

FCAF3D: Fully Convolutional Anchor-Free 3D Object Detection

Danila Rukhovich, Anna Vorontsova, and Anton Konushin

Samsung AI Center, Moscow

{d.rukhovich, a.vorontsova, a.konushin}@samsung.com

Abstract. Recently, promising applications in robotics and augmented reality have attracted considerable attention to 3D object detection from point clouds. In this paper, we present FCAF3D — a first-in-class fully convolutional anchor-free indoor 3D object detection method. It is a simple yet effective method that uses a voxel representation of a point cloud and processes voxels with sparse convolutions. FCAF3D can handle large-scale scenes with minimal runtime through a single fully convolutional feed-forward pass. Existing 3D object detection methods make prior assumptions on the geometry of objects, and we argue that it limits their generalization ability. To eliminate prior assumptions, we propose a novel parametrization of oriented bounding boxes that allows obtaining better results in a purely data-driven way. The proposed method achieves state-of-the-art 3D object detection results in terms of mAP@0.5 on ScanNet V2 (+4.5), SUN RGB-D (+3.5), and S3DIS (+20.5) datasets. The code and models are available at <https://github.com/samsunglabs/fcaf3d>.

Keywords: 3D object detection, anchor-free object detection, sparse convolutional networks

1 Introduction

3D object detection from point clouds aims at simultaneous localization and recognition of 3D objects given a 3D point set. As a core technique for 3D scene understanding, it is widely applied in autonomous driving, robotics, and AR.

While 2D methods ([26], [32]) work with dense fixed-size arrays, 3D methods are challenged by irregular unstructured 3D data of arbitrary volume. Consequently, the 2D data processing techniques are not directly applicable for 3D object detection, so 3D object detection methods ([10], [22], [19]) employ inventive approaches to 3D data processing.

Convolutional 3D object detection methods have scalability issues: large-scale scenes either require an impractical amount of computational resources or take too much time to process. Other methods opt for voxel data representation and employ sparse convolutions; however, these methods solve scalability problems at the cost of detection accuracy. In other words, there is no 3D object detection method that provides precise estimates *and* scales well.

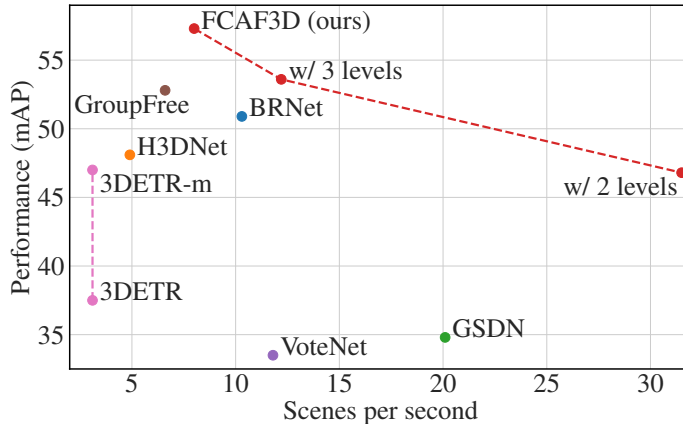


Fig. 1. mAP@0.5 scores on ScanNet against scenes per second. FCAF3D modifications (marked red) have different number of backbone feature levels. For each existing method, there is a FCAF3D modification surpassing this method in both detection accuracy and inference speed.

Besides being scalable and accurate, an ideal 3D object detection method should handle objects of arbitrary shapes and sizes without additional hacks and hand-tuned hyperparameters. We argue that prior assumptions on 3D object bounding boxes (e.g. aspect ratios or absolute sizes) restrict generalization and increase the number of hyperparameters and trainable parameters.

On the contrary, we do not want to rely on prior assumptions. We propose an anchor-free method that does not impose priors on objects and addresses 3D object detection with a purely data-driven approach. Moreover, we introduce a novel oriented bounding box (OBB) parametrization inspired by a Mobius strip that reduces the number of hyperparameters. To prove the effectiveness of our parametrization, we conduct experiments on SUN RGB-D with several 3D object detection methods and report improved results for all these methods.

In this paper, we present FCAF3D — a simple, effective, and scalable method for detecting 3D objects from point clouds. We evaluate the proposed method on ScanNet [7], SUN RGB-D [25], and S3DIS [1], demonstrating the solid superiority over the previous state-of-the-art on all benchmarks. On SUN RGB-D and ScanNet, our method surpasses other methods by at least 3.5% mAP@0.5. On S3DIS, FCAF3D outperforms the competitors by a huge margin.

Overall, our contribution is three-fold:

1. To our knowledge, we propose a first-in-class fully convolutional anchor-free 3D object detection method (FCAF3D) for indoor scenes.
2. We present a novel OBB parametrization and prove it to boost the accuracy of several existing 3D object detection methods on SUN RGB-D.

3. Our method significantly outperforms the previous state-of-the-art on challenging large-scale indoor ScanNet, SUN RGB-D, and S3DIS datasets in terms of mAP while being faster on inference.

2 Related Work

Recent 3D object detection methods are designed to be either indoor or outdoor. Indoor and outdoor methods have been developing almost independently, applying domain-specific data processing techniques. Many modern outdoor methods [30], [13], [35] project 3D points onto a bird-eye-view plane, thus reducing the task of 3D object detection to 2D object detection. Naturally, these methods take advantage of the fast-evolving algorithms for 2D object detection. Given a bird-eye-view projection, [14] processes it in a fully convolutional manner, while [31] exploits 2D anchor-free approach. Unfortunately, the approaches that proved to be effective for both 2D object detection and 3D outdoor object detection cannot be trivially adapted to indoor, as it would require an impracticable amount of memory and computing resources. To address performance issues, different 3D data processing strategies have been proposed. Currently, three approaches dominate the field of 3D object detection - voting-based, transformer-based, and 3D convolutional. Below we discuss each of these approaches in detail; we also provide a brief overview of anchor-free methods.

