Rotationally Equivariant 3D Object Detection

Hong-Xing Yu
Stanford University

Jiajun Wu
Stanford University

Li Yi
Tsinghua University, Shanghai Qi Zhi Institute

Abstract

Rotation equivariance has recently become a strongly desired property in the 3D deep learning community. Yet most existing methods focus on equivariance regarding a global input rotation while ignoring the fact that rotation symmetry has its own spatial support. Specifically, we consider the object detection problem in 3D scenes, where an object bounding box should be equivariant regarding the object pose, independent of the scene motion. This suggests a new desired property we call object-level rotation equivariance. To incorporate object-level rotation equivariance into 3D object detectors, we need a mechanism to extract equivariant features with local object-level spatial support while being able to model cross-object context information. To this end, we propose Equivariant Object detection Network (EON) with a rotation equivariance suspension design to achieve object-level equivariance. EON can be applied to modern point cloud object detectors, such as VoteNet and PointR-CNN, enabling them to exploit object rotation symmetry in scene-scale inputs. Our experiments on both indoor scene and autonomous driving datasets show that significant improvements are obtained by plugging our EON design into existing state-of-the-art 3D object detectors. Project website: https://kovenyu.com/EON/.

1. Introduction

3D object detection is a fundamental problem in various downstream applications including augmented reality, robotics, and autonomous driving. Research efforts in designing 3D object detection networks have shown great effectiveness for both indoor [17, 27] and outdoor scenes [21, 29]. However, existing 3D object detectors cannot explicitly treat object rotation equivariance in their designs. Object rotation equivariance in 3D detection is well-reflected in the rotation invariance of shape and equivariance of orientation. For example, no matter how an object is oriented in an input scene, the detection result (typically represented as an oriented bounding box) associated with the object should orient in the same way while retaining the same shape. Explicit
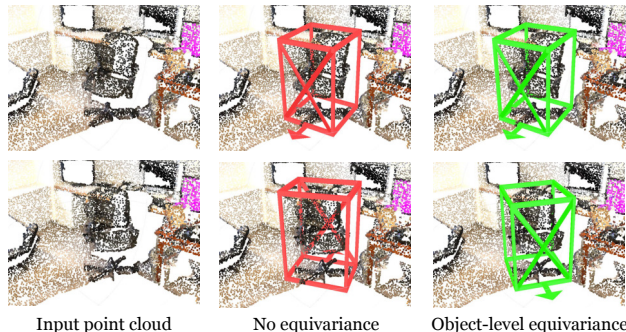

Figure 1. Rotation symmetry in object detection includes the equivariant orientation and invariant shape of the bounding box. The object-level rotation equivariant detector explicitly models these strong priors and intrinsically generates oriented bounding box to rotate following the object, while maintaining the same box shape. Non-equivariant detectors, however, may suffer from box shape changes and unaligned variations in box orientation.

modeling of these strong priors save the needs for expensive data augmentations, and can increase the expressivity and discriminative power of detection models without heavily increasing the number of parameters and introducing additional optimization challenges.

A recent trend to explicitly exploit rotation equivariance is through equivariant networks [4, 25, 26] (EN). The main idea is that the equivariant geometric features carry both shape information and orientation information separately by design. Rotation equivariant networks have been explored for object classification and pose estimation [2, 7, 13], but not yet for 3D object detection. A main challenge is that existing EN mostly explores rotation equivariance regarding the full visual input, while equivariance to rotations of a whole scene is not ideal for object detection, because individual object orientation can be independent of the scene. Thus, it is unclear how to achieve object-level rotation equivariance and how to benefit 3D object detection in cluttered scenes.

We identify two key technical challenges toward object-level equivariant model design: how to determine the object-level spatial support to extract locally equivariant features, and how to aggregate context information. The greater con-

text (such as nearby objects) is helpful to recognize objects especially in noisy or incomplete raw point clouds. However, the context information could easily break object-level rotation equivariance if not handled properly. For example, when detecting a chair, the presence of a nearby desk can provide useful context due to high co-occurrence probability. However, in case that the chair remains static while the desk changes its orientation, the chair features could also be affected unexpectedly.

We propose Equivariant Object detection Network (EON) to exploit object-level equivariance for 3D detection. Our core design is called **rotation equivariance suspension**. To properly determine the object-level spatial support, we let our model extract equivariant features only up to an intermediate stage. This is based upon the observation that most 3D detection networks extract features in a hierarchical manner [14, 17, 21] where early stages focus on local features while later stages cover more context-level information. Computing equivariant features only up to an intermediate stage produces local spatial support to rotation equivariance, and the model can adaptively learn to adjust its effective spatial support [15]. To allow aggregating context information, we *suspend* equivariant feature computation at the previous intermediate stage by decomposing each equivariant feature into an object orientation hypothesis (orientation information) and an invariant object-frame feature (shape information). Our model keeps aggregating object-frame features in the latter stages, and finally resumes the orientation information for object proposals. Since only object-frame features (i.e., without object orientation information) are aggregated after the intermediate stage, the greater contexts can be modeled without breaking object-level equivariance.

