General Geometry-aware Weakly Supervised 3D Object Detection

Guowen Zhang[1,2], Junsong Fan[2], Liyi Chen[1,2], Zhaoxiang Zhang[2,3,4], Zhen Lei[2,3,4,⋆], and Lei Zhang[1⋆]

[1] The Hong Kong Polytechnic University
{guowen.zhang, liyi.chen}@connect.polyu.hk, cslzhang@comp.polyu.edu.hk
[2] Centre for Artificial Intelligence and Robotics, HKISI, CAS
[3] Institute of Automation, Chinese Academy of Sciences
[4] School of Artificial Intelligence, University of Chinese Academy of Sciences
{junsong.fan, zhaoxiang.zhang, zlei}@ia.ac.cn

Abstract. 3D object detection is an indispensable component for scene understanding. However, the annotation of large-scale 3D datasets requires significant human effort. To tackle this problem, many methods adopt weakly supervised 3D object detection that estimates 3D boxes by leveraging 2D boxes and scene/class-specific priors. However, these approaches generally depend on sophisticated manual priors, which is hard to generalize to novel categories and scenes. In this paper, we are motivated to propose a general approach, which can be easily adapted to new scenes and/or classes. A unified framework is developed for learning 3D object detectors from RGB images and associated 2D boxes. In specific, we propose three general components: *prior injection module* to obtain general object geometric priors from LLM model, *2D space projection constraint* to minimize the discrepancy between the boundaries of projected 3D boxes and their corresponding 2D boxes on the image plane, and *3D space geometry constraint* to build a Point-to-Box alignment loss to further refine the pose of estimated 3D boxes. Experiments on KITTI and SUN-RGBD datasets demonstrate that our method yields surprisingly high-quality 3D bounding boxes with only 2D annotation. The source code is available at https://github.com/gwenzhang/GGA.

Keywords: Weakly supervised object detection · 3D object detection · General scenarios

1 Introduction

The primary objective of 3D object detection is to detect tight bounding boxes from point clouds in 3D space, which is essential in many applications, such as autonomous driving [20, 69] and robotics [21]. Large-scale annotated datasets [3, 14, 60] are critical to training high performance detectors. However, annotating an adequate number of 3D labels is highly labor-intensive and costly,

⋆ Corresponding authors.

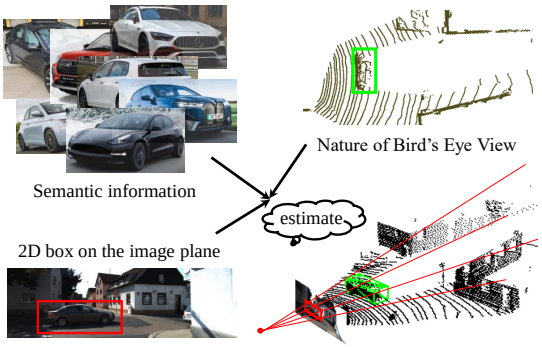

Fig. 1: The insight of our proposed method. Given RGB-D data, we utilize semantic information, 2D space projection constraints, and 3D space geometry constraints to estimate 3D boxes.

primarily due to the large workforce of human annotators [23, 58]. Therefore, seeking efficient proxy annotations for 3D object detection has become attractive [45, 63, 70, 74]. Previous works [43, 61] suggest that labeling 2D bounding boxes is 3 to 16 times faster than 3D bounding boxes. Additionally, many on-the-shelf large-scale models [24, 25] demonstrate excellent generalization capabilities for 2D objects, which can be a free source for 2D box surrogates. Therefore, in this work we are motivated to study how to utilize 2D boxes to train general-purpose 3D object detectors.

Some previous efforts attempted to learn 3D detectors from 2D boxes. For example, geometry-based methods [45, 70] leverage rule-based information from specific categories in LiDAR point clouds to convert 2D boxes to 3D boxes. Some semi-supervised methods [32, 39] employed some precisely annotated scenes and instances to regress 3D boxes from 2D boxes. Besides, some template-based methods [38, 75] adopt SDF [44] or synthetic data [4] for geometric alignment to convert 2D objects into 3D bounding boxes. These methods conduct good practice to use human prior knowledge and intuitive constraints for training. However, their designs are bound to specific object categories and/or model architectures, preventing them from generalizing to different scenes and categories.

To achieve the generalization of weakly supervised 3D detection, in this paper, we propose a General Geometry-Aware (GGA) method, which decomposes the problem into three sub-tasks of *prior injection*, *2D space projection constraints*, and *3D space geometry constraints*, as shown in Fig. 1. The prior injection provides the semantic shape information for the target categories to compensate for the 2D-3D information gap and reduce the data noise. Meanwhile, to ensure universality, it does not rely on any category-specific local geometries. We consider the ratio information as the prior and propose a semantic ratio loss (SRL) to inject ratio priors. The ratio is determined by the object's bird-eye view (BEV) width and height. The category-wise ratio information can be obtained either from humans or Large Language models (LLMs) [2, 64, 66]. While

for 2D space projection constraint, we propose a boundary projection loss (BPL) to ensure that the projected wireframe of the estimated 3D boxes aligns closely with the 2D boxes on the image. However, solely applying constraints on the 2D view space is insufficient for model convergence because there are multiple 3D boxes of different orientations and sizes that can project to the same 2D box. To deal with this problem, we propose a point-to-box alignment loss (PAL), which operates in 3D space and requires the predicted box to tightly wrap the nearby point cluster.

By integrating the three components, we finally arrive at a unified framework to handle the 2D-supervised 3D object detection problem. Our framework is not limited to specific model architectures or scene categories, and it is orthogonal to previous well-tested manual methods, enabling the integration of any category-specific methods. Results on the KITTI [14] and the SUN-RGBD [79] datasets demonstrate that our method is effective and it generalizes well across outdoor and indoor scenarios. In summary, the major contributions of our work are:

- We present a novel General Geometry-Aware (GGA) method for weakly supervised 3D object detection from 2D bounding boxes. It can significantly reduce the cost of 3D annotation.
- GGA exhibits promising generalization capabilities, allowing it to be easily extended to various novel scenarios and classes.
- Extensive experiments on both KITTI and SUN-RGBD benchmarks demonstrate that GGA can generate high-quality 3D boxes from 2D boxes.

