

SAILOR: Scaling Anchors via Insights into Latent Object Representation

Dušan Malić^{1,2}Christian Fruhwirth-Reisinger^{1,2}Horst Possegger¹Horst Bischof^{1,2}¹Institute of Computer Graphics and Vision, Graz University of Technology²Christian Doppler Laboratory for Embedded Machine Learning

{dusan.malic, christian.reisinger, possegger, bischof}@icg.tugraz.at

Abstract

LiDAR 3D object detection models are inevitably biased towards their training dataset. The detector clearly exhibits this bias when employed on a target dataset, particularly towards object sizes. However, object sizes vary heavily between domains due to, for instance, different labeling policies or geographical locations. State-of-the-art unsupervised domain adaptation approaches outsource methods to overcome the object size bias. Mainstream size adaptation approaches exploit target domain statistics, contradicting the original unsupervised assumption. Our novel unsupervised anchor calibration method addresses this limitation. Given a model trained on the source data, we estimate the optimal target anchors in a completely unsupervised manner. The main idea stems from an intuitive observation: by varying the anchor sizes for the target domain, we inevitably introduce noise or even remove valuable object cues. The latent object representation, perturbed by the anchor size, is closest to the learned source features only under the optimal target anchors. We leverage this observation for anchor size optimization. Our experimental results show that, without any retraining, we achieve competitive results even compared to state-of-the-art weakly-supervised size adaptation approaches. In addition, our anchor calibration can be combined with such existing methods, making them completely unsupervised.

1. Introduction

Acquiring and labeling data for the training of 3D object detectors requires considerable effort. The sheer size and the underlying unorganized structure of point clouds make this process cumbersome. Detecting objects in a point cloud can be difficult even for humans since an object may contain just a few points. Moreover, during labeling, an expert will usually examine a single LiDAR frame from multiple viewpoints to account for unavoidable occlusions and trun-

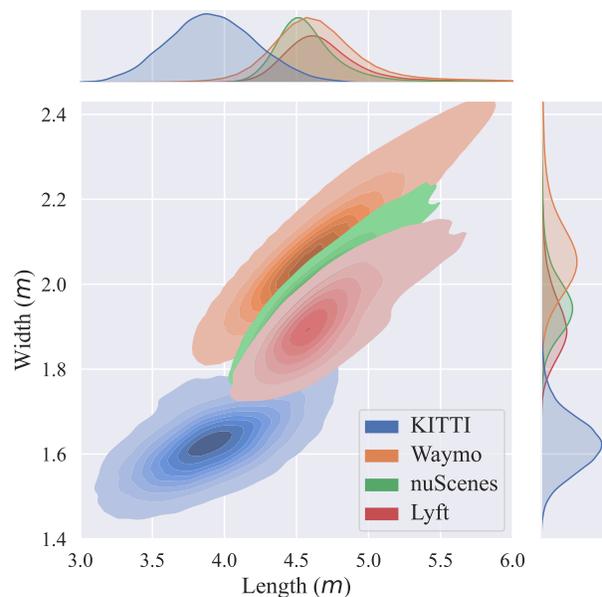


Figure 1: Box sizes, showing the domain gap across datasets. Boxes are generated by a fully-supervised oracle model trained on KITTI [8], Waymo [20], nuScenes [1] and Lyft [11], respectively.

cations, usually switching the context between an image and a point cloud. These interruptions substantially increase labeling time and, consecutively, labeling cost.

To mitigate the high labeling cost, our research community is making immense advances in unsupervised 3D domain adaptation [15, 23, 31]. State-of-the-art approaches typically rely on some variant of self-training [30, 31], input transformation [21, 24, 27], feature alignment [35] and/or tracking [7, 34]. These approaches share a common problem, initially reported by Wang *et al.* [23]: the discrepancy between object sizes, as depicted in Figure 1, introduces a

tremendous domain gap.

In 2D object detection, object sizes vary depending on the distance of an object from a sensor. This naturally occurring augmentation increases a detector’s robustness towards size variation. In 3D object detection, however, the size of an object does not correlate with the distance from the sensor. Instead, it depends on where the dataset has been acquired, *e.g.* on average, vehicles in the USA are larger than the vehicles in Europe [23], or on the labeling policy, *e.g.* Waymo Open Dataset [20] includes side mirrors of cars in its annotations, whereas KITTI [8] does not.