Voting-based methods. VoteNet [22] was the first method that introduced points voting for 3D object detection. VoteNet processes 3D points with PointNet [23], assigns a group of points to each object candidate according to their voted center, and computes object features from each point group. Among the numerous successors of VoteNet, the major progress is associated with advanced grouping and voting strategies applied to the PointNet features. BRNet [4] refines voting results with the representative points from the vote centers, which improves capturing the fine local structural features. MLCVNet [29] introduces three context modules into the voting and classifying stages of VoteNet to encode contextual information at different levels. H3DNet [33] improves the point group generation procedure by predicting a hybrid set of geometric primitives. VENet [28] incorporates an attention mechanism and introduces a vote weighting module trained via a novel vote attraction loss.

All VoteNet-like voting-based methods are limited by design. First, they show poor scalability: as their performance depends on the amount of input data, they tend to slow down if given larger scenes. Moreover, many voting-based methods implement voting and grouping strategies as custom layers, making it difficult to reproduce or debug these methods or port them to mobile devices.

Transformer-based methods. The recently emerged transformer-based methods use end-to-end learning and forward pass on inference instead of heuristics and optimization, which makes them less domain-specific. GroupFree [16] replaces VoteNet head with a transformer module, updating object query locations iteratively and ensembling intermediate detection results. 3DETR [19] was the first method of 3D object detection implemented as an end-to-end trainable

transformer. However, more advanced transformer-based methods still experience scalability issues similar to early voting-based methods. Differently, our method is fully-convolutional, thus being faster and significantly easier to implement than both voting-based and transformer-based methods.

3D convolutional methods. Voxel representation allows handling cubically growing sparse 3D data efficiently. Voxel-based 3D object detection methods ([12], [18]) convert points into voxels and process them with 3D convolutional networks. However, dense volumetric features still consume much memory, and 3D convolutions are computationally expensive. Overall, processing large scenes requires a lot of resources and cannot be done within a single pass.

GSDN [10] tackles performance issues with sparse 3D convolutions. It has encoder-decoder architecture, with both encoder and decoder parts built from sparse 3D convolutional blocks. Compared to the standard convolutional voting-based and transformer-based approaches, GSDN is significantly more memory-efficient and scales to large scenes without sacrificing point density. The major weakness of GSDN is its accuracy: this method is comparable to VoteNet in terms of quality, being significantly inferior to the current state-of-the-art [16].

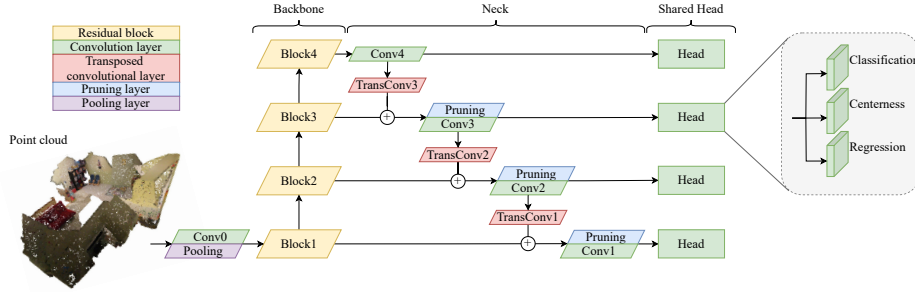


Fig. 2. The general scheme of the proposed FCAF3D. All convolutions and transposed convolutions are three-dimensional and sparse. This design allows processing the input point cloud in a single forward pass.

GSDN uses 15 aspect ratios for 3D object bounding boxes as anchors. If GSDN is trained in an anchor-free setting with a single aspect ratio, the accuracy decreases by 12%. Unlike GSDN, our method is anchor-free while taking advantage of sparse 3D convolutions.

RGB-based anchor-free object detection. In 2D object detection, anchor-free methods are solid competitors for the standard anchor-based methods. FCOS [26] addresses 2D object detection in a per-pixel prediction manner and shows a robust improvement over its anchor-based predecessor RetinaNet [15]. FCOS3D [27] trivially adapts FCOS by adding extra targets for monocular 3D object detection. ImVoxelNet [24] solves the same problem with an FCOS-like head built from standard (non-sparse) 3D convolutional blocks. We adapt the ideas from mentioned anchor-free methods to process sparse irregular data.

3 Proposed Method

Following the standard 3D detection problem statement, FCAF3D accepts N_{pts} RGB-colored points and outputs a set of 3D object bounding boxes. The FCAF3D architecture consists of a backbone, a neck, and a head (depicted in Fig. 2).

While designing FCAF3D, we aim for scalability, so we opt for a GSDN-like sparse convolutional network. For better generalization, we reduce the number of hyperparameters in this network that need to be manually tuned; specifically, we simplify sparsity pruning in the neck. Furthermore, we introduce an anchor-free head with a simple multi-level location assignment. Finally, we discuss the limitations of existing 3D bounding box parametrizations and propose a novel parametrization that improves both accuracy and generalization ability.

3.1 Sparse Neural Network

Backbone. The backbone in FCAF3D is a sparse modification of ResNet [11] where all 2D convolutions are replaced with sparse 3D convolutions. The family of sparse high-dimensional versions of ResNet was first introduced in [5]; for brevity, we refer to them as to HDResNet.