Our approach follows the modular design adopted by most bottom-up detectors, so that it can be easily plugged into state-of-the-art 3D object detectors. We have tested our method using various backbones and models on both indoor and outdoor 3D object detection benchmarks. We find that EON significantly boosts the performance of previous state-of-the-art 3D object detectors (+9.0 mAP on ScanNetV2, +3.1 mAP on SUN RGB-D, and +1.4 mAP on KITTI Dataset). In summary, our contributions are threefold:

- To our best knowledge, this is the first work to explore rotation equivariance for 3D object detection.

- We propose Equivariant Object detection Network (EON), incorporating a novel design dubbed rotation equivariance suspension to exploit object-level equivariance in 3D detection. Our EON can be easily plugged into state-of-the-art bottom-up 3D object detectors.

- On both indoor and outdoor datasets, we demonstrate the benefits of object-level equivariance by boosting performances of previous state-of-the-art 3D detectors.

2. Related Work

3D object detection. Most current state-of-the-art approaches to 3D object detection directly takes 3D data (such as point cloud, Lidar, and voxelized grids from them) as input and generates 3D oriented bounding boxes (OBBs) to represent the objects [11, 16, 17, 20–22, 29–32]. Most of them follow a bottom-up design, where a backbone network extracts a sparse set of regional features from the dense input data, and a detection head proposes candidate OBBs (one-stage) or regions of interests (two-stage) for further refinement. Seminal works include VoteNet [17] and its follow-ups. VoteNet's backbone design includes two PointNet++, with a voting in between to help the aggregation of object surface points. H3DNet [30] improves VoteNet predictions using 3D primitives and a geometric loss. MLCVNet [27] further allows VoteNet to aggregate global contextual information via self-attention. Another line of methods is directly inspired from 2D bottom-up detectors, such as PointRCNN [21], Voxel-Net [31], PointPillar [11] and CenterPoint [29]. However, these existing methods do not explicitly exploit object rotation equivariance in their models. Our method is grounded on this popular bottom-up modular design, and we focus on equipping these state-of-the-art detection models with object-level equivariance.

Rotation equivariance networks. From the seminal work of Group Equivariant Convolutional Network [4], leveraging group equivariance for deep networks becomes increasingly popular. Existing approaches toward rotation equivariance can be roughly divided into two categories: filter orbit-based and filter design-based. Deriving from the group equivariant convolution [3, 4], filter orbit-based methods discretize the rotation group and construct a set of group-transformed kernels (an "orbit") for the group equivariant computation [2, 12, 13]. Filter design-based methods design intrinsically rotation-equivariant basis functions (e.g., generalized Fourier basis functions [5, 25]) and compose their networks with these basis functions [5, 7, 8, 25]. Our method is inspired from the filter orbit-based Equivariant Point Network [2] (EPN) which introduces a tractable approximation to SE(3) group equivariant convolution on point clouds. However, EPN focuses on single object tasks and achieves equivariance regarding the full visual content. We target object-level equivariance for 3D detection in scenes. Recently, there are a few works attempting to leverage rotation equivariance for aerial image detection [10, 28]. While they focus on 2D images, we aim at 3D object-level equivariant detection.

3. Preliminary

To compute rotation equivariant features, our approach finds inspiration from a recent state-of-the-art equivariant network, Equivariant Point Network (EPN) [2], that is designed for single objects. We briefly review EPN and provide

an intuitive explanation for the equivariance property.

The key idea of EPN to achieve rotation equivariance is to augment each point feature vector $\mathbf{f} \in \mathbb{R}^C$ to an equivariant feature $\mathbf{f}_{\text{eqv}} \in \mathbb{R}^{C \times |G|}$ w.r.t. a discretized SO(3) subgroup G. Each vector element in the equivariant feature, $\mathbf{f}_{\text{eqv}}(g) \in \mathbb{R}^C$, is computed via rotating $\mathbf{f}$ by $g^{-1} \in G$ before passing it into a computational layer. One can verify that this rotate-and-compute operation indeed generates $\mathbf{f}_{\text{eqv}}$ that is equivariant to any $g \in G$. Intuitively this means that, if an input is rotated by some $g_0 \in G$, its equivariant feature $\mathbf{f}_{\text{eqv}}$ will undergo a *circular shift* without changing any values, becoming $\mathbf{f}'_{\text{eqv}}$ such that $\mathbf{f}'_{\text{eqv}}(g') = \mathbf{f}_{\text{eqv}}(g)$, where g' and g satisfy $g' = g_0 \circ g$. In other words, rotating the input $x \in \mathbb{R}^3$ by g_0 leads to a "rotation" by g_0 defined on the equivariant feature (in this case, the "rotation" is defined as shifting). For formal expressions and rigorous derivations, we refer readers to Chen et al. [2].*

Such equivariance network extracts very expressive equivariant features regarding input rotations. However, since the equivariance design is regarding the entire input point cloud without spatial scale concepts, EPN is not able to handle object-level equivariance when consuming a 3D scene.

Inspired by EPN, we achieve rotation equivariance via this feature augmenting strategy (sometimes referred to as feature orbits or filter orbits). Yet in contrast to EPN, our method allows learning *object-level* equivariant features in full scene-scale inputs for detection.

4. Equivariant Object detection Network

Our goal is to explicitly leverge object rotation equivariance in 3D detectors, to increase the network expressivity and discriminative power. We target at a modular design that allows us to directly equip state-of-the-art 3D detection models with object-level rotation equivariance.

