenes have been altered by CMAug. Instead, depth estimation is supervised implicitly by dense distillation.

Training details. Our framework is trained on a single NVIDIA RTX 3090 GPU with a batch size of 4. We follow a two-stage distillation strategy and train the framework for 30 epochs in stage 1 ($\mathcal{L}_{\text{FeatKD}}$ only) and 15 epochs in stage 2 ($\mathcal{L}_{\text{FeatKD}}$ and $\mathcal{L}_{\text{RespKD}}$), using both labelled and unlabelled data. The proposed CMAug is applied in both stages. More experimental details are provided in the supplementary material.

4.2. Comparisons with Prior Arts

Quantitative results. We make a detailed quantitative comparison of our method against recently published supervised and semi-supervised monocular 3D object detectors on both KITTI *test* and *val* sets. We report the results on the "Car" category since it is the most important category in the KITTI 3D dataset ("Pedestrian" and "Cyclist" results are provided in the supplementary material). As shown in Tab. 1, our method drastically boosts a CaDDN [45] model from AP_{3D} 13.41 to 19.09 (a 42.4% increase) on moderate cars owing to effective usage of extra LiDAR and unlabelled data. Our method surpasses all existing methods, supervised or semi-supervised, by considerable margins, including the current state-of-the-art CMKD [16]. Specifically, on KITTI *test*, ODM3D outperforms CMKD by 0.40 and 0.20 on AP_{3D} and AP_{BEV}, respectively. Larger performance gains are observed on KITTI *val*, where ODM3D is ahead of CMKD by 2.34 AP_{3D} on moderate cars and outperforms all other methods by at least 0.4 AP_{3D}.

Expt.	FD	RD	O-FD	O-RD	CMA	*Val* AP_{3D}@IoU=0.7		
						Easy	Mod.	Hard
1	✓	✓				32.67	21.54	18.79
2	✓	✓			✓	31.50	22.18	19.34
3	✓			✓		33.02	21.89	18.19
4		✓	✓			34.69	23.68	20.55
5			✓	✓		34.84	23.77	20.04
6			✓	✓	✓	**35.09**	**23.84**	**20.57**

Table 2. Ablation experiments on core components of our method. "FD" and "RD" stand for vanilla feature and response distillation, respectively. "O-" denotes their occupancy-guided variants. "CMA" denotes the proposed CMAug.

Expt.	OFD-V	ORD-V	OFD-G	ORD-G	*Val* AP_{3D}@IoU=0.7		
					Easy	Mod.	Hard
1	✓	✓			34.11	23.36	19.50
2	✓			✓	34.58	23.72	19.83
3		✓	✓		34.63	23.59	19.85
4			✓	✓	**34.84**	**23.77**	**20.04**

Table 3. Ablation experiments on the use of Gaussian smoothing in occupancy-guided feature and response distillation. "OFD-V" and "ORD-V" denote vanilla occupancy-guided feature and response distillation, respectively, whereas "-G" indicates their variants using Gaussian-smoothed occupancy masks.

Qualitative results. Fig. 6 visualises detections by our method and CMKD. In the first scene, it is clear that ODM3D precisely detects the two cars on the left and another car further down the street which are missed by CMKD. In the second scene, ODM3D picks up two cars on the right and one occluded by a tree with higher accuracy. These examples show that our method better handles faraway and occluded objects, which is likely owing to our occlusion-aware augmentation and foreground-attentive occupancy-guided distillation.

4.3. Ablation Studies

Effectiveness of core designs. Tab. 2 studies the effectiveness of each core component of our framework. It is clear that the proposed occupancy-guided feature distillation and occupancy-guided response distillation lead to increased detection results both individually (Expt. 1→3, 4) and collectively (Expt. 1→5) compared to baseline re-

Expt.	CD	LD	DD	CD-G	LD-G	DD-G	*Val* AP_{3D}@IoU=0.7		
							Easy	Mod.	Hard
1	✓	✓	✓				34.63	23.59	19.85
2		✓	✓	✓			34.66	23.71	19.98
3	✓		✓		✓		34.50	23.75	19.81
4			✓	✓	✓		**34.97**	23.75	20.03
5				✓	✓	✓	34.84	**23.77**	**20.04**

Table 4. Ablation experiments on the use of Gaussian smoothing for various heads in occupancy-guided response distillation. "CD", "LD" and "DD" denote vanilla classification, localisation and direction distillation, respectively, whereas "-G" denotes their variants using Gaussian-smoothed occupancy masks.

sults (Expt. 1). It can be inferred from Expt. 3 and 4 that improvements brought about by the use of BEV occupancy guidance (Expt. 1→5) are primarily accounted for by occupancy-guided feature distillation. This also suggests that high-quality BEV feature maps can be a prerequisite for accurate and effective response distillation and detection that take place downstream. The proposed CMAug leads to improved results on both CMKD [16] (Expt. 1→2) and occupancy-guided distillation (Expt. 5→6) baselines, highlighting its effectiveness and potential as a versatile, plug-and-play gadget for boosting joint LiDAR-RGB learning. Finally, the highest detection results are achieved using occupancy-guided feature distillation, occupancy-guided response distillation, and CMAug altogether (Expt. 6).