Object detectors are commonly evaluated using the Mean Average Precision (mAP) metric, which further utilizes Intersection over Union (IoU) for ground truth matching. In 2D detection, IoU is calculated given the area of the predicted and ground truth bounding box, whereas in 3D, it is derived using the volumes of the predicted and the ground truth cuboid. Given the additional dimension, the overall influence of the incorrect size prediction leads to an exponential accuracy decrease. This correlation makes 3D detectors extremely volatile to changes in the object sizes.

Statistical Normalization (SN) [23] has become the standard approach for bridging the size gap. It attempts to shift the source data statistics to the target statistics through deliberate scaling of the source annotations as a training augmentation. Random Object Scaling (ROS) [31] strives to overcome this size bias without directed scaling. Instead, it substantially augments the ground truth boxes via a wider range of scales. Nevertheless, both approaches exploit key target domain insights, which are usually not available in an unsupervised setting. First, they employ anchor sizes, manually optimized for the target domain. Second, they manually refine the magnitude of the augmentation scales. Lastly, the stochastic nature of these approaches does not allow a deterministic checkpoint selection, but rather the best performing one.

We propose SAILOR, a novel *unsupervised anchor calibration pipeline* which estimates optimal target anchors given a pretrained source model. Our objective is to identify target anchors under which the feature similarity between a source and a target domain is maximized. We first establish a reference feature database by gathering instance features from the source domain. In the next phase, we iteratively perturb the anchors by a small amount, and analogously compute a target feature database. The fitness of the target feature database to the reference feature database provides a stochastic gradient, which we employ in a stochastic optimization method. Our method achieves similar improvements as SN or ROS, but does not introduce additional model parameters, does not require retraining, and does not require any knowledge from the target domain at all.

Experimental results on the autonomous driving datasets KITTI [8], Waymo [20], nuScenes [1] and Lyft [11] indicate

large performance gains with minimal effort. We demonstrate that a simple exchange of the source anchors with our optimized ones, leads to large precision gains on the target domain. Our method is competitive even with the popular weakly-supervised approaches for bridging the object size domain gap, but costs only a fraction of the computation time, since we do not require retraining. Additionally, our optimized anchors can also be used as an unsupervised prior to extend weakly-supervised methods, turning them into fully unsupervised approaches.

2. Related Work

3D Object Detection State-of-the-art 3D detectors consist of a Region Proposal Network (RPN) followed by a detection head. RPN first abstracts an input point cloud with a feature extractor, which is commonly point-based [17], voxel-based [18, 28] or a hybrid between these two [16]. Afterwards, at each discrete location of the abstracted point cloud, RPN predicts an objectness probability and regresses coarse bounding boxes. The magnitude of this regression is either absolute [33] or relative to *anchor* values [5, 12, 16, 18, 28]. The anchors reduce the network’s regression space in such a way that the predictions become residuals from a matched anchor. The anchor values are commonly derived from the training data and have to be manually selected with care. The following detection head refines such coarse predictions into the final predictions.

Unsupervised 3D Domain Adaptation A 3D detection model trained on one dataset usually does not generalize well to new, unseen data. The source and target dataset difference is commonly referred to as the *domain gap*. Unsupervised domain adaptation aims to reduce the gap without any additional annotations.

Style-transfer approaches lessen the gap in the input space by transforming the target data into source-like data. Wei *et al.* [24] demonstrate the generalization capabilities of a model trained on dense data augmented by pseudo-sparse point clouds derived from the original dense data. Dropping LiDAR beams scales up with the size of a point cloud since it is not a costly operation. However, it is ambiguous to what extent the background transformation helps and if it justifies a more complex conversion. Modifying just the foreground objects by adding new points [21, 27] or reorganizing the existing ones [13] saves computational resources with a similar effect.