We ground our method on popular bottom-up detector designs, which have been adopted by many state-of-the-art point cloud object detectors, such as VoteNet [17], Point-Pillar [11], PointRCNN [21] and CenterPoint [29]. These methods generally include three modules: **seed feature extraction** in which a backbone processes point clouds into a dense set of features, **region context aggregation** that summarizes spatial regions (such as vote clustering [17], map-view convolution [11, 29], and RoI pooling [21]) to produce a sparse set of features, and **OBB (oriented bounding box) generation** that proposes candidates from each regional feature. We show an illustration on the top of Figure 2. We adapt these modules for object-level equivariance.

There are two key challenges in our designs. The first is how to determine a proper local spatial support for equivariant feature computation so that the rotation equivariance is restricted to the object level. The second is how to aggregate contexts without breaking object-level equivariance. We propose Equivariant Object detection Network (EON), which incorporates our **rotation equivariance suspension** mechanism to address both challenges. We show an illustration of EON in the bottom of Figure 2.

4.1. Overview

Our main idea is that we let our model compute equivariant features up to an intermediate stage that is expected to be associated with object-level features. After this intermediate stage, we aggregate context geometric information only, as context orientations can break object-level rotation equivariance. To do this, we decompose an equivariant feature into an orientation hypothesis and an invariant feature in the object frame. The following context aggregation will only happen on the invariant geometric features (i.e., equivariance "suspension"). Then, we generate OBB proposals from the aggregated invariant features, so the OBB proposals are also represented in the local object frame instead of the global scene frame. Finally, we rotate the object-frame proposals back to scene-frame OBBs using the object feature orientation (i.e., "resuming" the object orientation). See the bottom of Figure 2 for illustration.

One can conceptually verify the object-level equivariance in an ideal case. When an object rotates in a static scene, the associated object-level feature orientation changes in the same way, while the object-frame feature remains unchanged. Therefore, the object-frame OBB proposals are invariant to the rotation, and hence the scene-frame OBBs do not change their shapes while rotating equivariantly with the objects. Notice that such equivariant design not only facilitates the equivariant detection outcomes, but more importantly allows learning better geometric object features [2, 4, 7]. We ground our idea onto the bottom-up detector design (Figure 2).

Seed feature extraction. EON extracts equivariant seed features and suspends equivariance before the region aggregation stage where large-scale context information gets exchanged. Since this suspension happens in an intermediate stage, the deep detection network can adaptively adjust the effective spatial support [15].

Region context aggregation. To model contexts, EON decomposes each equivariant seed feature by predicting an orientation and extracting an object-frame seed feature. Ideally, object-frame seed features should be invariant to object poses and provide only geometric information for context aggregation. Therefore, only object-frame seed features are fed into the region aggregation module to capture useful context information such as co-occurrences of objects.

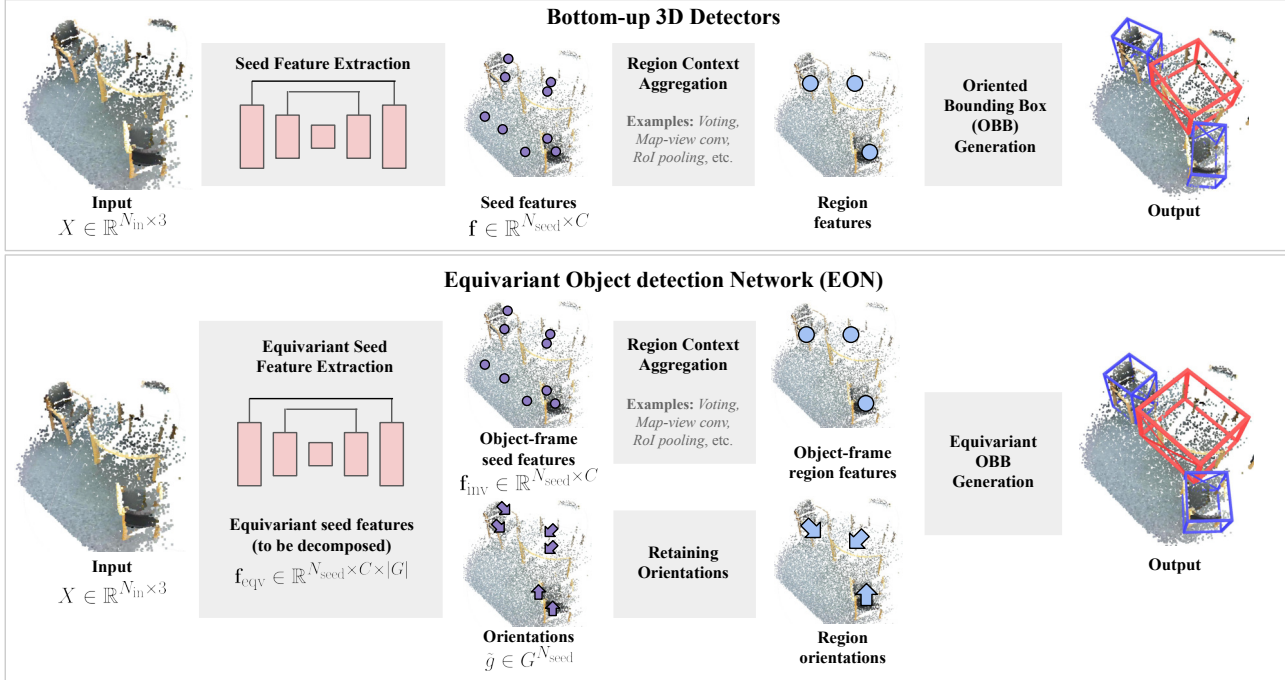**OBB generation.** Finally, the object-level equivariant OBB

