

# Quantifying Data Augmentation for LiDAR-based 3D Object Detection

Martin Hahner<sup>1</sup>, Dengxin Dai<sup>1</sup>, Alexander Liniger<sup>1</sup>, and Luc Van Gool<sup>1,2</sup>

**Abstract**—In this work, we shed light on different data augmentation techniques commonly used in Light Detection and Ranging (LiDAR) based 3D Object Detection. For the bulk of our experiments, we utilize the well known *PointPillars* [1] pipeline and the well established KITTI [2] dataset. We investigate a variety of *global* and *local* augmentation techniques, where *global* augmentation techniques are applied to the entire point cloud of a scene and *local* augmentation techniques are only applied to points belonging to individual objects in the scene. Our findings show that both types of data augmentation can lead to performance increases, but it also turns out, that some augmentation techniques, such as individual object translation, for example, can be counterproductive and can hurt the overall performance. We show that these findings transfer and generalize well to other state of the art 3D Object Detection methods and the challenging STF [3] dataset. On the KITTI [2] dataset we can gain up to 1.5% and on the STF [3] dataset up to 1.7% in 3D mAP<sub>40</sub> on the moderate car class.

## I. INTRODUCTION

The last years have seen tremendous progress in 3D Object Detection for autonomous driving and the task really only emerged in 2017, when the KITTI [2] dataset, originally introduced in 2012, was extended by novel benchmarks for 3D Object Detection including 3D and bird’s eye view (BEV) evaluation. Since then, many more publicly available LiDAR datasets with 3D bounding box annotations have followed. nuScenes [4], the Lyft Level5 Perception Dataset [5] and STF [3] are just some of them, others include [6], [7], [8] and [9]. With LiDAR sensors getting cheaper [10] and thereby becoming a viable option for autonomous driving, these LiDAR datasets are fundamental to improve the current state of the art in LiDAR based 3D Object Detection.

Further, the motivation behind the task of 3D Object Detection is that autonomous driving cars need to find a trajectory in the real world (a 3D space). Hence, 2D Object Detection is insufficient since 2D Object Detection only delivers a location and dimensions in the image plane. Even if 2D Object Detection results are combined with Monocular Depth Estimation methods such as [11], the performance is not as high as directly reasoning in 3D LiDAR point clouds. Also, Monocular 3D Object Detection methods like Pseudo-LiDAR [12] often re-purpose 3D Object Detection pipelines meant for point cloud processing. This indicates that point clouds are indeed a much better representation

to detect objects in 3D compared to utilizing purely image based pipelines. The authors of Pseudo-LiDAR [12] propose to convert image-based depth maps into pseudo-LiDAR representations (essentially fake point clouds). Newer state of the art Monocular 3D Object Detection methods that are not relying on such pseudo-LiDAR point clouds anymore, like DD3D [13] still lack far behind in performance compared to state of the art LiDAR based object detectors, *e.g.* PV-RCNN [14] (16.34% vs. 81.43% 3D mAP on moderate cars in the KITTI [2] dataset, respectively). Image based 3D Object Detection might just be too ill-posed, which makes us believe that in order to reach the desired level of safety for autonomous driving cars, we need a sensor on-board that is able to directly measure raw 3D depth information.

To specify the task at hand, what we are looking for in 3D Object Detection for autonomous driving are specifically seven degrees of freedom as opposed to only four in 2D Object Detection. In the task of 3D Object Detection we want to predict the center position of an object  $x_c, y_c, z_c$ , its dimensions  $w, l, h$  and the *yaw* angle  $\theta$  (rotation around the upright axis). Since data augmentation for 3D Object Detection has been found crucial by many works [15], [16], [17], we want to investigate data augmentation for LiDAR based 3D Object Detection in great detail. Given our findings (presented in Section IV), we believe that such an extensive augmentation study has been long overdue.

The contributions in our paper are three-fold, first and foremost, we present an in-depth study of augmentation methods for LiDAR based 3D Object Detection. This study allows practitioners to short cut time consuming experiments and get good results quicker. Second, we show some non-intuitive results, for example, translation of objects can reduce performance, or that excluding *hard* cases from training can increase the performance for these *hard* cases during evaluation. These findings pose new research questions and suggest some insight how 3D Object Detection networks might work internally. Finally, based on our study of augmentation methods, we propose a new augmentation policy that is able to increase the performance of several state of the art 3D Object Detection methods on two different LiDAR datasets: KITTI [2] and STF [3].

## II. RELATED WORK

In this Section, we review relevant work in the field of LiDAR-based 3D Object Detection (Section II-A) and LiDAR Data Augmentation (Section II-B).

<sup>1</sup>Martin Hahner, Dengxin Dai, Alexander Liniger and Luc Van Gool are all with the Toyota TRACE-Zurich team at the Computer Vision Lab, ETH Zurich, 8092 Zurich, Switzerland. `firstname.lastname@vision.ee.ethz.ch`

<sup>2</sup>Luc Van Gool is also with the Toyota TRACE-Leuven team at the Dept. of Electrical Engineering ESAT, KU Leuven, 3001 Leuven, Belgium. `luc.vangool@kuleuven.be`

### A. LiDAR-based 3D Object Detection

The earliest works in this line of work include MV3D [18], PIXOR [15] and AVOD [19]. They all divide the 3D space into a voxel grid. Then they use hand-crafted features (*e.g.* point count per voxel) and hand-crafted neighborhood features (*e.g.* maximum point count per pillar). In the end, they have a 3D tensor of neighborhood features  $(x, y, c)$  and apply regular 2D convolutions on them. The drawback of these methods mostly lies in the hand-crafted feature design.