2 Related Work

2.1 Outdoor 3D Object Detection

In outdoor scenes such as autonomous driving, point clouds are typically obtained from LiDAR sensors. Those scenes are often large-scale, sparse, and vary in density, which poses stricter requirements for designing efficient detection networks. There are three primary representations used for point cloud learning, point-based, range-view, and voxel-based. Voxel-based detectors [11, 15, 16, 27, 73, 76, 80] employ rasterization to convert point clouds into a volumetric representation, which is used to extract dense features by 3D convolution or Transformers. Point-based methods [30, 47–49, 56] directly extract semantic features from irregular points through a series of downsampling and set abstraction blocks. However, they are limited in the ability to process large-scale regions. Besides, some works convert point clouds into 2D birds-eye-view [27] or range view representations [12, 31, 40], which offer computational benefits compared to voxels. Some hybrid approaches [17, 35, 54, 55] combine voxel-based and point-based representations to balance the algorithm efficiency and accuracy.

2.2 Indoor 3D Object Detection

Different from outdoor scenes, indoor scenarios are characterized by cluttered and restricted environments. Objects are always dense and close to each other,

occupying a significant portion of scenes. Therefore, indoor 3D object detection has domain-specific data processing and approaches. Vote-based methods [8,46, 71,78] allocate groups of points to object candidates based on their voted center and extract object features from those point groups. Query-based approaches [34,41] emerge Transformers for end-to-end training, replacing the voting process with a forward pass during the inference stage. Similar to outdoor scenarios, voxel-based methods [19, 46, 53] convert points into voxel representations and process them using 3D convolution networks.

2.3 Weakly Supervised 3D Object Detection

To alleviate heavy 3D annotation costs, many weakly supervised methods and semi-supervised methods have been proposed. WS3D [39] introduces a weakly supervised framework for Lidar-based 3D object detection. It employs center clicks and precisely annotated scenes and instances to regress 3D bounding boxes, resembling a semi-supervised learning setting. Based on the shape prior derived from a pretrained DeepSDF [44] on synthetic datasets, SDF [74] can estimate the 3D geometry of cars detected in 2D images. FGR [70] utilizes a non-learning approach to generate pesudo 3D labels by leveraging the specific patterns observed in vehicle point clouds and the frustum. McCraith *et al.* [38] combine fixed 3D mesh template and yaw predictors to lift 2D mask to 3D box estimates. Weak3DM [45] predicted the 3D boxes from foreground LiDAR points based on hand-craft priors for monocular 3D object detection. Weakmono3D [63] incorporated projection consistency and multi-view consistency to estimate 3D boxes. Although those methods can lift 2D boxes to 3D box estimates, the sophisticated priors make them hard to extend to novel classes and scenarios.

3 Methods

In this section, we first present an overview in Section 3.1. Then, we introduce the data preparation, backbone, and proposal head of our proposed GGA framework in Section 3.2. Finally, we provide a detailed explanation of our proposed BPL, SRL, and PAL from Section 3.3 to Section 3.5.

3.1 Overview

An overview of our proposed GGA is shown in Fig. 2. Given an RGB image and its corresponding point clouds, which can be obtained from LiDAR scans or depth cameras, we aim to classify and localize objects within the scene in 3D space by using 2D bounding boxes B_{2d}. GGA adopts an end-to-end learning-based approach, with the scene point clouds and their corresponding 2D boxes on the image plane as inputs. For data preprocessing, we follow previous works [32, 45, 70] to select point clouds that can be projected into 2D boxes on the image plane as the initial representation of the corresponding objects, termed *In-Box-Points*. Then we estimate the minimum-perimeter 3D pseudo boxes from *In-Box-Points* to provide supervision. During training, we query the category-specific

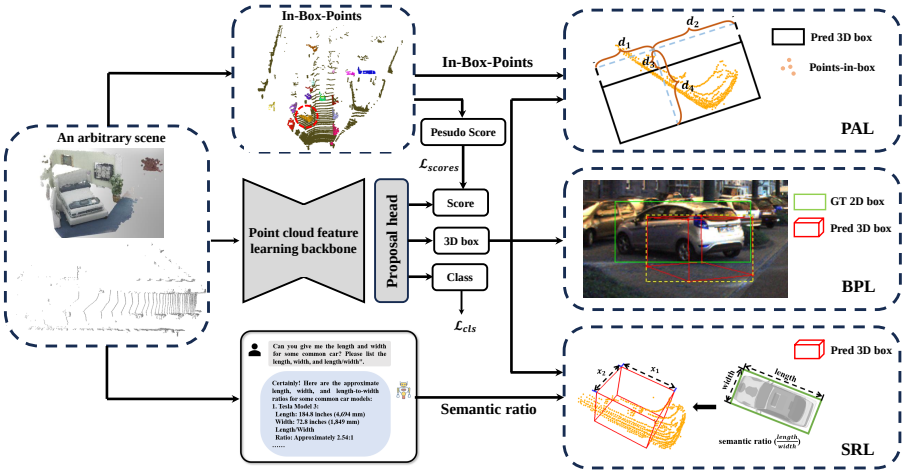

Fig. 2: Illustration of our proposed GGA approach. Initially, point clouds of a given scene are processed through the point cloud feature learning backbone and proposal head. Meanwhile, we extract the *In-Box-Points* from the 3D frustum corresponding to a 2D bounding box. The prior ratio information is obtained from GPT-4 and incorporated into the network through SRL. The PAL provides the 3D space geometry constraint, while the BPL enforces the 2D space projection constraint.

ratio information from LLMs (i.e., GPT-4 [2]) and inject such priors with the proposed *Semantic Ratio Loss* (SRL). Note that different from previous work [38, 45, 70, 74] using manually-collected priors, we fully leverage the LLMs which can provide general cues freely and easily. Besides, to provide supervision from the 2D perspective, we constrain the geometric relationship between 2D and 3D boxes through a *Boundary Projection Loss* (BPL). To further deal with the ill-posed problem of estimating 3D boxes from 2D boxes, we propose a *Point-to-box Alignment Loss* (PAL) which implicitly supervises 3D boxes and alleviates inherent ambiguity.

3.2 The Proposed GGA Framework

Data Preparation. To obtain the essential components and apply the unified framework of GGA, we first preprocess point clouds of the scene. Indoor point clouds, compared with outdoor scenarios, are more dense and abundant. To reduce computational complexity, we employ the Farthest Point Sampling [48] to reduce raw point clouds to 10^6 for indoor scenarios. Besides, following previous works [45, 68, 70], we remove ground points with the RANSAC algorithm [13] to reduce noise in the subsequent *In-Box-Points*.