Neck. Our neck is a simplified GSDN decoder. Features on each level are processed with one sparse transposed 3D convolution and one sparse 3D convolution. Each transposed sparse 3D convolution with a kernel size of 2 might increase the number of non-zero values by 2^3 times. To prevent rapid memory growth, GSDN uses the *pruning* layer that filters input with a probability mask.

In GSDN, feature level-wise probabilities are calculated with an additional convolutional scoring layer. This layer is trained with a special loss encouraging consistency between the predicted sparsity and anchors. Specifically, voxel sparsity is set to be positive if any of the subsequent anchors associated with the current voxel is positive. However, using this loss may be suboptimal, as distant voxels of an object might get assigned with a low probability.

For simplicity, we remove the scoring layer with the corresponding loss and use probabilities from the classification layer in the head instead. We do not tune the probability threshold but keep at most N_{vox} voxels to control the sparsity level, where N_{vox} equals the number of input points N_{pts} . We claim this to be a simple yet elegant way to prevent sparsity growth since reusing the same hyperparameter makes the process more transparent and consistent.

Head. The anchor-free FCAF3D head consists of three parallel sparse convolutional layers with weights shared across feature levels. For each location $(\hat{x}, \hat{y}, \hat{z})$, these layers output classification probabilities \hat{p} , bounding box regression parameters $\hat{\delta}$, and centerness \hat{c} , respectively. This design is similar to the simple and light-weight head of FCOS [26] but adapted to 3D data.

Multi-level location assignment. During training, FCAF3D outputs locations $\{\mathbf{r}_i = (\hat{x}_i, \hat{y}_i, \hat{z}_i, k_i)\}$. For a given set of boxes $\{\mathbf{b}_j\}$, we define a function $\mu(\mathbf{r})$ that matches a location \mathbf{r} with a corresponding box \mathbf{b} (if any). Here, $k \in \{1, 2, 3, 4\}$ is the feature level index. The assignment strategy used in FCOS [26] and ImVoxelNet [24] is suboptimal, and its alterations are widely explored

in 2D object detection [32], [9]. We propose a simplified strategy for sparse data that does not require tuning dataset-specific hyperparameters.

For a box \mathbf{b} , let $R_k(\mathbf{b})$ denote the locations inside \mathbf{b} at the k -th feature level. We consider the levels with more than N_{loc} locations:

$$\mathcal{K}(\mathbf{b}) := \{k : |R_k(\mathbf{b})| > N_{\text{loc}}\}.$$

We select the largest index $k(\mathbf{b})$ from $\mathcal{K}(\mathbf{b})$ if $\mathcal{K}(\mathbf{b}) \neq \emptyset$ and set $k(\mathbf{b})$ to 1 otherwise, obtaining a correspondence between a box \mathbf{b} and a location $\mathbf{r} \in R_{k(\mathbf{b})}(\mathbf{b})$. Finally, for each location \mathbf{r} , we select the corresponding box with the least volume:

$$\mu(\mathbf{r}) = \arg \min_{\mathbf{b}: \mathbf{r} \in R_{k(\mathbf{b})}(\mathbf{b})} V(\mathbf{b}).$$

Finally, we filter locations via *center sampling* [26], considering only the points near the bounding box center as positive matches (Sec. 5.4).

Through assignment, some locations $\{(\hat{x}, \hat{y}, \hat{z})\}$ are matched with ground truth bounding boxes $\mathbf{b}_{\hat{x}, \hat{y}, \hat{z}}$. Accordingly, these locations get associated with ground truth labels $p_{\hat{x}, \hat{y}, \hat{z}}$ and 3D centerness values $c_{\hat{x}, \hat{y}, \hat{z}}$. During inference, the scores $\hat{\mathbf{p}}$ are multiplied by 3D centerness \hat{c} just before NMS as proposed in [24].

Loss function. The overall loss function is formulated as follows:

$$L = \frac{1}{N_{\text{pos}}} \sum_{\hat{x}, \hat{y}, \hat{z}} (L_{\text{cls}}(\hat{\mathbf{p}}, p) + \mathbb{1}_{\{p_{\hat{x}, \hat{y}, \hat{z}} \neq 0\}} L_{\text{reg}}(\hat{\mathbf{b}}, \mathbf{b}) + \mathbb{1}_{\{p_{\hat{x}, \hat{y}, \hat{z}} \neq 0\}} L_{\text{cntr}}(\hat{c}, c)). \quad (1)$$

Here, the number of matched locations N_{pos} is $\sum_{\hat{x}, \hat{y}, \hat{z}} \mathbb{1}_{\{p_{\hat{x}, \hat{y}, \hat{z}} \neq 0\}}$. Classification loss L_{cls} is a focal loss, regression loss L_{reg} is IoU, and centerness loss L_{cntr} is binary cross-entropy. For each loss, predicted values are denoted with a hat.

3.2 Bounding Box Parametrization

The 3D object bounding boxes can be axis-aligned (AABB) or oriented (OBB). An AABB can be described as $\mathbf{b}^{\text{AABB}} = (x, y, z, w, l, h)$, while the definition of an OBB includes a *heading angle* θ : $\mathbf{b}^{\text{OBB}} = (x, y, z, w, l, h, \theta)$. In both formulas, x, y, z denote the coordinates of the center of a bounding box, while w, l, h are its width, length, and height, respectively.

