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

Weijia Zhang\textsuperscript{1}  Dongnan Liu\textsuperscript{1}  Chao Ma\textsuperscript{2}  Weidong Cai\textsuperscript{1}
\textsuperscript{1}University of Sydney  \textsuperscript{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 \textsuperscript{1}st in both KITTI validation and test benchmarks, significantly outperforming all existing monocular methods, supervised or semi-supervised, on both BEV and 3D detection metrics. Code will be released at \url{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.

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 \cite{Geiger2013IJRR}, 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 \cite{Gallagher2016CVPR, Girshick2017CVPR, Ren2017CVPR, Sun2017CVPR, Zhang2018CVPR}.

To mitigate this dilemma of 2D-to-3D inference, many existing M3OD approaches resort to incorporating explicit depth estimation for enhanced depth awareness \cite{Geiger2013IJRR, Zhang2018CVPR, Sun2017CVPR, Zhang2017CVPR}, lifting 2D images to 3D “pseudo-LiDAR” representation via off-the-shelf depth estimators \cite{Garg2016CVPR, Sun2017CVPR, Zhang2018CVPR}, or directly utilising matched depth maps in training for RGB-depth feature fusion \cite{Gallagher2016CVPR, Regneri2015CVPR, Riegler2015CVPR}. Other methods tackle the ill-posed M3OD problem by imposing geometric constraints such as inter-keypoint \cite{Gallagher2016CVPR, Regneri2015CVPR, Liu2017CVPR, Chen2018CVPR} or inter-object \cite{Chen2018CVPR, Zhang2018CVPR} 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 differently. 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-prior-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-
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.

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

3.2. LiDAR-Projected BEV Occupancy Mask

Given raw point cloud $L$ in a continuous 3D domain, we first project the voxelised $L$ into BEV to obtain a 3D BEV map $M' \in \mathbb{R}^{W_{BEV} \times H_{BEV} \times D}$ 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 $M'$. Specifically, we consider BEV representation $M'$ an indicator of point occupancy status in the 3D space. Each element in $M'$ can be regarded as a grid that corresponds to a 3D volume in the voxelised LiDAR space. We let an element in $M'$ 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 $M'$ along dimension $D$ to form our 2D BEV occupancy mask $M_{OCC} \in \mathbb{R}^{W_{BEV} \times H_{BEV}}$. Concretely, an element in $M_{OCC}$ is one if there is at least one active grid among all $D$ grids at this location in $M'$; an element in $M_{OCC}$ is zero if all $D$ grids at this location in $M'$ are empty.

In our experiments, voxelised point cloud $L$ has a shape of $(W = 1,210, H = 1,540, D = 40)$, which is determined by our choice of voxelisation resolution and point cloud range, and the intermediate BEV feature has $W_{BEV} = 140$ and $H_{BEV} = 188$. 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, $F_{BEV}^{Tea} \in \mathbb{R}^{W_{BEV} \times H_{BEV} \times C_{BEV}}$ and $F_{BEV}^{Stu} \in \mathbb{R}^{W_{BEV} \times H_{BEV} \times C_{BEV}}$, respectively. The mean square error (MSE) loss is adopted as feature distillation loss $\mathcal{L}_{Feat}$, on which the proposed 2D BEV occupancy mask $M_{OCC}$ 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:

$$
\mathcal{L}_{Feat\text{KD}} = \| (G(\sigma) \circ M_{OCC}) \circ \mathcal{L}_{Feat}(F_{BEV}^{Stu}, F_{BEV}^{Tea}) \|^2
$$