Figure 2. Overview of our technical approach. **Top**: We ground our method on popular bottom-up detector designs, including three modules (i.e., seed feature extraction, region aggregation, and OBB generation). **Bottom**: (*Equivariant seed feature extraction*) In our modular design, we maintain equivariance computation up to seed features which are expected to capture object-level information. (*Decomposed region context aggregation*) To maintain object-level equivariance, we decompose equivariant seed features into orientation hypotheses and object-frame invariant features. We aggregate the invariant geometric features, but not orientations (they are "suspended"). (*Equivariant OBB generation*) We generate object-frame OBBs from the region features and resume their orientation hypotheses to rotate them back to the scene frame.

generation module *resumes* the orientation information. It generates proposals in the object frame and transforms the proposals back to the scene frame using the predicted orientations. In the following, we describe EON's three modules in detail and how to plug them into modern 3D detectors.

4.2. Equivariant seed feature extraction

EON extracts equivariant seed features and expects them to capture object-level information. In the following we focus on a single point output of a single layer, in order to make the symbols concise. Extending them to full point clouds and to entire networks is straightforward.

Given a layer $\psi : \mathbb{R}^{N \times C'} \to \mathbb{R}^C$ that maps some input point features $X \in \mathbb{R}^{N \times C'}$ to a point feature $\mathbf{f} \in \mathbb{R}^C$ (such as a KPConv [24] or a PointNet [18]), its equivariant computation to map $X_{\text{eqv}} \in \mathbb{R}^{N \times C' \times |G|}$ to $\mathbf{f}_{\text{eqv}} = \psi(X_{\text{eqv}}) \in \mathbb{R}^{C \times |G|}$ is defined by:

$$\mathbf{f}_{\text{eqv}}(g) = \psi(X_{\text{eqv}}(g)) \in \mathbb{R}^C, g \in G, \qquad (1)$$

where G is a discrete subgroup of SO(3). For example, if one only considers rotation along a single axis, i.e., SO(2), a discretization can be $\{0, \pi/2, \pi, 3\pi/2\}$. And $X_{\text{eqv}} \in \mathbb{R}^{N \times C' \times |G|}$ denotes the input equivariant features

and the number of input points N can vary. For the first layer where the input $X \in \mathbb{R}^{N \times C'}$ is non-equivariant points, we pre-augment it to X_{eqv} such that $X_{\text{eqv}}(g) = T_g^{-1}[X]$, with T_g^{-1} denoting the inverse rotation of g. We follow EPN [2] to attach a 1D convolution over the $|G|$ channel for each layer ψ to further increase expressivity. Our formulation allows a drop-in replacement for most detection backbones. This equivariant computation is performed for the backbone seed feature extraction network.

Objectness-aware equivariant feature decomposition. We expect the equivariant seed feature extraction module to capture object-level geometric features. Thus, we suspend equivariant feature computation after this module and let the deep network to adaptively learn a proper effective receptive field [15]. To suspend equivariance, we decompose each equivariant feature $\mathbf{f}_{\text{eqv}}$ into an invariant feature represented in an object frame $\mathbf{f}_{\text{inv}}$, as well as the orientation $\tilde{g}$ of the object frame w.r.t. the scene frame. However, although object frame and orientation are well defined for foreground objects, these concepts do not hold for background stuff such as wall and road. Thus, we propose objectness-aware feature orbit decomposition which distinguishes between foreground and background points.

Specifically, we define foreground points as those inside

an OBB, and all other points as background points. To predict the objectness for a seed point, we may jointly train a segmentation head attached to the seed features. Then, our objectness-aware feature decomposition produces object-frame features as:

$$\mathbf{f}_{\text{inv}} = \begin{cases} \mathbf{f}_{\text{eqv}}(\tilde{g}) \in \mathbb{R}^C, & \text{if it is foreground} \\ \text{maxpool}(\{\mathbf{f}_{\text{eqv}}(g)\}_{g \in G}) \in \mathbb{R}^C, & \text{otherwise} \end{cases}$$

$$(2)$$

where $\tilde{g}$ denotes a predicted orientation given by an orientation classification head H, $\tilde{g} = \arg\max_g H(\mathbf{f}_{\text{eqv}}(g))$, and maxpool denotes max pooling over all $g \in G$. Notice we abuse the term "object-frame seed feature" to also cover background points. The prediction head H is jointly trained using a softmax classification loss. We generate the ground-truth orientation label by discretizing the orientation angle of each ground-truth OBB into $|G|$ bins, and then assigning the bin labels to all points inside the OBB. For background points, the orientations are undefined, and we ignore them in the rest of a forward pass.

4.3. Decomposed region context aggregation

In state-of-the-art 3D detectors, the context aggregation module summarizes features within a large spatial volume, allowing the following OBB generation to extract useful contextual information. Examples include the vote clustering in VoteNet [17], RoI pooling in PointRCNN [21], and Map-view convolution in CenterPoint [29] and PointPillar [11].