AABB parametrization. For AABBs, we follow the parametrization proposed in [24]. Specifically, for a ground truth AABB (x, y, z, w, l, h) and a location $(\hat{x}, \hat{y}, \hat{z})$, $\boldsymbol{\delta}$ can be formulated as a 6-tuple:

$$\begin{aligned} \delta_1 &= x + \frac{w}{2} - \hat{x}, \quad \delta_2 = \hat{x} - x + \frac{w}{2}, \quad \delta_3 = y + \frac{l}{2} - \hat{y}, \\ \delta_4 &= \hat{y} - y + \frac{l}{2}, \quad \delta_5 = z + \frac{h}{2} - \hat{z}, \quad \delta_6 = \hat{z} - z + \frac{h}{2}. \end{aligned} \quad (2)$$

The predicted AABB $\hat{\mathbf{b}}$ can be trivially obtained from $\boldsymbol{\delta}$.

Heading angle estimation. All state-of-the-art 3D object detection methods from point clouds address the heading angle estimation task as classification

followed by regression. The heading angle is classified into bins; then, the precise heading angle is regressed within a bin. For indoor scenes, the range from 0 to 2π is typically divided into 12 equal bins [22], [21], [33], [19]. For outdoor scenes, there are usually only two bins [30], [13], as the objects on the road can be either parallel or perpendicular to the road.

When a heading angle bin is chosen, the heading angle value is estimated through regression. VoteNet and other voting-based methods estimate the value of θ directly. Outdoor methods explore more elaborate approaches, e.g. predicting the values of trigonometric functions. For instance, SMOKE [17] estimates $\sin\theta$ and $\cos\theta$ and uses the predicted values to recover the heading angle.

Fig. 3 depicts indoor objects where the heading angle is unambiguous. Accordingly, ground truth angle annotations can be chosen randomly for these objects, making heading angle bin classification meaningless. To avoid penalizing the correct predictions that do not coincide with annotations, we use rotated IoU loss, as its value is the same for all possible choices of heading angle.



Fig. 3. Examples of objects with an ambiguous heading angle.

Thus, we propose OBB parametrization that considers the rotation ambiguity.

Proposed Mobius OBB parametrization. Considering the OBB with parameters $(x, y, z, w, l, h, \theta)$, let us denote $q = \frac{w}{l}$. If $x, y, z, w + l, h$ are fixed, it turns out that the OBBs with

$$(q, \theta), \left(\frac{1}{q}, \theta + \frac{\pi}{2}\right), (q, \theta + \pi), \left(\frac{1}{q}, \theta + \frac{3\pi}{2}\right) \quad (3)$$

define the same bounding box. We notice that the set of (q, θ) , where $\theta \in (0, 2\pi]$, $q \in (0, +\infty)$ is topologically equivalent to a Mobius strip [20] up to this equivalence relation. Hence, we can reformulate the task of estimating (q, θ) as a task of predicting a point on a Mobius strip. A natural way to embed a Mobius strip being a two-dimensional manifold to Euclidean space is the following:

$$(q, \theta) \mapsto (\ln(q) \sin(2\theta), \ln(q) \cos(2\theta), \sin(4\theta), \cos(4\theta)). \quad (4)$$

It is easy to verify that 4 points from Eq. 3 are mapped into a single point in Euclidean space (see Supplementary for details). However, the experiments reveal that predicting only $\ln(q) \sin(2\theta)$ and $\ln(q) \cos(2\theta)$ provides better results than predicting all four values. Thereby, we opt for a *pseudo* embedding of a Mobius strip to \mathbb{R}^2 . We call it *pseudo* since it maps the entire center circle of a Mobius strip defined by $\ln(q) = 0$ to $(0, 0)$. Accordingly, we cannot distinguish points with $\ln q = 0$. However, $\ln(q) = 0$ implies strict equality of w and l , which is rare in real-world scenarios. Moreover, the choice of an angle has a minor effect on the IoU if $w = l$; thereby, we ignore this rare case for the sake

of detection accuracy and simplicity of the method. Overall, we obtain a novel OBB parametrization:

$$\delta_7 = \ln \frac{w}{l} \sin(2\theta), \quad \delta_8 = \ln \frac{w}{l} \cos(2\theta). \quad (5)$$

In the standard parametrization 2, $\hat{\mathbf{b}}$ is trivially derived from δ . In the proposed parametrization, w, l, θ are non-trivial and can be obtained as follows:

$$w = \frac{sq}{1+q}, \quad l = \frac{s}{1+q}, \quad \theta = \frac{1}{2} \arctan \frac{\delta_7}{\delta_8}, \quad (6)$$

where ratio $q = e^{\sqrt{\delta_7^2 + \delta_8^2}}$ and size $s = \delta_1 + \delta_2 + \delta_3 + \delta_4$.

4 Experiments

4.1 Datasets

We evaluate our method on three 3D object detection benchmarks: ScanNet V2 [7], SUN RGB-D [25], and S3DIS [1]. For all datasets, we use mean average precision (mAP) under IoU thresholds of 0.25 and 0.5 as a metric.

ScanNet. The ScanNet dataset contains 1513 reconstructed 3D indoor scans with per-point instance and semantic labels of 18 object categories. Given this annotation, we calculate AABBs via the standard approach [22]. The training subset is comprised of 1201 scans, while 312 scans are left for validation.

SUN RGB-D. SUN RGB-D is a monocular 3D scene understanding dataset containing more than 10,000 indoor RGB-D images. The annotation consists of per-point semantic labels and OBBs of 37 object categories. As proposed in [22], we run experiments with objects of the 10 most common categories. The training and validation splits contain 5285 and 5050 point clouds, respectively.