PointNet [20] laid the foundations for many applications, not just for 3D Object Detection. Their work can also be applied to point cloud Classification, Part Segmentation and Semantic Segmentation. The main reason their work is so applicable to multiple domains is because its architecture is relatively lightweight, yet highly efficient and effective, it can process up to one million points per second.

VoxelNet [16] is one of the seminal works in the area of 3D Object Detection, as a first, this work presents a 3D Object Detection pipeline in an end-to-end fashion without any hand-crafted features. They apply a PointNet-like [20] architecture to every individual voxel leading to a 4D tensor  $(x, y, z, c)$ . Then they perform computationally intensive 3D convolutions to consolidate the  $z$ -dimension, yielding a 3D tensor  $(x, y, c')$ . After this consolidation step, they only use computationally much cheaper 2D convolutions to generate and regress their region proposals.

Another significant contribution to the area of LiDAR based 3D Object Detection is SECOND [17], they were able speed up the bottleneck of earlier works: 3D convolutions, by proposing sparse 3D convolutions. Building on top of [17], PointPillars [1] gets entirely rid of the  $z$ -dimension by dividing the 3D space into pillars instead of voxels. This change by itself gives a 10-100x speed-up compared to VoxelNet [16]. Additionally, to further speed up the novel encoder, they process each pillar only using a single  $1 \times 1$  convolution + max-pooling layer instead of a more complex PointNet-like [20] architecture.

Shi *et al.* achieved several recent milestones in 3D Object Detection. PointRCNN [21] is a two-stage architecture, where the first stage generates 3D bounding box proposals from a point cloud in a bottom-up manner and the second stage refines these 3D bounding box proposals in a canonical fashion. Part-A<sup>2</sup> [22] is part-aware in a sense that the network takes into account which part of the object a point belongs to. It leverages these intra-object part locations and can thereby achieve better results. PV-RCNN [14] and its successor PV-RCNN++ [23] are the latest of their works that simultaneously process (coarse) voxels and the raw points of the point cloud at the same time.

CenterPoint [24] is a two-stage architecture, where the first stage predicts centers of objects using a key-point detector and the second stage thereafter regresses 3D size and orientation of these objects. VoteNet [25] has not proven itself yet on an automotive dataset, but showed promising results on two indoor scene understanding datasets, ScanNet [26] and SUN RGB-D [27].

### B. LiDAR Data Augmentation

Ever since SECOND [17] introduced ground truth (GT) sampling (a.k.a. *GT-Sampling*, see Section III-D), data augmentation has been widely adopted for 3D Object Detection. PA-AUG [28] extends *GT-Sampling* by dividing the GT objects into sub-partitions and randomly augmenting each sub-partition, *e.g.* dropping all points in a partition. They demonstrate that their augmentation technique can improve the performance of PV-RCNN [14] by 0.47% on moderate cars in the KITTI [2] validation split. Similarly, [29] divide each ground truth object into six pyramids, one inward facing pyramid for each face of the 3D bounding box, and then independently augments each pyramid by random dropout, swap, and/or sparsify operations. This way, they can successfully boost the performance of their SE-SSD detector by 0.31% on moderate cars in the KITTI [2] validation split. [30] also builds on top of *GT-Sampling*, they propose an algorithm to mimic the diverging point pattern that occurs when objects are placed at different distances (due to diverging LiDAR beams) and can improve the performance of PV-RCNN [14] by 0.29% on moderate cars in the KITTI [2] validation split.

PointPainting [31] and PointAugmenting [32] are another line of work that can also be considered as LiDAR data augmentation. The idea of PointPainting [31] is to run an Image Segmentation network on the camera image first and then project the predicted segmentation class scores onto the LiDAR point cloud. PointAugmenting [32] takes this idea one step further and projects deep features of a 2D Object Detector onto the LiDAR point cloud instead. In both cases, the resulting augmented point cloud can still be processed by standard 3D Object Detection methods introduced in Section II-A, resulting in a higher performance compared to processing the original, non-augmented LiDAR point clouds.

Some works use physical modelling to augment clear weather point clouds with artificial fog [33] or artificial snowfall [34] and show improvements on point clouds collected in such real-world adverse conditions.

Also other fields that process point clouds, like point cloud classification make use of augmentation. PointMixup [35] *e.g.* interpolates point clouds from different classes to regularize the training process and PointAugment [36] provides an auto-augmentation framework for this task.

While in this paper, we investigate manually designed standard data augmentation policies, PPBA [37] attempts to automate this design process via an evolutionary algorithm. Their search algorithm narrows down the search space every iteration and adopts the best parameters discovered in previous iterations.

## III. METHOD

In this Section, we present all augmentation techniques investigated in this paper. They can be categorized into four categories:

- *global* augmentations (Section III-A),
- *local* augmentations (Section III-B),
- *GT-Filtering* (Section III-C) and
- *GT-Sampling* (Section III-D).