In our region context aggregation module, we aim to not only aggregate contextual geometric information, but also maintain object-level equivariance. Thus, we have two technical goals. The first goal is to aggregate invariant seed features for a sparse set of regions. We use the original region aggregation module from the detector. For every region, it takes as input the seed features inside the region, and outputs an object-frame invariant region feature.

The second goal is to retain the orientation of an object-of-interest for the region, and filter out context orientations. Ideally, the object-of-interest in a region is defined according to the proposal label assigned to the central point of the region, such as the target assignment methods used in PointRCNN [21] and CenterPoint [29]. In this case, the output object-of-interest orientation $\tilde{h}$ is simply the prediction $\tilde{g}$ at the central point of the region.

However, for detectors that use IoU threshold-based target assignment, the object-of-interest for a region is not naturally defined, because the proposals from a region are dynamically associated with nearby ground-truth OBBs. In this case, we take the mode orientation in the region. The idea is that if most foreground points in the region are from the same object, it is likely that the proposals are also mostly for that object. Thus we can capture its orientation by a mode

selection. For regions that do not have a significant mode, the proposals are likely to be low-quality and get filtered out, or simply not be associated to any targets. Thus, they would yield little negative effects on the outcomes.

4.4. Equivariant OBB generation

Given an object-frame region feature and its orientation, our object-level equivariant OBB generation module yields OBBs in the scene frame. This is divided into two steps. First, we generate OBBs in the object frame by using the original module. Then, we transform the OBBs back to the scene frame using the region orientation. Specifically, given the center (or corner points) $c_{\text{inv}} \in \mathbb{R}^3$ and the orientation θ_{inv} of an object-frame OBB, we transform the box to scene frame by replacing the orientation with $\theta = \theta_{\text{inv}} - \tilde{h}$ and replacing the center with $c = R^{-1}c_{\text{inv}}$ where R is the rotation matrix for $\tilde{h}$.

5. Experiments

In this section we apply our EON design to modern detection models with various backbones and we show experimental results on several indoor and outdoor benchmarks.

5.1. Experiment setup

Datasets. We adopt ScanNetV2 dataset [6] and SUN RGB-D dataset [23] for indoor 3D detection. ScanNetV2 provides 1513 indoor scans with semantic segmentation labels. For benchmarking oriented object detection, we use the detection labels from Scan2CAD [1] which provides CAD models with oriented bounding boxes aligned with common objects in the scans. We report performances on 9 categories which have more than 200 instances and put all others into the "Others" class. SUN RGB-D dataset has around 5K RGB-D images annotated with amodal oriented boxes for 37 object classes. We follow the same setup as VoteNet [17] and report performances on 10 classes for SUN RGB-D. For outdoor 3D detection, we use KITTI [9] for evaluation, which contains 7481 training samples and 7518 testing samples. We follow the original evaluation protocol in KITTI.

Implementation details. We implement our EON design for three 3D detectors including VoteNet [17] and a transformer-based state-of-the-art method from Liu et.al. [14] for indoor scenes, and PointRCNN [21] for outdoor scenes. We plug in our proposed modules to replace their original modules, and denote the resultant models as EON-VoteNet, EON-Liu et.al., and EON-PointRCNN. As for the seed feature extraction module, all three methods originally use PointNet++ [19] as their backbones. To demonstrate the applicability of our EON to different backbones, we use KPConv [24] as the backbone for VoteNet. We show the KPConv backbone architecture in the supplement. Then we replace the original seed feature extraction modules as described in Sec. 4.2. To

predict orientation for each equivariant seed feature, we use a two-layer MLP for the head H_o. Since most objects are subject to gravity constraints, we only consider one degree of freedom (i.e., yaw) in the rotation group. We discretize it into 4 bins for classification. To predict the foreground segmentation, we use another two-layer MLP head for binary classification.

The region aggregation module of VoteNet consists of a voting stage which generates a spatial translation ("vote") for each seed point, a grouping stage using the updated spatial coordinates, and a summarization stage to generate region features. Since the grouping happens in the scene frame, we inversely rotate the vote to the scene frame for the grouping. The region orientation $\tilde{h}$ is determined by mode selection. PointRCNN's region aggregation module is an RoI pooling where the object-of-interest is well defined. Its orientation $\tilde{h}$ uses the seed point that generates the RoI.

For all other settings such as the detection heads, input resolution, preprocessing, hyper-parameters and training configurations, we follow the implementations provided by the authors[†]. We also use the same settings for our EON for fair comparisons.

5.2. Indoor datasets

For 3D detection on indoor scenes, we compare our EON-VoteNet to VoteNet on the ScanNetV2 and the SUN RGB-D datasets.

ScanNetV2. For ScanNetV2 dataset, we show comparison results on validation set in Table 1. As we can see, EON outperforms the vanilla methods significantly, with a boost in mAP by 6.3% for VoteNet and 9.0% for the transformer-based Liu et.al. [14]. Notably, the performance gains are particularly significant for objects that are "thin" from the top view, such as display (+13.0%/+37.0% AP for VoteNet/Liu et.al.). We suppose a major reason be that for these objects, the accurate orientation prediction is crucial, and thus object-level equivariance design kicks in to help. Significant boost is also obtained for small objects like trash-bin (+17.9%/7.4% AP), which might be due to more accurate vote orientations brought by object-level equivariance, leading to better pinning-down of small objects' center positions.