Effectiveness of occupancy-guided distillation designs. The benefits of smoothed BEV occupancy masks are illustrated in ablation experiments in Tab. 3. Performance drops take place with Gaussian smoothing ablated in either feature (Expt. 4→2) or response (Expt. 4→3) distillation, and worst performance is observed when it is absent in both distillations (Expt. 4→1). We further study the effect of Gaussian smoothing in distilling each detection head in Tab. 4. The results show that smoothed occupancy masks result in improved detection when applied either individually to each detection head (Expt. 1→2,3) or to multiple heads in combinatorial ways (Expt. 1,2,3→4,5).

Effectiveness of CMAug designs. From Tab. 5, naively applying the commonly adopted MixedAug [26], which performs simple IoU-based collision tests, leads to significantly degraded detection performance. Replacing the IoU score with the proposed OAIS immediately gives rise to drastic increases in performance (a notable 21.2% increase on moderate cars), turning data augmentation's contribution from negative to positive and surpassing the baseline. Besides, filtering by PV sizes also boosts detection performance on moderate and hard cars.

Generalisation studies of CMAug. We further apply our CMAug to representative and high-performance multi-modal 3D object detectors VFF [26], FocalsConv [5] and LoGoNet [24]. As shown in Tab. 6, by simply replacing their default augmentation, the proposed CMAug consistently boosts the detection accuracy of these multi-modal detectors, with all other settings unchanged. These results corroborate our previous arguments that the identified "IoU defect" is universal among 3D object detectors employing cross-modality GT-sampling, and our proposed fix to it generalises beyond monocular detectors.

Method	*Val* AP_{3D}@IoU=0.7		
	Easy	Mod.	Hard
Baseline	**32.67**	21.54	18.79
+ MixedAug [26]	27.77	18.58	16.44
Improvements	*-4.90*	*-2.96*	*-2.35*
+ OAIS	32.42	21.87	18.99
+ MinPxFilter	32.57	**22.21**	**19.43**
Improvements	*-0.10*	*+0.76*	*+0.64*

Table 5. Ablation experiments on the components of CMAug and performance comparison between CMAug and MixedAug [26].

Method	*Val* AP_{3D}@IoU=0.7		
	Easy	Mod.	Hard
VFF–Voxel-RCNN [26]	**92.80**	83.53	82.78
+ CMAug	92.64	**83.68**	**82.97**
Improvements	*-0.16*	*+0.15*	*+0.19*
FocalsConv–PV-RCNN [5]	91.89	85.31	83.10
+ CMAug	**92.55**	**85.53**	**83.33**
Improvements	*+0.66*	*+0.22*	*+0.23*
LoGoNet [24]	91.92	85.02	82.88
+ CMAug	**92.16**	**85.20**	**83.02**
Improvements	*+0.24*	*+0.18*	*+0.14*

Table 6. A generalisation study on CMAug applied to multi-modal 3D object detectors. Baseline results are produced from official code for a fair assessment of CMAug's effectiveness.

5. Conclusion

In this paper, we proposed ODM3D, a novel knowledge distillation framework that alleviates the foreground sparsity issue in autonomous driving scenes for enhanced semi-supervised monocular 3D object detection. We showed that exploiting the inherent ground-truth 3D occupancy knowledge in point clouds significantly benefits knowledge distillation in both feature and prediction spaces, as the network is encouraged to attend to regions that more likely contain objects. We also demonstrated that our proposed cross-modal data augmentation strategy not only enriches supervisory signals throughout the cross-modality learning process, but also generates more realistic and learner-friendly augmented scenes. Extensive experiments on the KITTI dataset have validated the effectiveness of the proposed method.