S3DIS. Stanford Large-Scale 3D Indoor Spaces dataset contains 3D scans of 272 rooms from 6 buildings, with 3D instance and semantic annotation. Following [10], we evaluate our method on furniture categories. AABBs are derived from 3D semantics. We use the official split, where 68 rooms from *Area 5* are intended for validation, while the remaining 204 rooms comprise the training subset.

4.2 Implementation Details

Hyperparameters. For all datasets, we use the same hyperparameters except for the following. First, the size of output classification layer equals the number of object categories, which is 18, 10, and 5 for ScanNet, SUN RGB-D, and S3DIS. Second, SUN RGB-D contains OBBs, so we predict additional targets δ_7 and δ_8 for this dataset; note that the loss function is not affected. Last, ScanNet, SUN RGB-D, and S3DIS contain different numbers of scenes, so we repeat each scene 10, 3, and 13 times per epoch, respectively.

Similar to GSDN [10], we use the sparse 3D modification of ResNet34 named HDResNet34 as a backbone. The neck and the head use the outputs of the

backbone at all feature levels. In initial point cloud voxelization, we set the voxel size to 0.01m and the number of points N_{pts} to 100,000. Respectively, N_{vox} equals to 100,000. Both ATSS [32] and FCOS [26] set N_{loc} to 3^2 for 2D object detection. Accordingly, we select a feature level so bounding box covers at least $N_{loc} = 3^3$ locations. We select 18 locations by center sampling. The NMS IoU threshold is 0.5.

Training. We implement FCAF3D using the MMDetection3D [6] framework. The training procedure follows the default MMDetection [3] scheme: training takes 12 epochs and the learning rate decreases on the 8th and the 11th epochs. We employ the Adam optimizer with an initial learning rate of 0.001 and weight decay of 0.0001. All models are trained on two NVidia V100 with a batch size of 8. Evaluation and performance tests are run on a single NVidia GTX1080Ti.

Method	Presented at	ScanNet		SUN RGB-D		S3DIS	
		mAP@0.25	mAP@0.5	mAP@0.25	mAP@0.5	mAP@0.25	mAP@0.5
VoteNet[22]	ICCV'19	58.6	33.5	57.7	-	-	-
3D-MPA[8]	CVPR'20	64.2	49.2	-	-	-	-
HGNet[2]	CVPR'20	61.3	34.4	61.6	-	-	-
MLCVNet[29]	CVPR'20	64.5	41.4	59.8	-	-	-
GSDN[10]	ECCV'20	62.8	34.8	-	-	47.8	25.1
H3DNet[33]	ECCV'20	67.2	48.1	60.1	39.0	-	-
BRNet[4]	CVPR'21	66.1	50.9	61.1	43.7	-	-
3DETR[19]	ICCV'21	65.0	47.0	59.1	32.7	-	-
VENet[28]	ICCV'21	67.7	-	62.5	39.2	-	-
GroupFree [16]	ICCV'21	69.1 (68.6)	52.8 (51.8)	63.0 (62.6)	45.2 (44.4)	-	-
FCAF3D	-	71.5 (70.7)	57.3 (56.0)	64.2 (63.8)	48.9 (48.2)	66.7 (64.9)	45.9 (43.8)

Table 1. Results of FCAF3D and existing indoor 3D object detection methods that accept point clouds. The best metric values are marked bold. FCAF3D outperforms previous state-of-the-art methods: GroupFree (on ScanNet and SUN RGB-D) and GSDN (on S3DIS). The reported metric value is the best one across 25 trials; the average value is given in brackets.

Evaluation. We follow the evaluation protocol introduced in [16]. Both training and evaluation are randomized, as the input N_{pts} are randomly sampled from the point cloud. To obtain statistically significant results, we run training 5 times and test each trained model 5 times independently. We report both the best and average metrics across 5×5 trials: this allows comparing FCAF3D to the 3D object detection methods that report either a single best or an average value.

5 Results

5.1 Comparison with State-of-the-art Methods

We compare FCAF3D with previous state-of-the-arts on three indoor benchmarks in Tab. 1. As one might observe, FCAF3D achieves the best results on all benchmarks. The performance gap is especially tangible in terms of mAP@0.5: our method surpasses previous state-of-the-art by 4.5% on ScanNet and 3.7% on

SUN RGB-D. On S3DIS, FCAF3D outperforms weak state-of-the-art by a huge margin. Overall, the proposed method is consistently better than existing methods, setting a new state-of-the-art for indoor 3D object detection. The examples of ScanNet, SUN RGB-D, and S3DIS point clouds with predicted bounding boxes are depicted in Fig. 4, 5, 6.

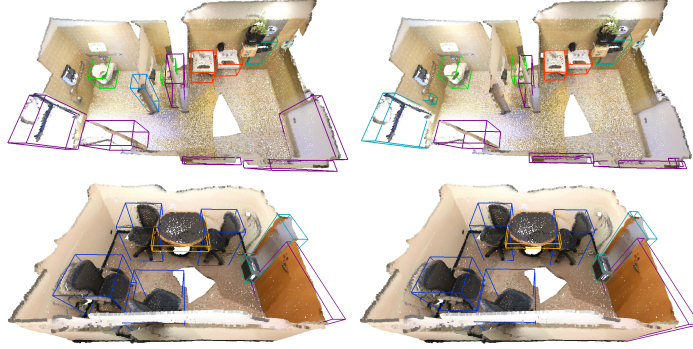


Fig. 4. The point cloud from ScanNet with AABBs. The color of a bounding box denotes the object category. Left: estimated with FCAF3D, right: ground truth.