In-Box-Points. Given the Point-Image calibration parameters P_{proj}, 3D point clouds can be projected to the image plane. We select those points within

the 2D bounding boxes as the valid points for the corresponding object. Additionally, the selected points often contain outliers that belong to nearby objects and backgrounds. Following previous methods [38, 45, 70], we use a pre-trained 2D instance segmentation network [25] to filter outliers from valid foreground points in indoor scenarios. The outdoor valid points are denoised by the region growing algorithm [1]. The resulting point clouds, denoted as *In-Box-Points*: $\mathcal{P}_{ibp} \in \mathbb{R}^{N\times 3}$, represent the salient points associated with objects. Furthermore, we take the minimum-perimeter 3D box that encases all *In-Box-Points* for each object as our initial pseudo box.

Backbone. Recent indoor and outdoor 3D object detection networks consistently utilize domain-specific backbones and heads. Therefore, our study employs one-stage CenterPoint [73] and FCAF3D [52] for outdoor and indoor scenarios, respectively. Both of them are anchor-free methods. Specifically, CenterPoint uses a 3D encoder to quantize the point clouds into regular bins and then extract features from those voxels. The backbone utilized in FCAF3D is a sparse adaptation of ResNet [18], wherein all 2D convolutions are replaced by sparse 3D convolutions.

Proposal head. As shown in Fig. 2, the proposal head mainly outputs the objectness scores, predicted 3D boxes, and object classes. Due to inherent distinctions, the objectness scores of CenterPoint and FACF3D differ. CenterPoint generates a K-channel heatmap, with each channel representing one class. FACF3D, like FCOS [65], utilizes centerness to represent the presence of an object at a specific location. We use initial pseudo boxes to generate pseudo objectness scores, providing the necessary supervision. For unified subsequent processing, we standardize the predicted 3D box properties to ensure consistency in both indoor and outdoor regressions: a location refinement $(x_r, y_r, z_r) \in \mathbb{R}^3$, a 3D size $(l, w, h) \in \mathbb{R}^3$, and a yaw rotation angle $(sin(\alpha), cos(\alpha)) \in \mathbb{R}^2$. For classification, FCAF3D generates classification probability $\hat{p}$, while CenterPoint determines the object class based on class-specific heatmaps. Finally, by combining the activation position and predicted box properties, we can derive the predicted 3D boxes as:

$$B^p_{3d} : (x, y, z, l, w, h, sin(\alpha), cos(\alpha)) \in \mathbb{R}^8, \tag{1}$$

where the coordinates (x, y, z) represent the center of the predicted 3D box.

3.3 Boundary Projection Loss

With the Point-Image calibration parameters, a 2D bounding box can be lifted to a frustum, establishing a 3D search space for the object of interest [47, 70]. This frustum can impose 2D space projection constraints on the predicted 3D boxes, thereby reducing the searching space. Specifically, in outdoor scenarios, the minimum bounding rectangles of projected 3D bounding boxes on the image plane tend to align closely with the 2D bounding boxes. Inspired by this observation, we propose BPL to locate objects' positions from an image perspective. In particular, we project the 3D boxes onto the image as follows:

$$(x^p_{min}, y^p_{min}, x^p_{max}, y^p_{max}) = \mathbf{Proj}(\mathbf{Corners}(B^p_{3d}), P_{proj}), \tag{2}$$

where **Corners**$(\cdot)$ denotes obtaining the eight corners' coordinates of each predicted 3D box. Besides, **Proj**$(\cdot)$ represents the projection of these corners onto the image by calibration parameters. $(x^p_{min}, y^p_{min}, x^p_{max}, y^p_{max})$ represents the boundaries of the minimum bounding rectangle of those projected corners. As shown in Fig. 2, we minimize the discrepancy between the boundaries of the minimum bounding box and the corresponding 2D box as follows:

$$\mathcal{L}_{\text{BPL}} = L(x^p_{min}, x_{min}) + L(x^p_{max}, x_{max}) + L(y^p_{min}, y_{min}) + L(y^p_{max}, y_{max}), \quad (3)$$

where $L(\cdot, \cdot)$ is the L_1 loss and $(x_{min}, y_{min}, x_{max}, y_{max})$ means the boundaries of B_{2d}. BPL is applied to all the objects of interest and computes their average.

Different from outdoor scenarios, the indoor objects are usually closed to the camera, and their images are captured with diverse camera poses, resulting in a significant alignment gap between projected 3D boxes and 2D boxes. To address the problem, we project initial pseudo boxes obtained from *In-Box-Points* onto the image plane to obtain the pseudo 2D box. Then the minimum bounding rectangle that encompasses the pseudo and original 2D box servers as the new 2D constraint. Overall, we have addressed this issue using a simple yet effective approach, without resorting to any additional information.

3.4 Semantic Ratio Loss

While the BPL can offer 2D space projection constraints for predicted 3D boxes, it remains challenging to accurately perceive the shape of objects. Previous methods commonly rely on synthetic data [38, 74] or category-specific local geometries [45, 70] as shape priors to help learning semantic information. However, integrating sophisticated hand-craft rules brings challenges for generalization. We observe that fine-grained priors in the network are unnecessary; basic ratio information is sufficient to help the model capture object shapes. Therefore, we propose a semantic ratio loss to exploit the ratio information for model optimization. In detail, we take the off-the-shelf GPT-4 to get the required information, noting that large language models contain countless relevant knowledge [2,64,66]. Given the predicted 3D box parameters of $(x, y, z, l, w, h, \alpha)$, we aim to apply SRL to supervise the objects' bird-eye view width and height, *i.e.*, $\frac{l}{w}$. We find that compared to explicitly specifying the output meaning, taking the shorter one as width while another as height achieves better performance. Particularly, we obtain the predicted ratio with $\frac{min(l,w)}{max(l,w)}$. SRL is formulated as:

$$\mathcal{L}_{\text{SRL}} = L(\frac{min(l, w)}{max(l, w)}, r), \quad (4)$$

where r is the average ratio of various examples generated from GPT-4 and $L(\cdot, \cdot)$ represents the L_1 loss. With the prior information from SRL, GGA can converge faster and achieve better performance.