SUN RGB-D. We also validate the effectiveness of our EON on SUN RGB-D dataset, in which the single-view derived point clouds are highly incomplete compared to the scans in ScanNetV2. SUN RGB-D is more difficult as the partial point clouds pose challenges not only on detection itself, but also on orientation estimation. We show comparison results on SUN RGB-D dataset in Table 2. We draw the same observation as on ScanNetV2 that our EON-VoteNet

outperforms VoteNet on all categories, leading to a boost in mAP by 3.1%, despite the difficulty in orientation prediction. Similarly, the performance gain is more significant for thin object categories such as bookshelf (+8.9% AP) and bathtub (+10.2% AP).

5.3. Outdoor dataset

For outdoor scenes, we evaluate our EON on KITTI 3D object detection dataset by equipping PointRCNN with our module designs, denoted as EON-PointRCNN. We show the comparison on the KITTI validation set in Table 3. We observe that our EON-PointRCNN outperforms the vanilla model at all difficulty levels for Car and Pedestrian. The performance boost is most significant on the pedestrian category, where the AP gains are 3.02%, 5.15%, 2.90% for Easy, Moderate, and Hard subsets, respectively. We conjecture that this might be because the object-level rotation equivariance is more useful when object orientations have a large variation, which is the case for indoor objects and pedestrians in outdoor scenes. Thus, our EON design has significant performance gains on indoor scenes and pedestrians on KITTI. As for cars and cyclists, they are mostly axis-aligned with less orientation variations. Thus, the performance gains are less significant.

Notice that EON-PoinRCNN and EON-VoteNet adopt different backbone networks (PointNet++ and KPConv, respectively). The consistent improvements across indoor/outdoor scenes and different models validate the wide applicability of our EON as a modular design.

5.4. Analysis

To provide some insights on how our EON helps in 3D detection, we conduct several analysis experiments on ScanNetV2 dataset. We show further analysis on using category-level pose estimation, orientation discretization, suspension versus pooling, object rotation augmentation, and time and numbers of parameters in the supplementary material.

Where to suspend equivariance suspension. Our core design in EON is suspending equivariance at the region aggregation stage to extract object-level equivariant features. We evaluate our design on VoteNet by suspending equivariance at earlier or latter stages, including suspending at the second last backbone layer (denoted as "PreEON") and at the proposal generation stage ("FullEON"). "FullEON" keeps equivariant computation across the whole feature extraction, thus it is using full scene-level equivariance. We show the results in Table 4.

We draw two major observations. First, our EON compares much favorably with using full scene-level equivariance. This shows the rotation equivariance to full visual content does not help much in object detection, as the target object can rotate arbitrarily when the rest of the scene context remains unchanged. Second, our design to suspend equiv-

---

[†]VoteNet: `github.com/facebookresearch/votenet`, commit: 2f6d6d3. Liu et.al.: `https://github.com/zeliu98/Group-Free-3D`, commit: ef8b7bb. PointRCNN: `github.com/open-mmlab/OpenPCDet`, version: 0.3.0, commit: c9d31d3.
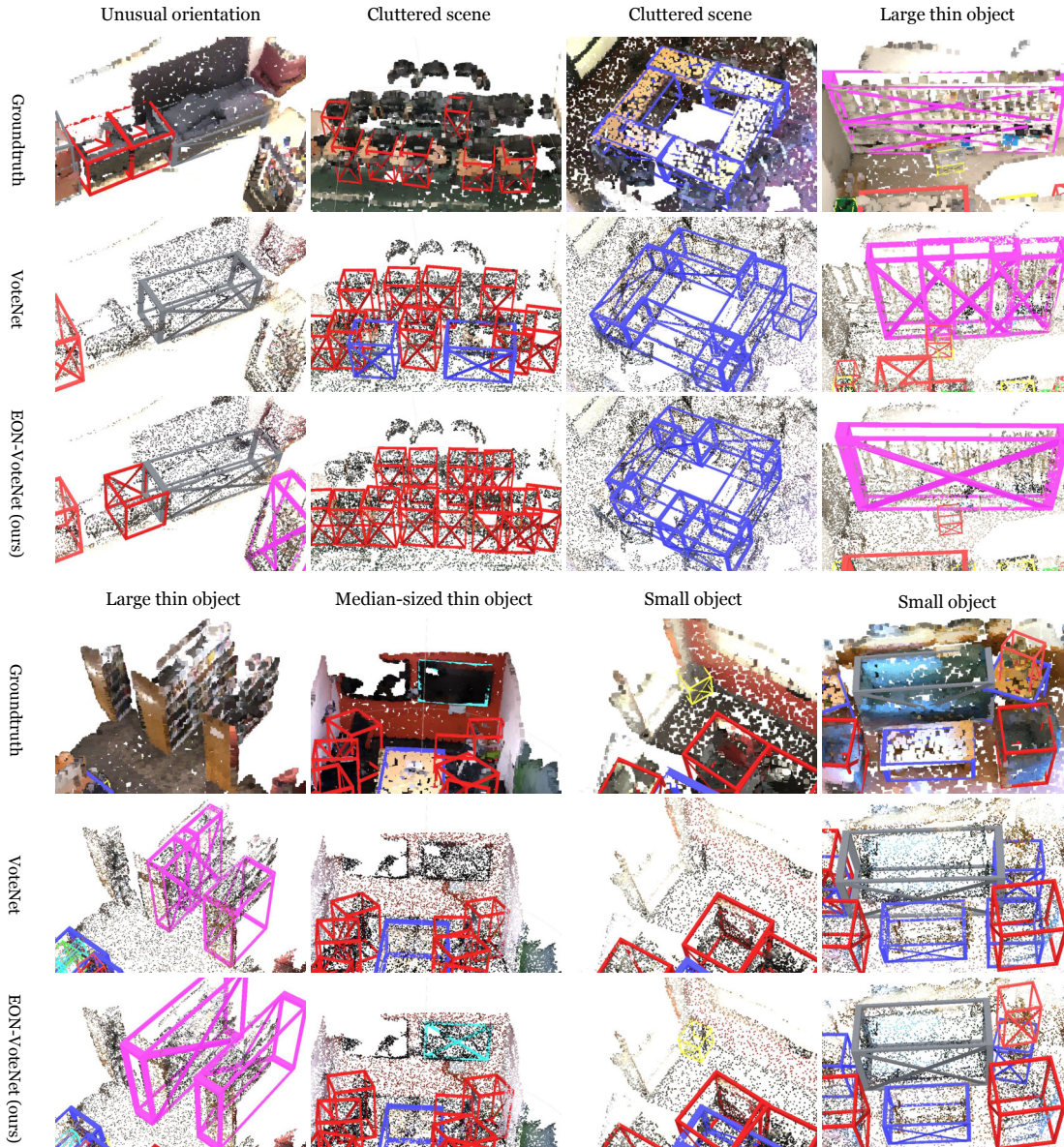
Figure 3. Qualitative results on ScanNetV2 dataset. Each row depicts a typical case where the object-level equivariance design shows promising improvements. Color information is not used for models, but only for visualization.