The sequential acquisition of autonomous driving datasets allows exploiting temporal consistency for the adaptation. Some self-training approaches [15, 29, 34] refine per-frame detection via Multi Object Tracking (MOT) [25] and are thus able to improve naïve retraining by reducing the false positive and negative rate. Even without explicit tracking, aggregating several LiDAR frames into a

common reference frame already leads to denser static objects. Detecting such objects is a trivial task for state-of-the-art 3D detectors [3, 16, 28]. By propagating high-quality detections to all aggregated frames, [22] generates pseudo labels for even the more complex (sparse and static) samples. FAST3D [7] further extends this idea to moving objects, by leveraging scene flow [26] for aggregation.

Besides tracking and flow pitfalls, *e.g.* association, domain gap, *etc.*, temporal information may not always be available. Generating pseudo labels on a per-frame basis has also been beneficial for self-training [2]. Leveraging the gradual improvement on the presented pseudo labels, memory ensembles [30, 31] enhance the robustness by cataloging the predictions during training. An auxiliary loss [32, 35] can further reduce the inevitable noise induced by the pseudo labels. Still, the self-training is going to amplify the noise. This is a well-known side effect called *inductive bias*, which is generally resolved with the student-teacher paradigm [10, 14].

The self-training approaches are not able to independently overcome the object size bias induced by the source data. The extent of the object size gap immensely influences the performance of the final model. Therefore, state-of-the-art self-training approaches explicitly utilize mechanisms to mitigate the cross-domain size mismatch.

Overcoming Object Size Bias Mainstream methods exploit target domain knowledge to overcome the object size bias. Wang *et al.* [23] propose employing the target domain statistics in two ways. Output Transformation (OT) directly modifies a predicted bounding box size by adding a residual. The difference between an average source and a target sample defines this residual. Statistical Normalization (SN) scales source ground truth boxes and points inside during training. Analogously to OT, the scaled intensity is derived from the source and target dataset statistics. To avoid precise target domain statistics that are usually unavailable, Random Object Scaling (ROS) [31] implements heavy size augmentation on the source ground truth objects. However, increasing the network’s search space leads to a frequent size prediction mismatch.

Contrary to the weakly-supervised approaches, we propose a completely unsupervised method, which specifically optimizes a model’s anchors for the target domain. Moreover, our approach does not require retraining and thus costs just a fraction of computational time. We achieve a similar effect as these weakly-supervised methods, since more suitable target anchors alleviate the model’s extrapolation efforts and implicitly provide improved detections.

3. Unsupervised Anchor Calibration

We consider a 3D detection model $H(\cdot | \Theta) = D \circ RPN$, with a Region Proposal Network RPN and a de-

tection head D . Its parameters Θ are trained on a labeled source dataset $\mathcal{S} = \{(\mathbf{P}_i^s, \mathbf{Y}_i^s)\}_{i=1}^{N_s}$, which contains LiDAR point clouds \mathbf{P}^s and a respective set of labeled instances \mathbf{Y}^s . The target dataset $\mathcal{T} = \{\mathbf{P}_i^t\}_{i=1}^{N_t}$, however, contains only unlabeled point clouds. Besides the trainable parameters, the model is additionally defined via its hyperparameters. The anchors $\Psi = \{\psi^{(x)}, \psi^{(y)}, \psi^{(z)}, \psi^{(w)}, \psi^{(l)}, \psi^{(h)}, \psi^{(\theta)}\} \subset \Theta$ provide a starting point for the regression head, where the network predicts residuals relative to an anchor, instead of regressing the absolute values. The anchors’ width, length and height are usually handpicked to match the average size of annotated objects from \mathcal{S} .