3.5 Points-to-Box Alignment Loss

A 3D box can be parameterized by multiple properties, such as location, dimension, and rotation. Methods following the fully supervised paradigm can easily disentangle these properties with annotations. However, in a weakly supervised manner, there is no general method available to disentangle those properties. The inherent ambiguity prevents the network from learning accurate 3D boxes. For example, if the predicted rotation deviates by θ degrees from the ground truth, other predicted properties (*e.g.*, length and width) can be adjusted to satisfy the requirements. Hence, we propose to use *In-Box-Points* to add implicit supervision for decoupling relevant information in 3D space.

Generally speaking, 3D boxes should cover all the foreground points. We apply the 3D space geometry supervision to predicted 3D boxes based on the BEV location of $\mathcal{P}_{ibp}$. As shown in Fig. 2, we first compute the distance $(d_i^1, d_i^2, d_i^3, d_i^4)$ from each point in $\mathcal{P}_{ibp}$ to the four edges of the predicted 3D in the BEV. Therefore, the 3D space constraint on $\mathcal{P}_{ibp}$ is formulated as:

$$\mathcal{L}_{\text{PAL}_1} = \sum_{i=1}^{N}\Big(\sum_{j\in\{1,2\}} \phi(d_i^j - \frac{l}{2}) + \sum_{k\in\{2,3\}} \phi(d_i^k - \frac{w}{2})\Big), \tag{5}$$

where l, w is the predicted length and width and ϕ is the activation function ReLU. In addition, RGB-D data capture only a single view of objects, resulting in point clouds that cover only one side of the objects. Thus, point clouds tend to be concentrated near the boundaries of 3D boxes in BEV. Inspired by this observation, we propose using this prior to perform implicit supervision. We further minimize the shortest distance between each point and the four edges of the predicted box in the BEV as follows:

$$\mathcal{L}_{\text{PAL}_2} = \sum_{i=1}^{N} \min(d_i^1, d_i^2, d_i^3, d_i^4). \tag{6}$$

3.6 Network Optimization

We train the network using a combination of our proposed losses, objectness scores loss $\mathcal{L}_{\text{score}}$, and cross-entropy loss $\mathcal{L}_{\text{cls}}$ for classification probability:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{\text{BPL}} + \lambda_2 \cdot \mathcal{L}_{\text{SRL}} + \lambda_3 \cdot (\mathcal{L}_{\text{PAL}_1} + \mathcal{L}_{\text{PAL}_2}) + \lambda_4 \cdot \mathcal{L}_{\text{score}} + \lambda_5 \cdot \mathcal{L}_{\text{cls}}, \tag{7}$$

where λ_{1-5} are the loss weight parameters. $\mathcal{L}_{\text{score}}$ denotes the heatmap regression loss in CenterPoint and centerness loss in FCAF3D. Finally, following the common protocol of weakly supervised learning [6, 45, 70], we use the predicted 3D boxes as pseudo labels to train the 3D object detection network under a fully supervised setting.

Table 1: Results on the KITTI val detection benchmark for monocular 3D object detection. Symbol '-' means that result is not available in the original paper. † means that methods are evaluated with metric $\mathbf{AP}_{40}$. Others are based on $\mathbf{AP}_{11}$.

Methods	Supversion	$\mathbf{AP_{BEV}}/\mathbf{AP_{3D}}^{\text{Car(Iou=0.7)}}$		
		Easy.	Mod.	Hard.
FQNet [33]	3D Box	9.50/5.98	8.02/5.50	7.71/4.75
Deep3DBox [42]	3D Box	9.99/5.85	7.71/4.10	5.30/3.84
OFTNet [51]	3D Box	11.06/4.07	8.79/3.27	8.91/3.29
RoI-10D [37]	3D Box	14.50/10.25	9.91/6.39	8.73/6.18
MonoDIS [57]	3D Box	24.26/18.05	18.43/14.98	16.95/13.42
MonoPSR [26]	3D Box	20.63/12.75	18.67/11.48	14.45/8.59
D4LCN [10]	3D Box	26.00/19.38	20.73/16.00	17.46/12.94
MonoRun [5]	3D Box	- /20.02	- /14.65	- /12.61
PatchNet [36]	3D Box	32.30/25.76	21.25/17.72	19.04/15.62
PGD [67]	3D Box	30.56/24.35	23.67/18.34	20.84/16.90
AutoSDF [74]	2D Box	15.70/1.23	10.52/0.54	-/-
WeakM3D [45]	2D Box	24.89/17.06	16.47/11.63	14.09/11.17
PGD+GGA	2D Box	27.70/17.71	20.39/13.37	18.04/10.89
DCD† [29]	3D Box	32.55/23.81	21.50/15.90	18.25/13.21
MonoDTR† [22]	3D Box	33.33/24.52	25.35/18.57	21.68/15.51
MonoDETR† [77]	3D Box	- /28.84	- /20.61	- /16.38
MonoDETR+GGA†	2D Box	30.07/21.18	21.49/14.96	18.23/12.25

Table 2: Results on the KITTI test detection benchmark for monocular 3D object detection.

Methods	Supversion	$\mathbf{AP_{BEV}}/\mathbf{AP_{3D}}^{\text{Car(Iou=0.7)}}$		
		Easy.	Mod.	Hard.
FQNet [33]	3D Box	5.40/2.77	3.23/1.51	2.46/1.01
GS3D [28]	3D Box	8.41/4.47	6.08/2.90	4.94/2.47
ROI-10D [37]	3D Box	9.78/4.32	4.91/2.02	3.74/1.46
MonoDIS [57]	3D Box	17.23/10.37	13.19/7.94	11.12/6.40
WeakM3D [45]	2D Box	11.82/5.03	5.66/2.26	4.08/1.63
WeakMono3D [63]	2D Box	12.31/6.98	8.80/4.85	7.81/4.45
PGD+GGA	2D Box	17.42/10.42	10.21/6.08	8.09/4.65

4 Experiments

4.1 Experimental Setup

Dataset and evaluation metric. We conduct the monocular 3D object detection, Lidar-based 3D object detection, and indoor 3D object detection experiments on the SUN-RGBD [79] and KITTI [14] benchmarks. SUN-RGBD benchmark contains more than 10,000 indoor RGB-D images. The training and validation splits contain 5,285 and 5,050 point cloud scenes. Following previous work [52], we evaluate on 10 classes. The KITTI dataset includes 3,712 images for training, 3,769 images for validation, and 7,518 images for testing. The KITTI benchmark provides evaluations for cars, pedestrians and cyclists. The evaluation metric is the mean average precision (mAP) with 3D intersection-over-union (IOU) of 0.25 for SUN-RGBD and 0.25/0.5/0.7 for KITTI.