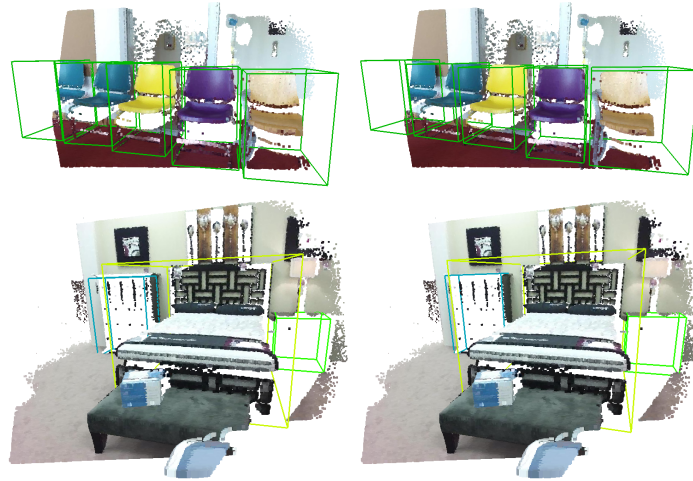


Fig. 5. The point cloud from SUN RGB-D with OBBs. The color of a bounding box denotes the object category. Left: estimated with FCAF3D, right: ground truth.



Fig. 6. The point cloud from S3DIS with AABBs. The color of a bounding box denotes the object category. Left: estimated with FCAF3D, right: ground truth.

5.2 Object Geometry Priors

To study geometry priors, we train and evaluate existing methods with proposed modifications. We experiment with 3D object detection methods accepting data of different modalities: point clouds, RGB images, or both, to see whether the effect is data-specific or universal. VoteNet and ImVoteNet have the same head and are trained with the same losses. Among them, there are 4 prior losses: size classification loss, size regression loss, direction classification loss, and direction regression loss. Both classification losses correspond to targets parametrized using priors (per-category mean object sizes and a set of angle bins). Similar to FCAF3D, we replace the aforementioned losses with a rotated IoU loss with Mobius parametrization 5. To give a complete picture, we also try a *sin-cos* parametrization used in the outdoor 3D object detection method SMOKE [17].

The rotated IoU loss decreases the number of trainable parameters and hyperparameters, including geometry priors and loss weights. This loss has already been used in outdoor 3D object detection [34]. Recently, [6] reported results of VoteNet trained with axis-aligned IoU loss on ScanNet.

Tab. 2 shows that replacing the standard parametrization with Mobius one boosts VoteNet and ImVoteNet mAP@0.5 by approximately 4%. ImVoxelNet does not use a classification+regression scheme to estimate heading angle but predicts its value directly in a single step. Since the original ImVoxelNet uses the rotated IoU loss, we do not need to remove redundant losses, only to change the parametrization. Again, the Mobius parametrization helps to obtain the best results, even though the superiority is minor.

5.3 Comparison to GSDN

For a fair comparison, we test our method against GSDN in the most similar scenario. We use a voxel size of 0.05m, ensuring we do not benefit from using

Method	Input	mAP@0.25	mAP@0.5
VoteNet[22]	PC	57.7	-
Reimpl.[6]		59.1	35.8
Reimpl. w/ IoU loss			
w/ naive param.		61.1 (60.3)	38.4 (37.7)
w/ <i>sin-cos</i> param.		60.7 (59.8)	37.1 (36.4)
w/ Mobius param.		61.1 (60.5)	40.4 (39.5)
ImVoteNet[21]	RGB +PC	63.4	-
Reimpl.[6]		64.0	37.8
Reimpl. w/ IoU loss			
w/ naive param.		64.2 (63.9)	39.1 (38.3)
w/ <i>sin-cos</i> param.		64.6 (64.0)	39.9 (37.8)
w/ Mobius param.		64.6 (64.1)	40.8 (39.8)
ImVoxelNet[24]	RGB	40.7	-
w/ naive param.		41.3 (40.4)	13.8 (13.0)
w/ <i>sin-cos</i> param.		41.3 (40.5)	13.2 (12.8)
w/ Mobius param.		41.5 (40.6)	14.6 (14.0)
FCAF3D	PC		
w/ naive param.		63.8 (63.5)	46.8 (46.2)
w/ <i>sin-cos</i> param.		63.9 (63.6)	48.2 (47.3)
w/ Mobius param.		64.2 (63.8)	48.9 (48.2)

Table 2. Results of several 3D object detection methods that accept inputs of different modalities, with different OBB parametrization on SUN RGB-D. The FCAF3D metric value is the best across 25 trials; the average value is given in brackets. For other methods, we report results from the original papers and also the results obtained through our experiments with MMDetection3D-based re-implementations (marked as Reimpl). PC stands for point cloud.

more detailed and informative spatial information; the voxel sizes at different feature levels are also of the same sizes (0.2, 0.4, 0.8, 1.6).

As shown in Tab. 3, the best results are obtained by the original *accurate* FCAF3D with the HDResNet34 backbone, the voxel size of 0.01m, and the default stride of the first convolution of 2: in this setting, FCAF3D outperforms GSDN by a huge margin (mAP@0.25 of 70.7 against 62.8, mAP@0.5 of 56.0 against 34.8). With the same stride of 1, the same voxel size and the same voxel sizes at feature levels, FCAF3D slightly outperforms GSDN in mAP@0.25 (64.2 against 62.8), having a notable gain in mAP@0.5 (46.2 against 34.8).