The size discrepancy between objects in \mathcal{S} and \mathcal{T} induces an apparent domain gap, as illustrated in Figure 1. With our *unsupervised anchor calibration*, we adapt the anchor sizes to the target dataset \mathcal{T} without any supervision. As depicted in Figure 2, using a model $H(\cdot | \Theta)$ pretrained on the source data, we first construct a reference feature database (Section 3.1) by accumulating the proposal feature vectors from the source domain \mathcal{S} . Then, we iteratively perturb the model’s anchor sizes by a small amount ε , while leaving *all* other parameters unchanged ($\psi^{(x)}, \psi^{(y)}, \psi^{(z)}, \psi^{(w)} + \varepsilon^{(w)}, \psi^{(l)} + \varepsilon^{(l)}, \psi^{(h)} + \varepsilon^{(h)}, \psi^{(\theta)}$) and, analogous to the reference feature database, compute a target feature database from \mathcal{T} . The fitness of such a target feature database to the source feature database (Section 3.2) provides a stochastic gradient, which we utilize to adapt the model’s anchor sizes (Section 3.3). Our approach yields anchors which are specifically tailored to the given model for the target domain, without any retraining.

3.1. Reference Feature Database

Using the model $H(\cdot | \Theta)$, pretrained on the source data \mathcal{S} , we first generate a reference feature database

$$\mathcal{F}^s = \{RPN(\mathbf{P}^s | \Theta) \delta_{[\hat{c} > \tau]} | \forall \mathbf{P}^s \in \mathcal{S}\}. \quad (1)$$

We select the RPN features of the frame \mathbf{P}^s where the respective prediction score \hat{c} , obtained from the classification head, exceeds a certain threshold τ . Ultimately, \mathcal{F}^s is a latent feature database of the source database samples.

Depending on the size of the source dataset, \mathcal{F}^s can have a large memory footprint. Therefore, to compress a potentially abundant reference feature database we fit a Gaussian Mixture Model (GMM) to the data. A GMM consists of K weighted Gaussian distributions, where the probability of observing a sample f is defined as

$$p(f | \Omega) = \sum_{i=1}^K \omega_i \cdot \mathcal{N}(f | \mu_i, \Sigma_i), \quad \text{where} \quad (2)$$

$$\mathcal{N}(f | \mu_i, \Sigma_i) = \frac{\exp\left(-\frac{1}{2}(f - \mu_i)^T \Sigma_i^{-1} (f - \mu_i)\right)}{\sqrt{(2\pi)^K |\Sigma_i|}}. \quad (3)$$