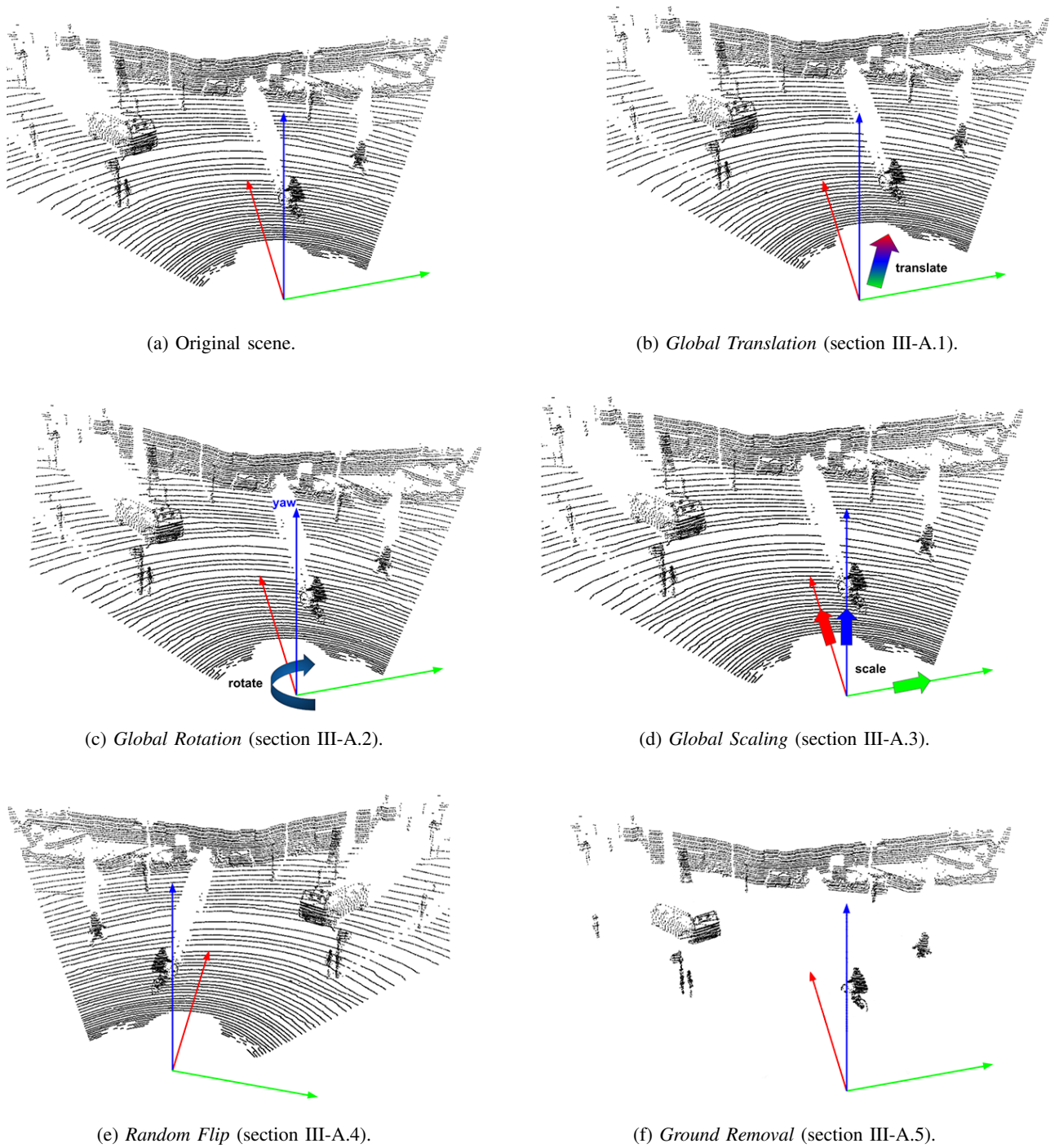


Fig. 1: Visualization of all *global* augmentation techniques investigated in this paper.

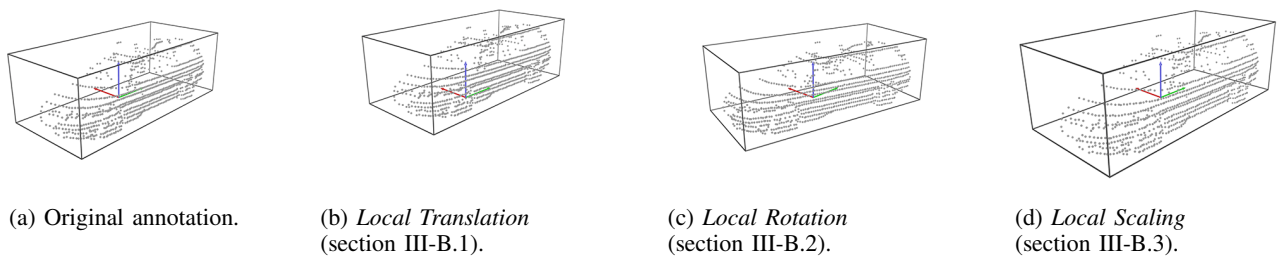


Fig. 2: Visualization of all *local* augmentation techniques investigated in this paper.

### A. Global Augmentations

Global augmentations are applied to the all points in the point cloud  $P = \{p_0, \dots, p_n\}$  and all annotations  $A = \{a_0, \dots, a_m\}$  simultaneously. A visualization of all global augmentation techniques investigated in this paper are shown in Fig. 1 applied to an example scene (Fig. 1a).

#### 1) Global Translation:

Global translation (Fig. 1b) means that we are translating every point  $p(x, y, z) \in P$  such that an augmented point  $p^*$  has the form  $p(x + \Delta x, y + \Delta y, z + \Delta z)$ . Simultaneously, we shift every annotation  $a(x_c, y_c, z_c, w, l, h, \theta) \in A$  such that an augmented annotation  $a^*$  has the form  $a(x_c + \Delta x, y_c + \Delta y, z_c + \Delta z, w, l, h, \theta)$ . Therefore we independently sample  $\Delta x, \Delta y$  and  $\Delta z$  from a normal distribution  $N(0, \sigma^2)$  where  $\sigma$  can take the following values  $\sigma^2 \in \{0.1, 0.2, 0.4\}m$ .