Table 3: Results on KITTI test BEV detection benchmark for Lidar-based 3D object detection methods.

Methods	Supversion	Car (Iou=0.7)			Pedestrain (Iou=0.5)			Cyclist (Iou=0.5)		
		Easy.	Mod.	Hard.	Easy.	Mod.	Hard.	Easy.	Mod.	Hard.
MV3D [7]	3D Box	86.02	76.90	68.49	-	-	-	-	-	-
VoxelNet [80]	3D Box	89.35	79.26	77.39	39.48	33.69	31.50	61.22	48.36	44.37
Second [72]	3D Box	88.07	79.37	77.95	55.10	46.27	44.76	73.67	56.04	48.78
F-PointNet [47]	3D Box	88.70	84.00	75.33	51.21	44.89	40.23	72.27	56.12	49.01
PointPillars [27]	3D Box	88.35	86.10	79.83	52.08	43.53	41.49	75.78	59.07	52.92
PointRCNN [56]	3D Box	89.28	86.04	79.02	49.43	41.78	38.63	81.36	67.23	59.35
PointRCNN+GGA	2D Box	77.25	63.27	54.70	12.07	11.27	10.47	61.06	46.75	41.84

Table 4: AP@0.25 scores for 10 object categories from the SUN-RGBD.

Methods	Supervision	bathtub	bed	bkshelf	chair	desk	dresser	nstand	sofa	table	toilet	mAP
DSS [59]	3D Box	44.2	78.7	11.9	61.2	20.5	6.4	15.4	53.5	50.3	78.9	42.1
VoteNet [46]	3D Box	74.4	83.0	28.8	75.3	22.0	29.8	62.2	64.0	47.3	90.1	57.7
H3DNet [78]	3D Box	73.8	85.6	31.0	76.7	29.6	33.4	65.5	66.5	50.8	88.2	60.1
GroupFree [34]	3D Box	80.0	87.8	32.5	79.4	32.6	36.0	66.7	70.0	53.8	91.1	63.0
FCAF3D [52]	3D Box	79.0	88.3	33.3	81.1	34.0	40.1	71.9	69.7	53.0	91.3	64.2
BoxPC [62]	2D Box	29.5	60.9	6.7	36.0	20.2	27.3	50.9	46.4	28.4	65.3	37.2
FCAF3D+GGA	2D Box	55.4	69.9	22.4	59.1	22.5	31.3	59.3	58.9	34.8	71.4	48.5

Implementation details. In our experiments, we adopt the CenterPoint [73] and FCAF3D [52] as the backbone for indoor and outdoor scenarios, respectively. We implement our method with mmdetection3d [9] and train it with the AdamW optimizer. The RANSAC threshold is set to 0.04 and 0.2 for indoor and outdoor scenarios, respectively. The five parameters λ_{1-5} in the loss function are set to 2e-3, 2e-3, 1e-4, 1 and 1 for SUN-RGBD dataset. Following CenterPoint, we omit $\mathcal{L}_{\text{cls}}$ and set λ_{1-4} as 0.3, 0.1, 0.1 and 5 for KITTI benchmark. We train our network for 120 epochs for outdoor scenarios and 12 epochs for indoor scenarios, respectively. For the Lidar-based and monocular 3D object detection network, we leverage the generated pseudo 3D bounding boxes to train models (PointRCNN [56], PGD [67], and MonoDETR [77]) to obtain the final detection results. For indoor scenarios, we utilize FCAF3D [52].

4.2 Experimental Results

Outdoor scenarios. Table 1 shows the evaluation results on the KITTI validation set for monocular 3D object detection, using metrics of AP_{3D} and AP_{BEV}. Our proposed method significantly outperforms other weakly supervised methods. MonoDETR with our generated pseudo label achieves 21.18/14.96/12.25 on AP_{3D} for the easy, moderate and hard categories, which are +4.12, +3.33, and + 0.31 better than WeakM3D. Moreover, although our method does not utilize any 3D bounding box annotations, it outperforms several previous state-of-the-art fully supervised methods. For instance, our approach outperforms MonoRun [5] by +1.16 and +0.31 for AP_{3D} in the easy and moderate settings. In Table 2, we compare the metrics on the test set for cars with other fully supervised and weakly supervised methods. Our approach clearly outperforms the state-of-the-

Table 5: Comparisons of different weakly supervised methods on KITTI val set. Symbol '×' means that the method fails to handle the category.

Input	Methods	$AP_{BEV}/AP_{3D^{Car(Iou=0.5)}}$			$AP_{BEV}/AP_{3D^{Cyclist(Iou=0.25)}}$			$AP_{BEV}/AP_{3D^{Pedestrian(Iou=0.25)}}$		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Image	VS3D [50]	31.59/22.62	20.59/14.43	16.28/10.91	×	×	×	×	×	×
	AutoSDF [74]	50.51/38.31	30.97/19.90	23.72/14.83	×	×	×	×	×	×
	WeakM3D [45]	58.20/50.16	38.02/29.94	30.17/23.11	×	×	×	×	×	×
	WeakMono3D [63]	54.32/49.37	42.83/39.01	**40.07/36.34**	-	-	-	-	-	-
	GGA	**61.45/53.49**	**43.95/39.50**	37.09/34.22	25.03/23.68	15.77/14.87	14.35/13.99	10.96/9.98	9.76/8.75	8.49/7.87
LiDAR	VS3D [50]	74.50/40.32	66.71/37.36	57.55/31.09	×	×	×	×	×	×
	AutoSDF [74]	77.84/62.25	59.75/42.23	-	×	×	×	×	×	×
	Lifting [38]	86.52/83.45	86.22/79.53	75.53/71.01	×	×	×	×	×	×
	GGA	**95.25/94.83**	**87.73/85.19**	**78.84/76.18**	88.10/86.59	65.68/65.34	61.20/60.78	66.86/65.06	65.61/63.68	61.88/60.01

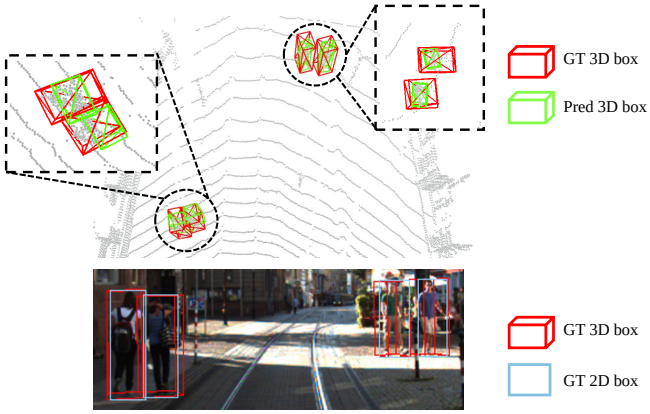

Fig. 3: Generated pseudo 3D boxes of pedestrians by our proposed GGA.

art weakly supervised methods WeakMono3D on all metrics. Furthermore, our method achieves comparable results with some fully supervised methods.