The *balanced* FCAF3D with a lighter backbone with three feature levels, the voxel size of 0.05m, and the stride of 1 outperforms GSDN in *both accuracy and speed*. Overall, we argue that FCAF3D addresses the 3D object detection in a more efficient way and thus should be preferred.

Moreover, we address the speed issues with the most lightweight HDResNet34:2 backbone with two feature levels. The *fast* FCAF3D modification with HDResNet34:2 processes 30 scenes per second, while GSDN is able to handle only 20 scenes. While improving the inference speed, we maintain the superior

accuracy: with the voxel size of 0.02m, FCAF3D with a HDResNet34:2 backbone still outperforms GSDN in both mAP@0.25 and mAP@0.5.

Finally, it can be observed that mAP@0.5 drops dramatically by 12% if GSDN is trained in an anchor-free setting (which is equivalent to using one anchor). In other words, GSDN demonstrates a poor performance without domain-specific guidance in the form of anchors; hence, we claim this method to be inflexible and non-generalizable.

Method	Backbone	Voxel size [m]	Stride	Feature level voxel sizes [m]	Scenes per sec.	mAP	
						0.25	0.5
GSDN[10]		0.05	1		20.1	62.8	34.8
w/o anchors	HDResNet34	0.05	1	0.2,0.4,0.8,1.6	20.4	56.3	22.7
FCAF3D		0.05	1		17.0	64.2	46.2
FCAF3D (<i>accurate</i>)	HDResNet34	0.01	2	0.08,0.16,0.32,0.64	8.0	<u>70.7</u>	<u>56.0</u>
FCAF3D (<i>balanced</i>)	HDResNet34:3	0.05	1	0.2,0.4,0.8	22.9	62.9	43.9
FCAF3D (<i>fast</i>)	HDResNet34:2	0.02	2	0.16,0.32	31.5	63.1	46.8

Table 3. Results of fully convolutional 3D object detection methods that accept point clouds on ScanNet. The FCAF3D results better than the results of the original GSDN (with anchors) are marked bold. The all-best results are underlined.

5.4 Ablation Study

In this section, we discuss the FCAF3D design choices and investigate how they affect metrics when applied independently in ablation studies. We run experiments with varying voxel size, the number of points in a point cloud N_{pts} , the number of locations selected by center sampling, and with and without centerness. The results of ablation studies are aggregated in Tab. 4 for all benchmarks.

Voxel size. Expectedly, with an increasing voxel size, accuracy goes down. We try voxels of 0.03, 0.02, and 0.01 m. We do not experiment with smaller values since inference would take too much time. We attribute the notable gap in mAP between voxel sizes of 0.01 and 0.02 m to the presence of *almost flat* objects, such as doors, pictures, and whiteboards. Namely, with a voxel size of 2 cm, the head would output locations with 16 cm tolerance, but the *almost flat* objects could be less than 16 cm by one of the dimensions. Accordingly, we observe a decrease in accuracy for larger voxel sizes.

Number of points. Similar to 2D images, subsampled point clouds are sometimes referred to as *low-resolution* ones. Accordingly, they contain less information than their *high-resolution* versions. As can be expected, the fewer the points, the lower is detection accuracy. In this series of experiments, we sample 20k, 40k, and 100k points from the entire point cloud, and the obtained metric values revealed a clear dependency between the number of points and mAP. We do not consider larger N_{pts} values to be on a par with the existing methods (specifically, GSDN [10] uses all points in a point cloud, GroupFree [16] samples 50k points, VoteNet [22] selects 40k points for ScanNet and 20k for SUN

RGB-D). We use $N_{\text{vox}} = N_{\text{pts}}$ to guide pruning in the neck. When N_{vox} exceeds 100k, the inference time increases due to growing sparsity in the neck, while the accuracy improvement is negligible. So we restrict our grid search for N_{pts} with 100k and use it as a default value regarding the obtained results.

Ablating parameter	Value	ScanNet		SUN RGB-D		S3DIS	
		mAP@0.25	mAP@0.5	mAP@0.25	mAP@0.5	mAP@0.25	mAP@0.5
Voxel size	0.01	71.5 (70.7)	57.3 (56.0)	64.2 (63.8)	48.9 (48.2)	66.7 (64.9)	45.9 (43.8)
	0.02	66.3 (65.8)	49.4 (48.6)	62.3 (62.0)	46.3 (45.5)	61.0 (58.5)	43.8 (38.5)
	0.03	59.6 (59.2)	42.6 (41.6)	60.4 (59.7)	41.6 (41.0)	55.4 (53.3)	38.6 (35.0)
Number of points	20k	69.0 (68.1)	52.8 (52.0)	63.0 (62.5)	46.9 (46.5)	60.1 (58.8)	45.1 (40.1)
	40k	67.6 (66.7)	53.6 (52.2)	63.4 (63.1)	47.2 (46.6)	63.7 (61.2)	44.8 (42.2)
	100k	71.5 (70.7)	57.3 (56.0)	64.2 (63.8)	48.9 (48.2)	66.7 (64.9)	45.9 (43.8)
Centerness	No	71.0 (70.4)	56.1 (55.1)	63.8 (63.3)	48.2 (47.5)	67.9 (65.5)	46.0 (43.5)
	Yes	71.5 (70.7)	57.3 (56.0)	64.2 (63.8)	48.9 (48.2)	66.7 (64.9)	45.9 (43.8)
Center sampling	9	70.6 (70.1)	55.7 (55.0)	63.8 (63.3)	48.6 (48.2)	66.5 (63.6)	44.4 (42.5)
	18	71.5 (70.7)	57.3 (56.0)	64.2 (63.8)	48.9 (48.2)	66.7 (64.9)	45.9 (43.8)
	27	70.2 (69.7)	55.7 (54.1)	64.3 (63.8)	48.7 (47.9)	65.1 (63.2)	43.6 (41.7)

Table 4. Results of ablation studies on the voxel size, the number of points (which equals the number of voxels N_{vox} in pruning), centerness, and center sampling in FCAF3D. The better options are marked bold (actually, these are the default options used to obtain the results in Tab. 1 above). The reported metric value is the best across 25 trials; the average value is given in brackets.