#### 2) Global Rotation:

Global rotation (Fig. 1c) means that we are rotating every point  $p(x, y, z) \in P$  around the upright yaw axis by an angle  $\alpha$  drawn from a uniform distribution  $U(-\beta, +\beta)$  where  $\beta \in \{\pi/8, \pi/4, \pi/2\}$ . Simultaneously, we rotate every annotation  $a(x_c, y_c, z_c, w, l, h, \theta) \in A$  such that the augmented annotation  $a^*$  has the form  $a(x_c, y_c, z_c, w, l, h, (\theta + \alpha) \bmod 2\pi)$ .

#### 3) Global Scaling:

Global scaling (Fig. 1d) means that we are scaling every point  $p(x, y, z) \in P$  in every direction by a scalar  $s$  drawn from a uniform distribution  $U(1 - t, 1 + t)$  where  $t \in \{0.05, 0.1, 0.25\}$  such that an augmented point  $p^*$  has the form  $p(s \cdot x, s \cdot y, s \cdot z)$ . Simultaneously, we scale every annotation  $a(x_c, y_c, z_c, w, l, h, \theta) \in A$  such that an augmented annotation  $a^*$  has the form  $a(s \cdot x_c, s \cdot y_c, s \cdot z_c, s \cdot w, s \cdot l, s \cdot h, \theta)$ .

#### 4) Random Flip:

Random flip (Fig. 1e) means that we flip every point  $p(x, y, z) \in P$  by a 50% chance on the forward facing  $x$ -axis. So if we flip the whole point cloud, an augmented point  $p^*$  has the form  $p(x, -y, z)$ . Simultaneously, if we flip every point  $p(x, y, z) \in P$ , we also flip every annotation  $a(x_c, y_c, z_c, w, l, h, \theta) \in A$  such that an augmented annotation  $a^*$  has the form  $a(x_c, -y_c, z_c, w, l, h, (\theta + \pi) \bmod 2\pi)$ . We do not flip on the sideways facing  $y$ -axis because KITTI [2] and STF [3] only provide 3D bounding box annotations in the camera field of view. So it does not make sense to flip on the  $y$ -axis when on the other side, there are no targets to train for.

#### 5) Ground Removal:

Ground removal (Fig. 1f) means that we remove every point  $p(x, y, z) \in P$  where  $z$  is smaller than a threshold  $\epsilon \in \{1\text{st}, 5\text{th}, 10\text{th}, 15\text{th}\}$  percentile of all  $z$ -values in the point cloud  $P$ . The idea behind this is to get rid of “background” points, because a LiDAR point cloud is typically very unbalanced in terms of “foreground” vs. “background”. In the KITTI [2] training set for example, there are on average 19,047 points per scene in the camera field of view and only 1,382 points ( $\approx 7, 25\%$ ) thereof lie inside 3D bounding box annotations (“foreground”). For the KITTI [2] validation set, this statistic does not look much different, there are on average 18,888

points per scene in the camera field of view and 1,641 points ( $\approx 8, 69\%$ ) thereof lie inside 3D bounding box annotations. This augmentation also has the uniqueness, that it is the only augmentation we investigate, that is applied during training and testing. This is to not mess up the distribution of “foreground” vs. “background” during the different stages. All other augmentations are only applied during training.

### B. Local Augmentations

Similar to the global augmentation techniques presented in the previous Section, local augmentations presented in this Section also involve translation, rotation and scaling. The only difference is, that here we do not apply those transformations to every point  $p(x, y, z) \in P$ , but only to individual annotations and the points that reside inside those annotations. Thereby we augment every annotation independently from every other, meaning that for every annotation we draw a different random value. A visualization of all local augmentation techniques investigated in this paper are shown in Fig. 2 with an example annotation (Fig. 2a).

#### 1) Local Translation:

Local translation (Fig. 2b) means that we are translating every point  $p(x, y, z) \in a$  such that an augmented point  $p^*$  has the form  $p(x + \Delta x, y + \Delta y, z + \Delta z)$ . Simultaneously, we shift the annotation  $a(x_c, y_c, z_c, w, l, h, \theta)$  such that an augmented annotation  $a^*$  has the form  $a(x_c + \Delta x, y_c + \Delta y, z_c + \Delta z, w, l, h, \theta)$ . Again, we independently sample  $\Delta x, \Delta y$  and  $\Delta z$  from a normal distribution  $N(0, \sigma^2)$ , with  $\sigma^2 \in \{0.05, 0.25, 0.5, 1\}m$ .

#### 2) Local Rotation:

Local rotation (Fig. 2c) means that we are rotating every point  $p(x, y, z) \in a$  around the upright yaw axis by an angle  $\alpha$  drawn from a uniform distribution  $U(-\beta, +\beta)$  where  $\beta \in \{\pi/20, \pi/10, \pi/4\}$ . Simultaneously, we rotate the annotation  $a(x_c, y_c, z_c, w, l, h, \theta)$  such that the augmented annotation  $a^*$  has the form  $a(x_c, y_c, z_c, w, l, h, (\theta + \alpha) \bmod 2\pi)$ .

#### 3) Local Scaling:

Local scaling (Fig. 2d) means that we are scaling every point  $p(x, y, z) \in P$  in every direction by a scalar  $s$  drawn from a uniform distribution  $U(1 - t, 1 + t)$  where  $t \in \{0.05, 0.1, 0.25\}$  such that an augmented point  $p^*$  has the form  $p(s \cdot x, s \cdot y, s \cdot z)$ . Simultaneously, we scale every annotation  $a(x_c, y_c, z_c, w, l, h, \theta) \in A$  such that an augmented annotation  $a^*$  has the form  $a(s \cdot x_c, s \cdot y_c, s \cdot z_c, s \cdot w, s \cdot l, s \cdot h, \theta)$ .