where $G(\sigma)$ is a Gaussian kernel with standard deviation $\sigma$ determined by the kernel size; $\circ$ is the convolution operator; $\circ$ denotes the Hadamard product operator; $\circ$ denotes the Hadamard product operator; $\circ$ denotes the Hadamard product operator; $\circ$ denotes the Hadamard product operator. In our experiments, voxelised point cloud $L$ has a shape of $(W = 1,210, H = 1,540, D = 40)$, which is determined by our choice of voxelisation resolution and point cloud range, and the intermediate BEV feature has $W_{BEV} = 140$ and $H_{BEV} = 188$. 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, $F_{BEV}^{Tea} \in \mathbb{R}^{W_{BEV} \times H_{BEV} \times C_{BEV}}$ and $F_{BEV}^{Stu} \in \mathbb{R}^{W_{BEV} \times H_{BEV} \times C_{BEV}}$, respectively. The mean square error (MSE) loss is adopted as feature distillation loss $\mathcal{L}_{Feat}$, on which the proposed 2D BEV occupancy mask $M_{OCC}$ 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:

$$
\mathcal{L}_{Feat\text{KD}} = \| (G(\sigma) \circ M_{OCC}) \circ \mathcal{L}_{Feat}(F_{BEV}^{Stu}, F_{BEV}^{Tea}) \|^2
$$

where $G(\sigma)$ is a Gaussian kernel with standard deviation $\sigma$ determined by the kernel size; $\circ$ is the convolution operator; $\circ$ denotes the Hadamard product operator; $\circ$ denotes the Hadamard product operator; $\circ$ denotes the Hadamard product operator. In our experiments, voxelised point cloud $L$ has a shape of $(W = 1,210, H = 1,540, D = 40)$, which is determined by our choice of voxelisation resolution and point cloud range, and the intermediate BEV feature has $W_{BEV} = 140$ and $H_{BEV} = 188$. 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, $F_{BEV}^{Tea} \in \mathbb{R}^{W_{BEV} \times H_{BEV} \times C_{BEV}}$ and $F_{BEV}^{Stu} \in \mathbb{R}^{W_{BEV} \times H_{BEV} \times C_{BEV}}$, respectively. The mean square error (MSE) loss is adopted as feature distillation loss $\mathcal{L}_{Feat}$, on which the proposed 2D BEV occupancy mask $M_{OCC}$ 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:

$$
\mathcal{L}_{Feat\text{KD}} = \| (G(\sigma) \circ M_{OCC}) \circ \mathcal{L}_{Feat}(F_{BEV}^{Stu}, F_{BEV}^{Tea}) \|^2
$$

where $G(\sigma)$ is a Gaussian kernel with standard deviation $\sigma$ determined by the kernel size; $\circ$ is the convolution operator; $\circ$ denotes the Hadamard product operator; $\circ$ denotes the Hadamard product operator; $\circ$ denotes the Hadamard product operator. In our experiments, voxelised point cloud $L$ has a shape of $(W = 1,210, H = 1,540, D = 40)$, which is determined by our choice of voxelisation resolution and point cloud range, and the intermediate BEV feature has $W_{BEV} = 140$ and $H_{BEV} = 188$. 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 $[0.32m, 0.32m, 4m]$.
focal loss (QFL) [25], smooth L1 loss, and cross-entropy loss for $L_{\text{Cls}}$, $L_{\text{Loc}}$, and $L_{\text{Dir}}$, respectively. The occupancy-guided response distillation loss $L_{\text{RespKD}}$ 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 $L_{\text{FeatKD}}$ and $L_{\text{RespKD}}$. 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:

$$\text{OAIS}(B_1, B_2) = \frac{\text{Area}(B_1 \cap B_2)}{\text{Area}(\text{Max}_D(B_1, B_2))}$$

where $B_1$ and $B_2$ are 2D boxes projected from 3D boxes with respective depth values; $\text{Max}_D(\cdot)$ is an operator that selects the box with a larger depth value (a random selection if equal depth values). For the case $B_1 \cap B_2 \neq \emptyset$, 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.

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 0.3AP$_{3D}$ 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-
are pasted into the scene in a far-to-near order. In Sec.

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 [51]. 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...
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 ($L_{\text{FeatKD}}$ only) and 15 epochs in stage 2 ($L_{\text{FeatKD}}$ and $L_{\text{RespKD}}$), 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 $AP_{3D}$.

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-
results (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.

**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


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