References

[1] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *ICCV*, 2019. 2

[2] Garrick Brazil, Gerard Pons-Moll, Xiaoming Liu, and Bernt Schiele. Kinematic 3d object detection in monocular video. In *ECCV*, 2020. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[4] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. In *CVPR*, 2021. 1, 2

[5] Yukang Chen, Yanwei Li, Xiangyu Zhang, Jian Sun, and Jiaya Jia. Focal sparse convolutional networks for 3d object detection. In *CVPR*, 2022. 2, 3, 5, 8

[6] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. Monopair: Monocular 3d object detection using pairwise spatial relationships. In *CVPR*, 2020. 1, 2

[7] Yi-Nan Chen, Hang Dai, and Yong Ding. Pseudo-stereo for monocular 3d object detection in autonomous driving. In *CVPR*, 2022. 2

[8] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3d object detection. In *ECCV*, 2022. 3, 5

[9] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, and Wanli Ouyang. Monodistill: Learning spatial features for monocular 3d object detection. In *ICLR*, 2022. 2, 6

[10] Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning depth-guided convolutions for monocular 3d object detection. In *CVPR*, 2020. 1, 2

[11] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 2

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 32(11):1231–1237, 2013. 2, 6

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1, 6

[14] Xiaoyang Guo, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector. In *ICCV*, 2021. 2

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Workshops*, 2014. 2

[16] Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *ECCV*, 2022. 1, 2, 3, 6, 7, 8

[17] Jordan S. K. Hu, Tianshu Kuai, and Steven L. Waslander. Point density-aware voxels for lidar 3d object detection. In *CVPR*, 2022. 3

[18] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H. Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *CVPR*, 2022. 1, 2, 6

[19] Peixiang Huang, Li Liu, Renrui Zhang, Song Zhang, Xinli Xu, Baichao Wang, and Guoyi Liu. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *CVPR*, 2022. 3

[20] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019. 3

[21] Jianing Li, Ming Lu, Jiaming Liu, Yandong Guo, Li Du, and Shanghang Zhang. Bev-lgkd: A unified lidar-guided knowledge distillation framework for bev 3d object detection. In *arXiv preprint arXiv:2212.00623*, 2022. 5

[22] Peixuan Li. Monocular 3d detection with geometric constraints embedding and semi-supervised training. In *arXiv preprint arXiv:2009.00764*, 2020. 3

[23] Peixuan Li, Huaici Zhao, Pengfei Liu, and Feidao Cao. Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving. In *ECCV*, 2020. 1, 2

[24] Xin Li, Tao Ma, Yuenan Hou, Botian Shi, Yuchen Yang, Youquan Liu, Xingjiao Wu, Qin Chen, Yikang Li, Yu Qiao, and Liang He. Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion. In *CVPR*, 2023. 2, 3, 5, 8

[25] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *NeurIPS*, 2020. 5

[26] Yanwei Li, Xiaojuan Qi, Yukang Chen, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Voxel field fusion for 3d object detection. In *CVPR*, 2022. 2, 3, 5, 8

[27] Zhuoling Li, Zhan Qu, Yang Zhou, Jianzhuang Liu, Haoqian Wang, and Lihui Jiang. Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection. In *CVPR*, 2022. 6

[28] Zhenyu Li, Zhipeng Zhang, Heng Fan, Yuan He, Ke Wang, Xianming Liu, and Junjun Jiang. Augment and criticize: Exploring informative samples for semi-supervised monocular 3d object detection. In *arXiv preprint arXiv:2303.11243*, 2023. 6

[29] Qing Lian, Peiliang Li, and Xiaozhi Chen. Monojsg: Joint semantic and geometric cost volume for monocular 3d object detection. In *CVPR*, 2022. 1, 2, 6

[30] Qing Lian, Botao Ye, Ruijia Xu, Weilong Yao, and Tong Zhang. Exploring geometric consistency for monocular 3d object detection. In *CVPR*, 2022. 3, 6

[31] Qing Liang, Yanbo Xu, Weilong Yao, Yingcong Chen, and Tong Zhang. Semi-supervised monocular 3d object detection by multi-view consistency. In *ECCV*, 2022. 3

[32] Wei Liu, Dragomir Anguelov, Mumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 4

[33] Xianpeng Liu, Nan Xue, and Tianfu Wu. Learning auxiliary monocular contexts helps monocular 3d object detection. In *AAAI*, 2022. 1, 2

[34] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *CVPR Workshops*, 2020. 2

[35] 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 *ICCV*, 2021. 1, 2, 6

[36] Xinzhu Ma, Shinan Liu, Zhiyi Xia, Hongwen Zhang, Xingyu Zeng, and Wanli Ouyang. Rethinking pseudo-lidar representation. In *ECCV*, 2020. 1, 3

[37] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Xin Fan, and Wanli Ouyang. Accurate monocular object detection via color-embedded 3d reconstruction for autonomous driving. In *ICCV*, 2019. 1, 2

[38] Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, and Wanli Ouyang. Delving into localization errors for monocular 3d object detection. In *CVPR*, 2021. 2

[39] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3d object detection. In *ICCV*, 2021. 3

[40] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *ICCV*, 2021. 1, 2, 6

[41] Dennis Park, Jie Li, Dian Chen, Vitor Guizilini, and Adrien Gaidon. Depth is all you need for monocular 3d detection. In *ICRA*, 2023. 6