Table 3 shows the performance of our approach on Lidar-based 3D object detection for AP_{BEV} on KITTI test set. Compared with images, point clouds can provide more precise location information and achieve better performance. Compared to fully supervised methods, the performance gap between the easy and hard categories for GGA is notably larger. This is because predicting pseudo labels for objects with few foreground points is difficult. It is worth noting that the performance of pedestrian detection is comparatively lower. The phenomenon occurs because the annotation process of 3D bounding boxes involves inherent ambiguity. While the ground truth 3D annotation is highly accurate on KITTI, it may not present the only possible solution. Additionally, we give the pedestrian pseudo label generated by our GGA approach in Fig. 3. It can be observed that our predicted 3D boxes tend to be enclosed by 3D ground truth boxes. Compared with the 3D ground truth, our predicted pseudo 3D boxes may be more tightly aligned with the point clouds. It is challenging to determine which type of annotation is better. Besides, we give the results on the metrics $iou = 0.25$ for the pedestrian in Table 5.

Table 6: Ablation study on the three proposed loss functions on the KITTI val set for Lidar-based 3D object detection.

BPL	SRL	PAL	Car (Iou=0.7)			Car (Iou=0.5)		
			Easy.	Mod.	Hard.	Easy.	Mod.	Hard.
✓	×	×	7.92	5.74	4.60	66.33	56.11	48.90
✓	×	✓	26.39	17.73	14.21	78.11	64.93	56.96
✓	✓	×	42.21	28.02	27.45	72.94	61.07	54.35
✓	✓	✓	**56.02**	**40.35**	**33.15**	**84.03**	**67.81**	**58.73**

Table 7: Ablation study on the semantic ratio on the KITTI val set for Lidar-based 3D object detection.

Semantic Ratio	Car (Iou=0.7)			Car (Iou=0.5)		
	Easy	Mod.	Hard	Easy	Mod.	Hard
1.9	47.35	33.82	28.69	78.54	64.08	**59.65**
2.0	46.25	32.82	27.86	78.71	64.17	57.40
2.1	44.44	30.74	29.94	73.53	61.72	56.85
2.2	46.90	33.34	28.67	76.05	62.31	56.10
2.3	53.00	37.21	31.74	80.68	65.69	57.23
2.4	**56.02**	**40.35**	33.15	**84.03**	**67.81**	58.73
2.5	51.20	36.56	31.09	78.02	63.30	56.66
2.6	51.32	39.31	**33.33**	76.64	62.98	57.11
2.7	54.74	39.74	32.90	77.37	63.70	57.19
2.8	44.05	32.13	27.26	74.45	61.84	55.64
2.9	42.56	32.81	26.96	70.48	59.38	53.86

Indoor scenarios. For the indoor scenarios, we evaluate our method on the SUN-RGBD validation set, using metrics of $AP_{0.25}$. Tab. 4 shows that our proposed framework achieves promising results and outperforms prior fully supervised method DSS. Compared with the semi-supervised BoxPC [62], one can see that even without using any auxiliary 3D boxes, our method significantly outperforms it in all categories.

Comparison with other weakly supervised methods. We compare our methods with other weakly supervised methods on both Lidar-based and monocular 3D object detection. As shown in Tab. 5, other weakly supervised 3D object detection methods are limited to vehicle categories because they rely on sophisticated manual priors. Although our method is not specifically designed for the car category, GGA still achieves the best results in almost all metrics. For monocular 3D object detection, GGA's performance is slightly lower than WeakMono3D in the hard samples. This is because WeakMono3D takes advantage of 2D direction labels, which are useful for estimating rotation. Other weakly supervised 3D object detection methods are limited to vehicle categories because they rely on sophisticated manual priors.

4.3 Ablation Study

To demonstrate the effects of our proposed method, we perform ablation studies for GGA in outdoor scenarios. GGA is trained on the KITTI dataset, and we evaluate the quality of the generated pseudo labels against 3D ground truth. In addition, we conduct sensitivity analyses on ratio parameters to demonstrate the robustness of our method.

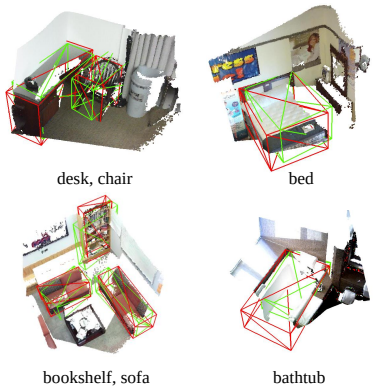

Fig. 4: Visualization of generated 3D boxes for indoor scenarios.

Fig. 5: Visualization of generated 3D boxes in bird's eye view for outdoor scenarios.

As shown in Table 6, we select the backbone with BPL as the initial starting point because we observe that solely relying on SRL or PAL makes it difficult for the model to converge. The 2D perspective constraint is considered essential for object localization. The combination of BPL and PAL brings improvements of +18.47 and +9.61 over using only BPL on $AP_{iou=0.7}$ for the easy and hard samples. Besides, combining BPL with SRL can bridge the 2D and 3D geometry gap and achieve +34.29 and + 22.85 $AP_{iou=0.7}$ gains over BPL. This indicates that semantic prior information is crucial in perceiving the shape of objects. Finally, utilizing all the proposed losses can achieve the best performance. PAL provides additional refinement to the predicted 3D boxes by further adjusting their location, rotation and dimensions in conjunction with BPL and SRL. Compared to using only BPL and PAL, incorporating SRL adds supplementary geometric information and enhances perceptual capabilities. The ablation study demonstrates the complementary nature of the three proposed losses, resulting in improved 3D box predictions.

We further analyze the effect of the ratio information in Equation 4 on the quality of generated pseudo labels. In Table 7, we present evaluation results of

generated pseudo 3D boxes on the KITTI validation set. GGA achieves optimal results when the semantic ratio is 2.4. Our proposed method demonstrates a comparable average $AP_{0.7}$ performance ranging from 39.62 to 43.17 across all difficulty levels when the semantic ratio falls within 2.3 to 2.7, implying that our method exhibits certain robustness to the ratio.