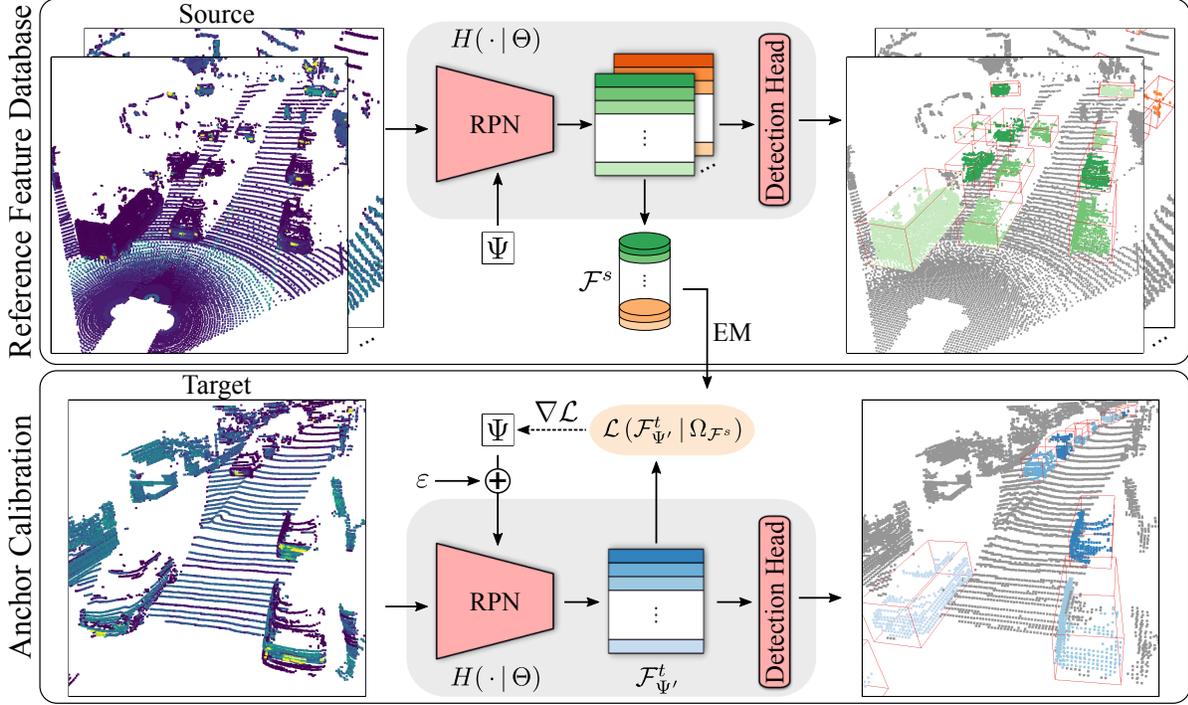


Figure 2: Overview of SAILOR: Given a model $H(\cdot | \Theta)$ with parameters Θ trained on the source data and anchors $\Psi \subset \Theta$ optimized for the same data, we first extract a source feature database \mathcal{F}^s by accumulating Region Proposal Network’s (RPN) predictions. We then fit a Gaussian Mixture Model (GMM) parameters $\Omega_{\mathcal{F}^s}$ to \mathcal{F}^s using Expectation Maximization (EM) [4]. During anchor calibration on the target domain, using the same model, we iteratively introduce small noise ε to the source anchors and extract a temporary target feature database \mathcal{F}_{Ψ}^t . We utilize a stochastic gradient $\nabla \mathcal{L}$, estimated from the fitness $\mathcal{L}(\mathcal{F}_{\Psi}^t | \Omega_{\mathcal{F}^s})$, to directly adapt the source anchors to the target domain. We neither perform any further adaption, nor require any retraining.

Here, ω_i weighs the i^{th} Gaussian distribution, which is defined by its mean vector μ_i and covariance matrix Σ_i . We fit the GMM parameters $\Omega_{\mathcal{F}^s} = \{(\omega_i, \mu_i, \Sigma_i)\}_{i=1}^K$ to a source feature database \mathcal{F}^s using Expectation-Maximization [4].

3.2. Target Fitness Quantification

The reference probability model $\Omega_{\mathcal{F}^s}$ describes ideal latent instances, *i.e.* high-dimensional feature vectors that are abstractions of a complete object without background noise. The inference on the target data, using the source model, will inevitably lead to predictions which include background noise or do not contain the object completely. However, we can reduce such noisy predictions by selecting adequate anchor sizes. In this section, we show how we quantify the quality of the selected anchors.

Using the source model $H(\cdot | \Theta)$, we compute the target database \mathcal{F}^t from the target dataset \mathcal{T} , analogous to Equation (1). However, during this inference, we ignore the network’s size residual predictions ($\Delta^{(w)}, \Delta^{(l)}, \Delta^{(h)}$) and use only matched anchor sizes, *i.e.* $(\psi^{(x)} + \Delta^{(x)}, \psi^{(y)} +$

$\Delta^{(y)}, \psi^{(z)} + \Delta^{(z)}, \psi^{(w)}, \psi^{(l)}, \psi^{(h)}, \psi^{(\theta)} + \Delta^{(\theta)})$. This isolates the influence of the selected anchors, because otherwise, the consecutive regression stage of the detector would obfuscate the changes.

To quantify the fitness of the anchors, we leverage the joint probability density

$$p(\mathcal{F}^t | \Omega_{\mathcal{F}^s}) = \prod_{f_i^t \in \mathcal{F}^t} p(f_i^t | \Omega_{\mathcal{F}^s}), \quad (4)$$

where $\Omega_{\mathcal{F}^s}$ are the parameters of the reference probability model and the features f_i^t are independent and identically distributed. To avoid numerical instabilities we do not explicitly employ the joint probability but instead use the per-sample average log-likelihood

$$\mathcal{L}(\mathcal{F}^t | \Omega_{\mathcal{F}^s}) = \frac{1}{|\mathcal{F}^t|} \sum_{f_i^t \in \mathcal{F}^t} \log p(f_i^t | \Omega_{\mathcal{F}^s}). \quad (5)$$

The sum of the logarithms is numerically stable and the cardinality normalization ensures independence of the number of target features.