### C. GT-Filtering

GT-Filtering is straightforward. One way to filter ground truth objects is to filter them based on their difficulty. In KITTI [2] (and STF [3] accordingly), there are three predefined difficulties: *easy*, *moderate* and *hard*, where some annotations are also labeled as *unknown*.

Another simple filter operation is to exclude annotations from training when they contain less than a certain amount of LiDAR points inside of them. In our experiments, we investigated a minimum threshold of 1, 5 and 10 points.

policy #	Global Translation	Global Rotation	Global Scaling	Random Flip	Ground Removal	Local Translation	Local Rotation	Local Scaling	filter out unknown	filter out hard	filter out moderate	filter by min. points	GT-Sampling	car 3D mAP <sub>40</sub>		
														easy	moderate	hard
0	X	X	X	X	X	X	X	X	X	X	X	X	X	71.98	59.29	55.94
1	0.1m	X	X	X	X	X	X	X	X	X	X	X	X	+6.31	+5.55	+3.93
2	<b>0.2m</b>	X	X	X	X	X	X	X	X	X	X	X	X	<b>+7.21</b>	<b>+6.64</b>	<b>+5.62</b>
3	0.4m	X	X	X	X	X	X	X	X	X	X	X	X	+4.54	+5.18	+4.47
4	X	$\pi/8$	X	X	X	X	X	X	X	X	X	X	X	+7.97	+9.10	+7.90
5	X	$\pi/4$	X	X	X	X	X	X	X	X	X	X	X	+10.34	+9.89	+8.75
6	X	$\pi/2$	X	X	X	X	X	X	X	X	X	X	X	<b>+11.51</b>	<b>+10.73</b>	<b>+10.90</b>
7	X	X	[0.95, 1.05]	X	X	X	X	X	X	X	X	X	X	+7.24	<b>+6.96</b>	<b>+5.43</b>
8	X	X	[0.90, 1.10]	X	X	X	X	X	X	X	X	X	X	+6.43	+5.72	+4.87
9	X	X	[0.75, 1.25]	X	X	X	X	X	X	X	X	X	X	<b>+8.10</b>	+5.43	+3.19
10	X	X	X	✓	X	X	X	X	X	X	X	X	X	<b>+7.02</b>	<b>+6.93</b>	<b>+7.58</b>
11	X	X	X	X	1%	X	X	X	X	X	X	X	X	+0.48	+1.79	+1.17
12	X	X	X	X	5%	X	X	X	X	X	X	X	X	-2.75	-1.51	-1.80
13	X	X	X	X	10%	X	X	X	X	X	X	X	X	-3.19	-1.83	-1.97
14	X	X	X	X	15%	X	X	X	X	X	X	X	X	-4.72	-3.83	-4.01
15	X	X	X	X	X	<b>0.05m</b>	X	X	X	X	X	X	X	+1.79	+0.71	+1.14
16	X	X	X	X	X	0.25m	X	X	X	X	X	X	X	+0.76	+0.52	-0.24
17	X	X	X	X	X	0.50m	X	X	X	X	X	X	X	-0.50	-0.08	-1.35
18	X	X	X	X	X	1.00m	X	X	X	X	X	X	X	-0.77	-0.33	-2.02
19	X	X	X	X	X	X	$\pi/20$	X	X	X	X	X	X	+6.08	<b>+4.66</b>	<b>+4.09</b>
20	X	X	X	X	X	X	$\pi/10$	X	X	X	X	X	X	+4.77	+3.63	+2.73
21	X	X	X	X	X	X	$\pi/4$	X	X	X	X	X	X	<b>+7.87</b>	+4.57	+3.06
22	X	X	X	X	X	X	X	[0.95, 1.05]	X	X	X	X	X	<b>+4.98</b>	<b>+3.99</b>	<b>+3.18</b>
23	X	X	X	X	X	X	X	[0.90, 1.10]	X	X	X	X	X	+4.82	+2.83	+1.67
24	X	X	X	X	X	X	X	[0.75, 1.25]	X	X	X	X	X	+4.79	+1.09	-1.49
25	X	X	X	X	X	X	X	X	✓	X	X	X	X	-0.16	+0.67	+0.54
26	X	X	X	X	X	X	X	X	✓	✓	X	X	X	+1.19	+0.09	+1.16
27	X	X	X	X	X	X	X	X	✓	✓	✓	X	X	+1.79	+0.14	-0.45
28	X	X	X	X	X	X	X	X	X	X	X	1	X	+0.96	+1.01	+0.67
29	X	X	X	X	X	X	X	X	X	X	X	5	X	-0.10	+0.29	+0.84
30	X	X	X	X	X	X	X	X	X	X	X	10	X	+0.76	+0.52	-0.48
31	X	X	X	X	X	X	X	X	X	X	X	X	5	+2.71	+2.86	+2.83
32	X	X	X	X	X	X	X	X	X	X	X	X	10	+1.29	+2.59	+1.70
33	X	X	X	X	X	X	X	X	X	X	X	X	15	-0.19	+1.29	+0.64
34	X	X	X	X	X	X	X	X	X	X	X	X	20	+1.72	+1.31	+0.48
35	X	X	X	X	X	X	X	X	X	X	X	X	25	+0.51	+1.01	+0.32
36	<b>0.2m</b>	$\pi/4$	[0.95, 1.05]	✓	X	0.25m	$\pi/20$	X	✓	X	X	5	15	<b>+15.75</b>	<b>+17.72</b>	<b>+16.59</b>
37	<b>0.2m</b>	$\pi/4$	[0.95, 1.05]	✓	X	0.25m	$\pi/20$	X	✓	X	X	5	X	+11.45	+11.06	+10.60
38	X	$\pi/2$	X	X	X	X	X	X	X	X	X	X	15	+13.14	+14.88	+13.88
39	<b>0.2m</b>	$\pi/4$	[0.95, 1.05]	✓	X	X	$\pi/20$	X	✓	X	X	5	15	+15.95	+18.05	+18.30
40	<b>0.2m</b>	$\pi/4$	[0.95, 1.05]	✓	X	X	$\pi/20$	[0.95, 1.05]	✓	X	X	5	15	+16.08	+18.53	+18.51
41	<b>0.2m</b>	$\pi/4$	[0.95, 1.05]	✓	X	X	$\pi/20$	[0.95, 1.05]	✓	✓	X	5	15	<b>+16.95</b>	<b>+19.20</b>	<b>+18.54</b>
42	<b>0.2m</b>	$\pi/2$	[0.95, 1.05]	✓	X	X	$\pi/20$	[0.95, 1.05]	✓	✓	X	5	15	<b>+16.52</b>	<b>+18.83</b>	<b>+17.77</b>