4.4 Qualitative Results

Fig. 4 illustrates the generated pseudo 3D boxes by GGA in indoor scenarios. We see that the pseudo 3D boxes are very close to the ground truth 3D boxes. This indicates that our method can effectively learn the various attributes of 3D bounding boxes. In Fig. 5, we visualize the generated pseudo 3D boxes in outdoor scenes. In the left image, the cars exhibit diverse rotations, yet our method can effectively handle such situations. Furthermore, most objects in the left and middle images are significantly occluded, and our approach still accurately predicts the 3D boxes in such challenging scenarios. The right image demonstrates that GGA can generate high-quality 3D boxes for both pedestrians and cyclists. The visualizations of both indoor and outdoor scenarios validate the generalization ability of GGA across various categories and scenes.

5 Limitation

In this paper, we aim to build a bridge between 2D bounding boxes and high-quality 3D boxes for diverse objects. However, those objects that are far away from the camera contain very few point clouds. The limited point clouds lead to a weak robustness against noise, and it is hard to predict the attributes (*e.g.*, rotation, location, dimension) for 3D boxes with only a few scattered points. In the future, we plan to explore how to transfer knowledge from objects containing dense points to objects containing very few points.

6 Conclusion

In this paper, we focused on the generalization issue for weakly supervised 3D object detection. Specifically, we proposed a novel general geometry-aware approach to replace the sophisticated hand-craft rules to derive 3D boxes from 2D. To address the challenge of training on diverse scenes and categories, we indicated that the key is to set up a unified framework for prior injection, 2D space projection constraint and 3D space geometry constraint. First, we introduced BPL, which serves as a 2D space projection constraint derived from the image plane. Then, we used ratio information as prior information to help the network to learn semantic knowledge. Last, we proposed the PAL to refine the box pose in bird's eye view. Extensive experiments on the KITTI and SUN-RGBD benchmarks demonstrated that our method achieved high performance and it could be easily generalized to novel categories and scenes.

Acknowledgement This work was supported in part by the InnoHK Program.

References

1. Adams, R., Bischof, L.: Seeded region growing. IEEE Transactions on pattern analysis and machine intelligence **16**(6), 641–647 (1994) 6
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 2, 5, 7
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020) 1
4. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015) 2
5. Chen, H., Huang, Y., Tian, W., Gao, Z., Xiong, L.: Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10379–10388 (2021) 9, 10
6. Chen, L., Lei, C., Li, R., Li, S., Zhang, Z., Zhang, L.: Fpr: False positive rectification for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1108–1118 (2023) 8
7. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1907–1915 (2017) 10
8. Cheng, B., Sheng, L., Shi, S., Yang, M., Xu, D.: Back-tracing representative points for voting-based 3d object detection in point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8963–8972 (2021) 4
9. Contributors, M.: MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d (2020) 10
10. Ding, M., Huo, Y., Yi, H., Wang, Z., Shi, J., Lu, Z., Luo, P.: Learning depth-guided convolutions for monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition workshops. pp. 1000–1001 (2020) 9
11. Fan, L., Pang, Z., Zhang, T., Wang, Y.X., Zhao, H., Wang, F., Wang, N., Zhang, Z.: Embracing single stride 3d object detector with sparse transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8458–8468 (2022) 3
12. Fan, L., Xiong, X., Wang, F., Wang, N., Zhang, Z.: Rangedet: In defense of range view for lidar-based 3d object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2918–2927 (2021) 3
13. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981) 5
14. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012) 1, 3, 9

15. He, C., Li, R., Zhang, G., Zhang, L.: Scatterformer: Efficient voxel transformer with scattered linear attention. arXiv preprint arXiv:2401.00912 (2024) 3
16. He, C., Li, R., Zhang, Y., Li, S., Zhang, L.: Msf: Motion-guided sequential fusion for efficient 3d object detection from point cloud sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5196–5205 (2023) 3
17. He, C., Zeng, H., Huang, J., Hua, X.S., Zhang, L.: Structure aware single-stage 3d object detection from point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11873–11882 (2020) 3
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 6
19. Hou, J., Dai, A., Nießner, M.: 3d-sis: 3d semantic instance segmentation of rgb-d scans. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4421–4430 (2019) 4
20. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17853–17862 (2023) 1
21. Huang, J., Yong, S., Ma, X., Linghu, X., Li, P., Wang, Y., Li, Q., Zhu, S.C., Jia, B., Huang, S.: An embodied generalist agent in 3d world. arXiv preprint arXiv:2311.12871 (2023) 1
22. Huang, K.C., Wu, T.H., Su, H.T., Hsu, W.H.: Monodtr: Monocular 3d object detection with depth-aware transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4012–4021 (2022) 9
23. Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q., Yang, R.: The apolloscape open dataset for autonomous driving and its application. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2702–2719 (2019) 2
24. Ke, L., Ye, M., Danelljan, M., Liu, Y., Tai, Y.W., Tang, C.K., Yu, F.: Segment anything in high quality. arXiv preprint arXiv:2306.01567 (2023) 2
25. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 2, 6
26. Ku, J., Pon, A.D., Waslander, S.L.: Monocular 3d object detection leveraging accurate proposals and shape reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11867–11876 (2019) 9
27. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019) 3, 10
28. Li, B., Ouyang, W., Sheng, L., Zeng, X., Wang, X.: Gs3d: An efficient 3d object detection framework for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1019–1028 (2019) 9
29. Li, Y., Chen, Y., He, J., Zhang, Z.: Densely constrained depth estimator for monocular 3d object detection. In: European Conference on Computer Vision. pp. 718–734. Springer (2022) 9
30. Li, Z., Wang, F., Wang, N.: Lidar r-cnn: An efficient and universal 3d object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7546–7555 (2021) 3