Smaller values of \mathcal{L} indicate unfit predictions, *e.g.* if target features include more background clutter or less object cues compared to the source features, and is maximized for anchors which are optimal for the target domain. We present an experiment which demonstrates this behavior in Figure 3.

3.3. Anchor Calibration

Using the reference feature database \mathcal{F}^s and fitness quantification \mathcal{L} , we can now compute the optimal anchors Ψ^* for a model $H(\cdot | \Theta)$ by optimizing

$$\Psi^* = \arg \min_{\Psi} -\mathcal{L}(\mathcal{F}_{\Psi}^t | \Omega_{\mathcal{F}^s}), \quad (6)$$

where $\Psi^* \subset \Theta$ represents the optimal anchors for the target dataset \mathcal{T} . We outline our approach in Algorithm 1.

Algorithm 1 Pseudocode of our SAILOR method

Input: 3D detection model $H(\cdot | \Theta)$ and anchors Ψ optimized on a source dataset \mathcal{S} ; Labeled source \mathcal{S} and unlabeled target \mathcal{T} dataset

Output: Optimized anchors Ψ^* for \mathcal{T}

Generate \mathcal{F}^s ▷ Section 3.1

Fit GMM parameters $\Omega_{\mathcal{F}^s}$ to \mathcal{F}^s with EM

$\Psi^* = \Psi$

while Termination criteria is not reached **do**

 Randomly sample small ε

$\Psi' = \Psi^* + \varepsilon$ ▷ Anchor size perturbation

 Generate $\mathcal{F}_{\Psi'}^t$ ▷ Section 3.2

 Update Ψ^* with $\nabla \mathcal{L}(\mathcal{F}_{\Psi'}^t | \Omega_{\mathcal{F}^s})$ ▷ Section 3.3

end while

Since Equation (6) is not differentiable w.r.t. the model hyperparameters, we do not use standard gradient methods. Instead, we utilize Differential Evolution (DE) [19], a stochastic optimization technique. At each iteration, given a population vector, DE constructs a mutation vector. A trial candidate is constructed in a crossover phase by mixing the mutation vector with a candidate solution. If the fitness of the trial candidate exceeds the current candidate solution, it becomes the next candidate solution. This is repeated until a convergence criterion is met.

To accelerate the optimization, we first perform a linear sweep for each parameter separately. Figure 3 depicts an example of this search. We freeze other anchor sizes, and while varying a single parameter, assess the fitness at each step. We use the parameter with the highest overall fitness as the initial candidate solution for the following joint optimization phase. The reduced search space expedites the final joint fine-tuning.

Starting from this initial candidate solution Ψ^* , we jointly optimize the length, width and height. For this, we generate the initial population vector by uniformly sampling

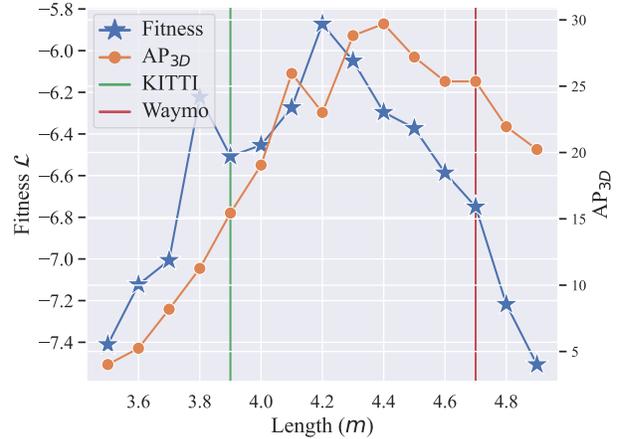


Figure 3: Comparing anchor fitness (Equation (5)) and Average Precision. The model is trained on Waymo [20], whereas the fitness and AP are assessed on KITTI [8]. Here, we fix width and height of the anchors and vary only the length. The fitness is computed without any annotation in the target domain and strongly correlates with the AP_{3D} . We denote ground truth Waymo and KITTI anchors in red and green, respectively.