[42] Liang Peng, Fei Liu, Zhengxu Yu, Senbo Yan, Dan Deng, Zheng Yang, Haifeng Liu, and Deng Cai. Lidar point cloud guided monocular 3d object detection. In *ECCV*, 2022. 2, 3, 6

[43] Liang Peng, Xiaopei Wu, Zheng Yang, Haifeng Liu, and Deng Cai. Did-m3d: Decoupling instance depth for monocular 3d object detection. In *ECCV*, 2022. 6

[44] Rui Qian, Divyansh Garg, Yan Wang, Yurong You, Serge Belongie, Bharath Hariharan, Mark Campbell, Kilian Q. Weinberger, and Wei-Lun Chao. End-to-end pseudo-lidar for image-based 3d object detection. In *CVPR*, 2020. 2

[45] Cody Reading, Ali Harakeh, Julia Chae, , and Steven L. Waslander. Categorical depth distribution network for monocular 3d object detection. In *CVPR*, 2021. 1, 2, 3, 6, 7

[46] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *CVPR*, 2020. 3

[47] Andrea Simonelli, Samuel Rota Bulò, Lorenzo Porzi, Manuel López Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *ICCV*, 2019. 2

[48] Andrea Simonelli, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, and Elisa Ricci. Are we missing confidence in pseudo-lidar methods for monocular 3d object detection? In *ICCV*, 2021. 1, 2, 6

[49] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds, 2020. https://github.com/open-mmlab/OpenPCDet. 7

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[51] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *CVPR*, 2021. 2, 3

[52] Li Wang, Liang Du, Xiaoqing Ye, Yanwei Fu, Guodong Guo, Xiangyang Xue, Jianfeng Feng, and Li Zhang. Depth-conditioned dynamic message propagation for monocular 3d object detection. In *CVPR*, 2021. 1, 2

[53] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *CVPR*, 2019. 2

[54] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *ICCV Workshops*, 2021. 2

[55] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *CoRL*, 2021. 1, 2

[56] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, 2019. 1, 2

[57] Xinshuo Weng and Kris Kitani. Monocular 3d object detection with pseudo-lidar point cloud. In *ICCV Workshops*, 2019. 1, 2

[58] Zizhang Wu, Yunzhe Wu, Jian Pu, Xianzhi Li, and Xiaoquan Wang. Attention-based depth distillation with 3d-aware positional encoding for monocular 3d object detection. In *AAAI*, 2023. 1, 2, 6

[59] Bin Xu and Zhenzhong Cheng. Multi-level fusion based 3d object detection from monocular images. In *CVPR*, 2018. 2

[60] Junkai Xu, Liang Peng, Haoran Cheng, Hao Li, Wei Qian, Ke Li, Wenxiao Wang, and Deng Cai. Mononerd: Nerf-like representations for monocular 3d object detection. In *ICCV*, 2023. 6

[61] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 2, 3, 5, 6, 7

[62] Lei Yang, Xinyu Zhang, Jun Li, Li Wang, Minghan Zhu, Chuan-Fang Zhang, and Huaping Liu. Mix-teaching: A simple, unified and effective semi-supervised learning framework for monocular 3d object detection. *IEEE TCSVT*, 33(11):6832–6844, 2023. 2, 3, 6

[63] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *CVPR*, 2022. 2

[64] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, , Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In *ICCV*, 2023. 1, 2, 3, 6

[65] Wenwei Zhang, Zhe Wang, and Chen Change Loy. Exploring data augmentation for multi-modality 3d object detection. In *arXiv preprint arXiv:2012.12741*, 2020. 2, 5

[66] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *CVPR*, 2021. 2, 6

[67] Zhaohui Zheng, Rongguang Ye, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, and Ming-Ming Cheng. Localization distillation for dense object detection. In *CVPR*, 2022. 2

[68] Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, and Chao Ma. Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view. In *CVPR*, 2023. 2

[69] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monocular 3d object detection: An extrinsic parameter free approach. In *CVPR*, 2021. 2

[70] Yunsong Zhou, Quan Liu, Hongzi Zhu, Yunzhe Li, Shan Chang, and Minyi Guo. Mogde: Boosting mobile monocular 3d object detection with ground depth estimation. In *NeurIPS*, 2022. 2, 6

[71] Yunsong Zhou, Hongzi Zhu, Quan Liu, Shan Chang, and Minyi Guo. Monoatt: Online monocular 3d object detection with adaptive token transformer. In *CVPR*, 2023. 2, 6

[72] Zheyuan Zhou, Liang Du, Xiaoqing Ye, Zhikang Zou, Xiao Tan, Li Zhang, Xiangyang Xue, and Jianfeng Feng. Sgm3d: Stereo guided monocular 3d object detection. *IEEE RA-L*, 7(4):10478–10485, 2022. 2