Method	Trash-bin	Display	Others	Bathtub	Chair	Cabinet	Bookshelf	Table	Sofa	Bed	mAP
VoteNet [17]	27.4	22.8	11.6	40.1	84.1	47.5	45.3	72.3	67.4	85.4	50.4
EON-VoteNet (ours)	**45.3**	**35.8**	**16.4**	**49.1**	**86.3**	**51.9**	**51.0**	**75.0**	**68.7**	**87.2**	**56.7**
Liu et.al. [14]	51.1	16.5	24.2	38.8	**88.0**	49.2	33.1	69.7	**63.7**	**89.7**	52.4
EON-Liu et.al. (ours)	**58.5**	**53.3**	**27.6**	**45.8**	84.9	**57.1**	**68.3**	**82.8**	45.8	89.4	**61.4**

Table 1. Performance (AP_{25}) comparison on ScanNet V2 validation set using Scan2CAD detection labels.

ariant computation at region aggregation yields best results, supporting our assumption that input to region aggregation roughly corresponds to an object level.

Qualitative comparisons. We show qualitative examples in Figure 3 to depict several typical cases where our EON shows great effectiveness.

In the first column, first row of Figure 3, we show that our EON-VoteNet is able to detect objects with unusual orientations. In the shown example, the black chairs are facing the wall, which is rare. Without having object-level equivari-

Method	Nightstand	Toilet	Chair	Bathtub	Dresser	Desk	Table	Bookshelf	Sofa	Bed	mAP
VoteNet [17]	58.4	89.9	74.5	68.8	34.6	26.1	48.5	22.3	65.9	84.9	57.4
EON-VoteNet (ours)	**60.0**	**91.6**	**75.9**	**79.0**	**35.4**	**27.2**	**49.6**	**31.2**	**68.1**	**86.7**	**60.5**

Table 2. Comparison of VoteNet with and without our EON design on SUN RGB-D dataset. Performances are measure by AP_{25}.

Method	Car (IoU=0.7)			Pedestrian (IoU=0.5)			Cyclist (IoU=0.5)		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
PointRCNN [21]	88.39	78.29	77.47	64.12	55.93	51.35	**87.81**	72.71	67.25
EON-PointRCNN (ours)	**89.11**	**78.61**	**77.55**	**67.14**	**61.08**	**54.25**	87.33	**73.36**	**67.41**

Table 3. Results on KITTI 3D detection validation set. IoU threshold for AP is 0.7 for Car and 0.5 for Pedestrian/Cyclist, respectively.

Method	mAP@0.25	mAP@0.5
PreEON-VoteNet	54.5	34.1
FullEON-VoteNet	52.7	26.9
EON-VoteNet (ours)	**56.7**	**36.5**

Table 4. Evaluation on where to suspend equivariance on Scan-NetV2 dataset with Scan2CAD detection labels.

Method	mAP@0.25	mAP@0.5
VoteNet [17]	50.4	28.3
EON-VoteNet (ours)	**56.7**	**36.5**
Oracle (ori.)	64.0	42.9
Oracle (seg.)	65.3	40.6
Oracle (ori. and seg.)	**69.8**	**47.0**

Table 5. Comparison to oracle model on ScanNetV2 dataset with Scan2CAD detection labels.

ance by design, VoteNet misses these chairs. Nevertheless, equipped with EON, it successfully detects a chair with a very unusual facing. Notice that EON-VoteNet is able to detect a bookshelf that is also in a less common orientation.