N_p anchor values $\Psi = \{\Psi_i\}_{i=1}^{N_p}$, where the sampling range is a percentage of the source anchor value. A mutation vector is constructed as

$$\Psi' = \Psi_0 + \eta \cdot (\Psi_{r_1} - \Psi_{r_2}), \quad (7)$$

where Ψ_0 is the population vector member with the best fitness score, r_1 and r_2 are two randomly selected indices, and η is the mutation amplitude constant. We terminate the optimization when the loss converges or when we reach a maximum number of iterations.

4. Experiments

We focus our evaluation on KITTI [8], Waymo [20] and nuScenes [1], and include Lyft [11] for further generalization evaluations. Inter-domain anchor calibration demonstrates the capability, whereas intra-domain calibration verifies the correctness of our approach. The variety of the data, *e.g.* sparse to dense (nuScenes \leftrightarrow Waymo) or a large size gap (KITTI \leftrightarrow Waymo), faithfully represents an arbitrary use-case. The source and target data are the training and validation splits of the datasets mentioned above, respectively. We always perform the evaluation following the evaluation protocol of the target dataset.

4.1. Comparison with the State-of-the-Art

To the best of our knowledge, there is no existing work in the unsupervised domain to compare to. Therefore,

Source	Method	Target		
		KITTI	Waymo	nuScenes
		AP _{3D} @R11 Car / Pedestrian / Cyclist	L1_AP Vehicle / Pedestrian / Cyclist	mAP Car / Pedestrian / Bicycle
KITTI	Source Anchors	78.59 / <u>54.58</u> / 64.53	3.55 / 8.69 / 9.93	14.66 / 0.0 / 0.0
	SN [†]	-	2.80 / 2.05 / 1.65	20.33 / 0.0 / 0.0
	OT [†]	-	6.51 / 13.94 / 16.81	14.65 / 0.0 / 0.0
	ROS [†]	-	3.32 / 2.75 / 2.29	<u>18.54</u> / 0.0 / 0.0
	Target Anchors [†]	-	11.69 / <u>10.40</u> / <u>12.15</u>	15.88 / 0.0 / 0.0
	Ours	<u>77.79</u> / 55.49 / <u>63.52</u>	<u>8.26</u> / 8.87 / 9.52	16.12 / 0.0 / 0.0
Waymo	Source Anchors	23.94 / 59.99 / 52.32	74.84 / 71.59 / 66.49	32.45 / 17.99 / 0.0
	SN [†]	24.11 / 63.51 / 52.83	-	34.05 / 17.24 / 0.0
	OT [†]	38.79 / 53.31 / 61.82	-	37.40 / <u>19.99</u> / 0.0
	ROS [†]	<u>43.00</u> / <u>62.39</u> / 51.09	-	35.91 / 18.62 / 0.0
	Target Anchors [†]	39.69 / 49.96 / <u>55.45</u>	-	<u>37.30</u> / 20.32 / 0.0
	Ours	58.02 / 61.60 / 53.04	<u>73.98</u> / <u>69.48</u> / <u>66.19</u>	32.34 / 17.27 / 0.0
nuScenes	Source Anchors	26.37 / 10.37 / <u>22.90</u>	29.59 / 2.99 / 2.66	59.93 / 11.03 / 0.23
	SN [†]	33.67 / 31.67 / 14.29	34.48 / 2.34 / <u>4.08</u>	-
	OT [†]	36.42 / 18.15 / 8.84	31.37 / 8.22 / 0.13	-
	ROS [†]	<u>56.28</u> / <u>26.88</u> / 19.76	<u>32.10</u> / <u>7.73</u> / 7.03	-
	Target Anchors [†]	57.93 / 18.57 / 11.68	31.70 / 6.69 / 0.41	-
	Ours	55.10 / 4.64 / 26.37	29.00 / 2.07 / 2.81	<u>59.41</u> / <u>10.98</u> / <u>0.12</u>