TABLE I: Results of our extensive augmentation study on the KITTI [2] validation set. Most significant improvements in **bold**. Augmentation policy of PointPillars [1] in **magenta** (#36). Our improved augmentation policies in **cyan** (#39–42).

#### D. GT-Sampling

GT-Sampling means that we are trying to sample additional ground truth objects, *e.g.* cars, from other scenes into the current scene. In order to do so, one has to iterate over all annotations once and construct a database of annotations and their corresponding points. This “trick” also aims at balancing the “foreground” vs. “background” imbalance described earlier in Sec. III-A.5. During training, one only has to check whether the additionally sampled annotations from the database do not collide with any of the ones originally present in the current scene. In our experiments we tried to additionally sample 5, 10, 15, 20 and 25 annotations. If there

is a collision, those annotations are discarded and only the ones that do not collide are kept. This means for example in the setting where we try to sample up to 15 cars, at most 15 annotations are additionally sampled.

## IV. EXPERIMENTS

In this Section, we present the findings of our extensive augmentation study summarized in Table I, II, III, and IV.

We always report 3D mAP<sub>40</sub> proposed by [39], the mean average precision computed over 40 instead of 11 recall operation points originally proposed by the Pascal VOC bench-

mark [40]. For more details, please refer to [39]. In this paper we focus on the *car* class and carry out our main experiments on the KITTI [2] validation set, following the common practice proposed in [41] to split the official training set into 3,712 training and 3,769 validation samples.

In order to investigate all the potential augmentation techniques introduced in Section III, we have to establish a proper baseline where no augmentation technique is applied (policy #0). To reduce the impact of randomness and to have a more meaningful comparison, we carried out all experiments listed in Table I, II, III, and IV exactly three times and report the numbers of the training run which has the highest 3D mAP<sub>40</sub> score on the *moderate* difficulty, the same metric that is used for ranking submissions on the official KITTI leader-board.

Diving into Table I, our first finding is, that there is a huge gap (17.72%) between the baseline with no augmentation (policy #0) and the augmentation policy used in PointPillars [1] (policy #36). This means that almost 1/4 (23%) of the performance of PointPillars [1] can be attributed to its sophisticated augmentation policy.

If we look at individual augmentation techniques, most notably we see that *Global Rotation* (policy #4-6) is the most effective augmentation with an increase of up and above 10% (policy #6). Followed by three other *global* augmentation techniques, *Global Translation* (policy #1-3), *Global Scaling* (policy #7-9) and *Random Flip* (policy #10), all contributing with a performance boost of around 7%.

To show how effective *Global Rotation* is as an augmentation technique, we can compare our best performing *Global Rotation* policy (policy #6), with the advanced augmentation policy of PointPillars [1] where only *GT-Sampling* is removed (resulting in policy #37), and we can see that the two policies perform similarly.

Only adding *GT-Sampling* to policy #6 (resulting in policy #38) we can already half the gap to the performance of the sophisticated policy (two vs. nine augmentation techniques) utilized in PointPillars [1] (policy #36). On the other hand, *GT-Sampling* applied by itself (policy #31-35) does not help as much as other individual augmentation techniques. It only shines in combination with other augmentation techniques, showcasing that the combination of the different augmentation techniques are not simply additive.

method	policy #	<i>easy</i>	<i>moderate</i>	<i>hard</i>
VoxelRCNN [38]	36	91.89	84.36	82.13
	41	<b>92.62</b>	<b>85.14</b>	<b>82.31</b>
PV-RCNN [14]	36	91.28	84.06	81.70
	41	<b>92.55</b>	<b>85.02</b>	<b>82.00</b>
CenterPoint [24]	36	91.63	83.92	<b>81.67</b>
	41	<b>92.04</b>	<b>84.69</b>	81.62
Part-A <sup>2</sup> [22]	36	91.83	82.49	80.17
	41	<b>92.38</b>	<b>83.24</b>	<b>80.28</b>
SECOND [17]	36	89.71	80.77	77.75
	41	<b>90.11</b>	<b>81.23</b>	<b>76.36</b>

TABLE II: Further 3D mAP<sub>40</sub> results of the *car* class on the KITTI [2] validation set.