31. Liang, Z., Zhang, Z., Zhang, M., Zhao, X., Pu, S.: Rangeioudet: Range image based real-time 3d object detector optimized by intersection over union. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7140–7149 (2021) 3
32. Liu, C., Qian, X., Huang, B., Qi, X., Lam, E., Tan, S.C., Wong, N.: Multimodal transformer for automatic 3d annotation and object detection. In: European Conference on Computer Vision. pp. 657–673. Springer (2022) 2, 4
33. Liu, L., Lu, J., Xu, C., Tian, Q., Zhou, J.: Deep fitting degree scoring network for monocular 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1057–1066 (2019) 9
34. Liu, Z., Zhang, Z., Cao, Y., Hu, H., Tong, X.: Group-free 3d object detection via transformers. 2021 ieee. In: CVF International Conference on Computer Vision (ICCV). pp. 2929–2938 (2021) 4, 10
35. Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel cnn for efficient 3d deep learning. Advances in Neural Information Processing Systems **32** (2019) 3
36. Ma, X., Liu, S., Xia, Z., Zhang, H., Zeng, X., Ouyang, W.: Rethinking pseudo-lidar representation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. pp. 311–327. Springer (2020) 9
37. Manhardt, F., Kehl, W., Gaidon, A.: Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2069–2078 (2019) 9
38. McCraith, R., Insafutdinov, E., Neumann, L., Vedaldi, A.: Lifting 2d object locations to 3d by discounting lidar outliers across objects and views. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2411–2418. IEEE (2022) 2, 4, 5, 6, 7, 11
39. Meng, Q., Wang, W., Zhou, T., Shen, J., Van Gool, L., Dai, D.: Weakly supervised 3d object detection from lidar point cloud. In: European Conference on computer vision. pp. 515–531. Springer (2020) 2, 4
40. Meyer, G.P., Laddha, A., Kee, E., Vallespi-Gonzalez, C., Wellington, C.K.: Lasernet: An efficient probabilistic 3d object detector for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12677–12686 (2019) 3
41. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2906–2917 (2021) 4
42. Mousavian, A., Anguelov, D., Flynn, J., Kosecka, J.: 3d bounding box estimation using deep learning and geometry. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 7074–7082 (2017) 9
43. Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V.: Extreme clicking for efficient object annotation. In: Proceedings of the IEEE international conference on computer vision. pp. 4930–4939 (2017) 2
44. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 165–174 (2019) 2, 4
45. Peng, L., Yan, S., Wu, B., Yang, Z., He, X., Cai, D.: Weakm3d: Towards weakly supervised monocular 3d object detection. arXiv preprint arXiv:2203.08332 (2022) 2, 4, 5, 6, 7, 8, 9, 11

46. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9277–9286 (2019) 4, 10
47. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum pointnets for 3d object detection from rgb-d data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 918–927 (2018) 3, 6, 10
48. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017) 3, 5
49. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017) 3
50. Qin, Z., Wang, J., Lu, Y.: Weakly supervised 3d object detection from point clouds. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4144–4152 (2020) 11
51. Roddick, T., Kendall, A., Cipolla, R.: Orthographic feature transform for monocular 3d object detection. arXiv preprint arXiv:1811.08188 (2018) 9
52. Rukhovich, D., Vorontsova, A., Konushin, A.: Fcaf3d: Fully convolutional anchor-free 3d object detection. In: European Conference on Computer Vision. pp. 477–493. Springer (2022) 6, 9, 10
53. Shen, X., Stamos, I.: Frustum voxnet for 3d object detection from rgb-d or depth images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1698–1706 (2020) 4
54. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10529–10538 (2020) 3
55. Shi, S., Jiang, L., Deng, J., Wang, Z., Guo, C., Shi, J., Wang, X., Li, H.: Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. International Journal of Computer Vision **131**(2), 531–551 (2023) 3
56. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 770–779 (2019) 3, 10
57. Simonelli, A., Bulo, S.R., Porzi, L., López-Antequera, M., Kontschieder, P.: Disentangling monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1991–1999 (2019) 9
58. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 567–576 (2015) 2
59. Song, S., Xiao, J.: Deep sliding shapes for amodal 3d object detection in rgb-d images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 808–816 (2016) 10
60. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020) 1
61. Tang, Y.S., Lee, G.H.: Transferable semi-supervised 3d object detection from rgb-d data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1931–1940 (2019) 2

62. Tang, Y.S., Lee, G.H.: Transferable semi-supervised 3d object detection from rgb-d data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1931–1940 (2019) 10, 12
63. Tao, R., Han, W., Qiu, Z., Xu, C.z., Shen, J.: Weakly supervised monocular 3d object detection using multi-view projection and direction consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17482–17492 (2023) 2, 4, 9, 11
64. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023) 2, 7
65. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019) 6
66. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 2, 7
67. Wang, T., Xinge, Z., Pang, J., Lin, D.: Probabilistic and geometric depth: Detecting objects in perspective. In: Conference on Robot Learning. pp. 1475–1485. PMLR (2022) 9, 10
68. Wang, Y., Chen, Y., ZHANG, Z.X.: 4d unsupervised object discovery. Advances in Neural Information Processing Systems **35**, 35563–35575 (2022) 5
69. Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14749–14759 (2024) 1
70. Wei, Y., Su, S., Lu, J., Zhou, J.: Fgr: Frustum-aware geometric reasoning for weakly supervised 3d vehicle detection. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 4348–4354. IEEE (2021) 2, 4, 5, 6, 7, 8
71. Xie, Q., Lai, Y.K., Wu, J., Wang, Z., Zhang, Y., Xu, K., Wang, J.: Mlcvnet: Multi-level context votenet for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10447–10456 (2020) 4
72. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018) 10
73. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11784–11793 (2021) 3, 6, 10
74. Zakharov, S., Kehl, W., Bhargava, A., Gaidon, A.: Autolabeling 3d objects with differentiable rendering of sdf shape priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12224–12233 (2020) 2, 4, 5, 7, 9, 11
75. Zakharov, S., Kehl, W., Bhargava, A., Gaidon, A.: Autolabeling 3d objects with differentiable rendering of sdf shape priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12224–12233 (2020) 2
76. Zhang, G., Fan, L., He, C., Lei, Z., Zhang, Z., Zhang, L.: Voxel mamba: Group-free state space models for point cloud based 3d object detection. arXiv preprint arXiv:2406.10700 (2024) 3
77. Zhang, R., Qiu, H., Wang, T., Guo, Z., Cui, Z., Qiao, Y., Li, H., Gao, P.: Monodetr: Depth-guided transformer for monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9155–9166 (2023) 9, 10

78. Zhang, Z., Sun, B., Yang, H., Huang, Q.: H3dnet: 3d object detection using hybrid geometric primitives. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. pp. 311–329. Springer (2020) 4, 10
79. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. Advances in neural information processing systems **27** (2014) 3, 9
80. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4490–4499 (2018) 3, 10