In the second columns we show results on a cluttered scene where many similar chairs are packed together with minor orientation differences among them. While the baseline VoteNet can detect most of the chairs, their shapes are distorted. This might be partially due to the different orientations among the chair instances. In contrast, our EON-VoteNet produces a nearly identical shape across the oriented chairs, showing the promise to explicitly handle object rotation symmetry. In the third column, we show another cluttered scene with 6 similar desks next to each other. Again, our EON-VoteNet is able to detect nearly equal-shaped desks, while VoteNet yields shape-varying boxes.

In the fourth column and the first two columns in the second row, we show the detection of thin objects (including displays and bookshelves) where EON-VoteNet improves VoteNet most significantly. This aligns with our quantitative observations in Tables 1 and 2. In the last two columns in the second row, we show that our EON-VoteNet is robust to detect small objects such as trash-bins.

Oracle case exploration. To explore the extent to which our EON can potentially bring improvements to the baseline methods, we show an oracle case exploration in Table 5. In the oracle models, we use the ground-truth orientation labels (discretized into 4 bins) and foreground segmentation in both training and testing. This oracle model shows a strong improvement of +19.4%/+18.7% for AP_{25} and AP_{50} over the baseline VoteNet. When using only ground-truth orientation labels or segmentation, the performance boosts are also promising. Thus, we might expect further improvements

along with better orientation prediction and segmentation prediction, possibly from external domain-tailored models.

6. Conclusion

In this work we explore how to leverage object-level rotation equivariance for 3D object detection. To this end, we propose Equivariant Object detection Network (EON) with our core design to suspend rotation equivariance at an intermediate stage in the feature learning backbone. We apply our design to various backbones and models for both indoor and outdoor scenes. Our experiments show that EON can consistently improve the state-of-the-art detectors, indicating the effectiveness of explicitly modeling object-level equivariance in detection models.

Limitations. One major limitation is the jointly learned orientation/segmentation predictions which might be less accurate than being treated separately. As suggested by the oracle experiments, while our current implementation can already benefit from the object-level equivariance, it is far from reaching the potential maximum gains. Nevertheless, our results suggest the promise of leveraging object-level equivariance for 3D detection. Another limitation is that we assume 1D rotation along the gravity axis. While extending to 3D rotation is straightforward, further exploring it might be beneficial for scenarios such as robotic manipulation.

References

[1] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X Chang, and Matthias Nießner. Scan2cad: Learning cad model alignment in rgb-d scans. In *CVPR*, 2019. 5

[2] Haiwei Chen, Shichen Liu, Weikai Chen, Hao Li, and Randall Hill. Equivariant point network for 3d point cloud analysis. In *CVPR*, 2021. 1, 2, 3, 4

[3] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *ICML*, 2019. 2

[4] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *ICML*, 2016. 1, 2, 3

[5] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv:1801.10130*, 2018. 2

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 5

[7] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas Guibas. Vector neurons: A general framework for so(3)-equivariant networks. *ICCV*, 2021. 1, 2, 3

[8] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so(3) equivariant representations with spherical cnns. In *ECCV*, 2018. 2

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 5

[10] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: a rotation-equivariant detector for aerial object detection. In *CVPR*, 2021. 2

[11] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019. 2, 3, 5

[12] Jiaxin Li, Yingcai Bi, and Gim Hee Lee. Discrete rotation equivariance for point cloud recognition. In *ICRA*, 2019. 2

[13] Xiaolong Li, Yijia Weng, Li Yi, Leonidas Guibas, A Lynn Abbott, Shuran Song, and He Wang. Leveraging se(3) equivariance for self-supervised category-level object pose estimation. *arXiv:2111.00190*, 2021. 1, 2

[14] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *ICCV*, 2021. 2, 5, 6, 7

[15] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *NeurIPS*, 2016. 2, 3, 4

[16] Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang. 3d object detection with pointformer. In *CVPR*, 2021. 2

[17] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *ICCV*, 2019. 1, 2, 3, 5, 7, 8

[18] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 4

[19] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv:1706.02413*, 2017. 3, 5

[20] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *CVPR*, 2020. 2

[21] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*, 2019. 1, 2, 3, 5, 8

[22] Weijing Shi and Raj Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *CVPR*, 2020. 2

[23] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 5

[24] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *CVPR*, 2019. 3, 4, 5

[25] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv:1802.08219*, 2018. 1, 2

[26] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *CVPR*, 2017. 1

[27] Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Yiming Zhang, Kai Xu, and Jun Wang. Mlcvnet: Multi-level context votenet for 3d object detection. In *CVPR*, 2020. 1, 2

[28] Guo-Ye Yang, Xiang-Li Li, Ralph R Martin, and Shi-Min Hu. Sampling equivariant self-attention networks for object detection in aerial images. *arXiv:2111.03420*, 2021. 2

[29] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, 2021. 1, 2, 3, 5

[30] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3dnet: 3d object detection using hybrid geometric primitives. In *ECCV*, 2020. 2

[31] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, 2018. 2

[32] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv:1908.09492*, 2019. 2