Investigating the remaining *global* augmentation technique *Ground Removal* (policy #11-14), which is intuitive and often performed in classical LiDAR pipelines [42], is rather disadvantageous than beneficial in our experiments. We speculate that this could be due to two things. First, by removing the lowest points of the scene (in such a naive way) a lot of potentially crucial neighborhood context around “foreground” objects gets deleted. Second, it could also mean that “background” points quite generally encode more valuable information than what is the established view on that matter. In order to clarify this though, further experiments would be necessary (see future work in Section V).

In general we see that *local* augmentation techniques are not as effective as *global* augmentation techniques. Although, while *Local Rotation* (policy #19-21) and *Local Scaling* (policy #22-24) can be considered beneficial, *Local Translation* (policy #15-18) contrarily is clearly not as beneficial as the other two *local* augmentation techniques and can even hurt performance if applied too aggressively.

*GT-Filtering* (policy #25-30) also does not make a big difference, interesting is maybe only the results of policy #26, where we filter all *hard* examples during training and still get a higher performance on them during evaluation (+1.16%). This could mean that if we keep *hard* examples during training, the network might get too distracted by those seemingly diverse *hard* examples and is not able to extract distinct features from them, thereby hurting its generalisation capabilities.

If we now take all those insights from policy #1-35 and apply them to the policy of PointPillars [1] (policy #36) we can further boost its performance. First, by removing the potentially hurtful *Local Translation* (resulting in policy #39), then adding *Local Scaling* (resulting in policy #40) and finally also leaving out *hard* examples during training (resulting in policy #41), we can further improve PointPillars’ [1] results by up to 1.5% on *moderate* and almost 2% on *hard* cars. Lastly in policy #42, where we combine all the individual best augmentation techniques, does not yield the best overall result, showcasing (again) that the combination of individual techniques are not just simply additive.

Table II shows that our findings from Table I are not just limited to PointPillars [1], but generalize well to five other state of the art 3D Object Detection methods.

method	policy #	<i>easy</i>	<i>moderate</i>	<i>hard</i>
VoxelRCNN [38]	36	39.86	39.39	36.20
	41	<b>40.77</b>	<b>41.12</b>	<b>36.55</b>
PV-RCNN [14]	36	<b>41.66</b>	41.94	37.60
	41	41.23	<b>42.67</b>	<b>38.41</b>
CenterPoint [24]	36	41.80	42.45	<b>38.42</b>
	41	<b>43.02</b>	<b>42.87</b>	38.30
Part-A <sup>2</sup> [22]	36	39.84	39.74	36.06
	41	<b>41.01</b>	<b>40.68</b>	<b>37.52</b>
SECOND [17]	36	<b>38.32</b>	38.16	35.17
	41	37.27	<b>38.84</b>	<b>35.84</b>

TABLE III: 3D mAP<sub>40</sub> results of the *car* class on the challenging STF [3] validation set.

method	policy #	<i>easy</i>	<i>moderate</i>	<i>hard</i>
VoxelRCNN [38]	36 + [28]	91.93	84.04	81.95
	36 + [29]	91.73	84.05	81.56
	36 + [30]	91.52	84.26	82.05
	41	<b>92.62</b>	85.14	82.31
	41 + [28]	<u>92.55</u>	<b>85.46</b>	<b>82.65</b>
	41 + [29]	92.52	84.90	82.16
PV-RCNN [14]	41 + [30]	92.46	84.99	82.00
	36 + [28]	90.60	84.06	81.98
	36 + [29]	91.78	84.40	82.16
	36 + [30]	91.80	84.34	82.12
	41	<b>92.55</b>	<u>85.02</u>	82.00
	41 + [28]	92.25	84.91	81.70
	41 + [29]	92.06	<b>85.14</b>	<b>82.41</b>
	41 + [30]	<u>92.27</u>	84.97	<u>82.29</u>

TABLE IV: Comparison to the augmentation methods introduced in [28], [29], and [30]. Again, we present 3D mAP<sub>40</sub> of the *car* class on the KITTI [2] validation set.

Table III demonstrates that the findings from Table I are also not just limited to the KITTI [2] dataset, but transfer well to the challenging STF [3] dataset. We believe that this is also applicable to other LiDAR datasets, such as nuScenes [4] and the Lyft Level5 Perception Dataset [5]. Even though these datasets are captured with different LiDAR sensors and hence can have drastically different number of overall LiDAR points per scene (Lyft Level5 Perception Dataset [5] with up to 190,000 points per scene vs. nuScenes [4] with only around 35,000 points per scene on average), the main issue, the imbalance between “foreground” and “background” points remains the same for all current LiDAR datasets.

Table IV shows our best augmentation policy compared to related work. We can see that our policy #41 can not be outperformed by [28], [29], or [30] applied to policy #36, the common default strategy introduced in PointPillars [1]. Only applying [28], [29], or [30] on top of our policy #41 in some cases can marginally push the performance even higher.

## V. CONCLUSION & FUTURE WORK

In this work, we provide many insights into the effectiveness of different augmentation techniques for LiDAR based 3D Object Detection. We demonstrate that our improved augmentation policy #41 transfers well to other 3D Object Detection methods and are not just limited to the KITTI [2] dataset. We hope that other practitioners can now take our findings off the shelf and apply them successfully to their work. Further, we uncovered how important data augmentation for LiDAR based 3D Object Detection really is and that it seemingly plays a significant role for many state of the art 3D Object Detection methods.

In future work, we want to investigate why *Ground Removal* can harm 3D Object Detection performance. One explanation for the deterioration could be that significant scene / neighborhood context is getting destroyed by removing the points belonging to the ground. However, more experiments are needed to test this hypothesis, *e.g.* by estimating the ground plane [43] or using semantic ground truth information provided by SemanticKITTI [44] or nuScenes-lidarseg [45].