Centerness. Using centerness improves mAP for the ScanNet and SUN RGB-D datasets. For S3DIS, the results are controversial: the better mAP@0.5 is balanced with a minor decrease of mAP@0.25. Nevertheless, we analyze the results altogether, so we can consider centerness a helpful feature with a small positive effect on the mAP, almost reaching 1% of mAP@0.5 on ScanNet.

Center sampling. Finally, we study the number of locations selected in center sampling. We select 9 locations, as proposed in FCOS [26], the entire set of 27 locations, as in ImVoxelNet [24], and 18 locations. The latter appeared to be the best choice according to mAP on all the benchmarks.

6 Conclusion

We presented FCAF3D, a first-in-class fully convolutional anchor-free 3D object detection method for indoor scenes. Our method significantly outperforms the previous state-of-the-art on the challenging indoor SUN RGB-D, ScanNet, and S3DIS benchmarks in terms of both mAP and inference speed. We also proposed a novel oriented bounding box parametrization and showed that it improves accuracy for several 3D object detection methods. Moreover, the proposed parametrization allows avoiding any prior assumptions about objects, thus reducing the number of hyperparameters. Overall, FCAF3D with our bounding box parametrization is accurate, scalable, and generalizable at the same time.

References

1. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1534–1543 (2016) [2](#), [8](#)
2. Chen, J., Lei, B., Song, Q., Ying, H., Chen, D.Z., Wu, J.: A hierarchical graph network for 3d object detection on point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 392–401 (2020) [9](#)
3. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019) [9](#)
4. 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) [3](#), [9](#)
5. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3075–3084 (2019) [5](#)
6. Contributors, M.: MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. <https://github.com/open-mmlab/mmdetection3d> (2020) [9](#), [11](#), [12](#)
7. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017) [2](#), [8](#)
8. Engelmann, F., Bokeloh, M., Fathi, A., Leibe, B., Nießner, M.: 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9031–9040 (2020) [9](#)
9. Ge, Z., Liu, S., Li, Z., Yoshie, O., Sun, J.: Ota: Optimal transport assignment for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 303–312 (2021) [6](#)
10. Gwak, J., Choy, C., Savarese, S.: Generative sparse detection networks for 3d single-shot object detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 297–313. Springer (2020) [1](#), [4](#), [8](#), [9](#), [13](#)
11. 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) [5](#)
12. 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](#)
13. 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](#), [7](#)
14. Li, B.: 3d fully convolutional network for vehicle detection in point cloud. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1513–1518. IEEE (2017) [3](#)

15. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) [4](#)
16. Liu, Z., Zhang, Z., Cao, Y., Hu, H., Tong, X.: Group-free 3d object detection via transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2949–2958 (2021) [3](#), [4](#), [9](#), [13](#)
17. Liu, Z., Wu, Z., Tóth, R.: Smoke: Single-stage monocular 3d object detection via keypoint estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 996–997 (2020) [7](#), [11](#)
18. Maturana, D., Scherer, S.: Voxnet: A 3d convolutional neural network for real-time object recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 922–928. IEEE (2015) [4](#)
19. 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) [1](#), [3](#), [7](#), [9](#)
20. Munkres, J.R.: Topology (2000) [7](#)
21. Qi, C.R., Chen, X., Litany, O., Guibas, L.J.: Imvotenet: Boosting 3d object detection in point clouds with image votes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4404–4413 (2020) [7](#), [12](#)
22. 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) [1](#), [3](#), [7](#), [8](#), [9](#), [12](#), [13](#)
23. 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](#)
24. Rukhovich, D., Vorontsova, A., Konushin, A.: Invoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. arXiv preprint arXiv:2106.01178 (2021) [4](#), [5](#), [6](#), [12](#), [14](#)
25. 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](#), [8](#)
26. 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) [1](#), [4](#), [5](#), [6](#), [9](#), [14](#)
27. Wang, T., Zhu, X., Pang, J., Lin, D.: Fcos3d: Fully convolutional one-stage monocular 3d object detection. arXiv preprint arXiv:2104.10956 (2021) [4](#)
28. Xie, Q., Lai, Y.K., Wu, J., Wang, Z., Lu, D., Wei, M., Wang, J.: Venet: Voting enhancement network for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3712–3721 (2021) [3](#), [9](#)
29. 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) [3](#), [9](#)
30. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. *Sensors* **18**(10), 3337 (2018) [3](#), [7](#)
31. 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](#)
32. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings

- of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9759–9768 (2020) [1](#), [6](#), [9](#)
33. Zhang, Z., Sun, B., Yang, H., Huang, Q.: H3dnet: 3d object detection using hybrid geometric primitives. In: European Conference on Computer Vision. pp. 311–329. Springer (2020) [3](#), [7](#), [9](#)
 34. Zhou, D., Fang, J., Song, X., Guan, C., Yin, J., Dai, Y., Yang, R.: Iou loss for 2d/3d object detection. In: 2019 International Conference on 3D Vision (3DV). pp. 85–94. IEEE (2019) [11](#)
 35. 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](#)