## REFERENCES

- [1] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, “Pointpillars: Fast encoders for object detection from point clouds,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [2] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [3] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, “Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [4] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuScenes: A multimodal dataset for autonomous driving,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [5] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, C. T. A. Kazakova, L. Platin-sky, W. Jiang, and V. Shet, “Lyft Level5 Perception Dataset,” <https://level5.lyft.com/dataset/>, 2019, accessed: 2022-02-28.
- [6] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, “The ApolloScape dataset for autonomous driving,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, 2018.
- [7] M. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays, “Argoverse: 3D tracking and forecasting with rich maps,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [8] A. Patil, S. Malla, H. Gang, and Y.-T. Chen, “The H3D dataset for full-surround 3D multi-object detection and tracking in crowded urban scenes,” in *IEEE International Conference on Robotics and Automation*, 2019.
- [9] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Cai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, “Scalability in perception for autonomous driving: Waymo open dataset,” *arXiv preprint:1912.04838*, 2019.
- [10] “Lidar is getting cheaper,” <https://cmte.ieee.org/futuredirections/2019/07/16/lidar-is-getting-cheaper/>, accessed: 2022-02-28.
- [11] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, “3d packing for self-supervised monocular depth estimation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [12] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Weinberger, “Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving,” in *CVPR*, 2019.
- [13] D. Park, R. Ambrus, V. Guizilini, J. Li, and A. Gaidon, “Is Pseudo-LiDAR needed for monocular 3D object detection?” in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [14] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, “PV-RCNN: Point-voxel feature set abstraction for 3D object detection,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [15] B. Yang, W. Luo, and R. Urtasun, “PIXOR: Real-time 3D object detection from point clouds,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [16] Y. Zhou and O. Tuzel, “VoxelNet: End-to-end learning for point cloud based 3D object detection,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [17] Y. M. Y. Yan and B. Li, “SECOND: Sparsely embedded convolutional detection,” *Sensors*, vol. 18, no. 10, p. 3337, 2018.
- [18] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, “Multi-view 3d object detection network for autonomous driving,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [19] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, “Joint 3D proposal generation and object detection from view aggregation,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.
- [20] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

- [21] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [22] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, "From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.
- [23] S. Shi, L. Jiang, J. Deng, Z. Wang, C. Guo, J. Shi, X. Wang, and H. Li, "PV-RCNN++: point-voxel feature set abstraction with local vector representation for 3D object detection," *arXiv preprint:2102.00463*, 2021.
- [24] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3D object detection and tracking," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [25] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3D object detection in point clouds," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [26] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [27] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [28] J. Choi, Y. Song, and N. Kwak, "Part-Aware data augmentation for 3D object detection in point cloud," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.
- [29] L. J. Wu Zheng, Weiliang Tang and C.-W. Fu, "SE-SSD: Self-ensembling single-stage object detector from point cloud," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [30] J. S. Hu and S. L. Waslander, "Pattern-Aware data augmentation for lidar 3d object detection," in *IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021.
- [31] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential fusion for 3D object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [32] C. Wang, C. Ma, M. Zhu, and X. Yang, "PointAugmenting: Cross-modal augmentation for 3D object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [33] M. Hahner, C. Sakaridis, D. Dai, and L. Van Gool, "Fog Simulation on Real LiDAR Point Clouds for 3D Object Detection in Adverse Weather," in *IEEE International Conference on Computer Vision (ICCV)*, 2021.
- [34] M. Hahner, C. Sakaridis, M. Bijelic, F. Heide, F. Yu, D. Dai, and L. Van Gool, "LiDAR Snowfall Simulation for Robust 3D Object Detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [35] Y. Chen, V. T. Hu, E. Gavves, T. Mensink, P. Mettes, P. Yang, and C. G. M. Snoek, "PointMixup: Augmentation for point clouds," in *European Conference on Computer Vision (ECCV)*, 2020.
- [36] R. Li, X. Li, P.-A. Heng, and C.-W. Fu, "PointAugment: An auto-augmentation framework for point cloud classification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [37] S. Cheng, Z. Leng, E. D. Cubuk, B. R. Zoph, C. Bai, J. Ngiam, Y. Song, B. Caine, V. Vasudevan, C. Li, Q. V. Le, J. Shlens, and D. Anguelov, "Improving 3D object detection through progressive population based augmentation," in *European Conference on Computer Vision (ECCV)*, 2020.
- [38] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel R-CNN: Towards high performance voxel-based 3D object detection," in *AAAI Conference on Artificial Intelligence*, 2021.
- [39] A. Simonelli, S. R. Buló, L. Porzi, M. Lopez-Antequera, and P. Kotschieder, "Disentangling monocular 3D object detection," in *IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [40] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [41] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3D object proposals for accurate object class detection," in *Advances in Neural Information Processing Systems*, 2015.
- [42] J. Kabzan, M. d. l. I. Valls, V. Reijgwart, H. F. C. Hendriks, C. Ehmke, M. Prajapat, A. Bühler, N. Gosala, M. Gupta, R. Sivanesan *et al.*, "Amz driverless: The full autonomous racing system," *arXiv preprint:1905.05150*, 2019.
- [43] A. Paigwar, Ö. Er kent, D. Sierra González, and C. Laugier, "GndNet: Fast Ground Plane Estimation and Point Cloud Segmentation for Autonomous Vehicles," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.
- [44] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences," in *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019.
- [45] W. K. Fong, R. Mohan, J. V. Hurtado, L. Zhou, H. Caesar, O. Beijbom, and A. Valada, "Panoptic nuScenes: A large-scale benchmark for lidar panoptic segmentation and tracking," *arXiv preprint:2109.03805*, 2021.