Table 1: Extensive comparison of Statistical Normalization (SN) [23], Output Transformation (OT) [23] and Random Object Scaling (ROS) [31] with our method on popular autonomous driving datasets KITTI [8], Waymo [20] and nuScenes [1]. We use the whole KITTI and nuScenes dataset for both training and evaluation and utilize 20% of Waymo v1.2 for training and full evaluation dataset. The model is trained on the source data using the configuration as in OpenPCDet [6]. We always evaluate the *last checkpoint* using the target dataset evaluation pipeline. For KITTI, we show moderate case of AP_{3D}@R11 for Car / Pedestrian / Cyclist thresholded at 0.7, 0.5 and 0.5, respectively. We report L1_AP for Vehicle / Pedestrian / Cyclist on Waymo, whereas for nuScenes, we provide mAP for the classes Car / Pedestrian / Bicycle. We mark weakly-supervised approaches with [†], since they require target object statistics.

we compare SAILOR¹ with the widely adapted weakly-supervised approaches that address the object size bias, *i.e.* Statistical Normalization (SN) [23], Output Transformation (OT) [23] and Random Object Scaling (ROS) [31].

For SN, we follow the original publication and set the model anchors to the average of the source and target domain and scale the labeled source bounding boxes and points inside using the difference between the statistics. Similarly, ROS exploits this knowledge, but in a more coarse way. Knowing that KITTI objects are smaller than Waymo’s, we can apply the appropriate scaling for ROS during the adaptation. We retrain both models according to the source configuration. OT does not require training but instead uses the source model and directly adds the dif-

ference to the predictions. We refer the reader to the supplementary material, where we list the exact anchor configuration for reproducibility.

During the evaluation, we directly employ the anchors optimized by SAILOR. We do not perform any retraining whatsoever. Table 1 demonstrates the potential of our method, given a Part-A² source model. In the case of Waymo → KITTI, we observe an improvement of 34 AP_{3D}@R11 for the Car class, beating even the weakly-supervised approaches by a large margin. We depict this improvement in Figure 4. Pedestrian and Cyclists are mostly unchanged, except for a slight improvement, mainly due to having similar anchors. We report similar behavior in the nuScenes → KITTI experiment, where we observe an improvement of almost 28 AP_{3D}@R11. We also note that the slight precision increase for the Cyclists class stems from

¹<https://github.com/maliced/sailor>

Task	Method	AP _{3D} @R11
Waymo → Lyft	Source Anchors	49.88 / 37.17 / 25.62
	SN [†]	44.96 / 33.98 / 17.00
	OT [†]	51.58 / 37.96 / 15.32
	ROS [†]	49.39 / 34.06 / 25.07
	Target Anchors [†]	51.57 / 37.02 / 16.15
	Ours	49.31 / 38.27 / 26.01
KITTI → Lyft	Source Anchors	20.61 / 13.94 / 14.24
	SN [†]	28.27 / 13.36 / 10.81
	OT [†]	20.83 / 16.83 / 10.40
	ROS [†]	25.86 / 16.33 / 8.05
	Target Anchors [†]	26.12 / 17.24 / 17.24
	Ours	26.48 / 17.27 / 14.20

Table 2: Results of the adaptation tasks to Lyft. We report AP_{3D}@R11 for the classes Car / Pedestrian / Bicycle.

the significant height difference (due to different labeling policies) between the two datasets.

In situations where a source model shows inferior performance and the size domain gap is large, *e.g.* KITTI → Waymo and KITTI → nuScenes, we still accomplish substantial relative gain. Calibrating KITTI anchors on Waymo, we achieve a relative improvement over 200%. This tremendous increase is significantly better than the weakly-supervised SN and ROS, which even degrade the performance slightly. Note that the KITTI → Waymo scenario is well-known to be extremely challenging and has thus been often neglected from evaluations, with very few exceptions, *e.g.* [7]. The overall low performance on KITTI → nuScenes stems from a small source dataset and the sparsity of the nuScenes point clouds, where the detector struggles. Thus, there are only very few objects with insufficient latent semantics to apply our method.

The datasets that share similar anchors, *e.g.* Waymo ↔ nuScenes, do not exhibit a substantial change in the overall evaluation score. Since SAILOR introduces no additional hyperparameters, the weakly-supervised approaches perform favorable in this scenario, as they can exploit the target domain statistics. However, note that similar improvements can also be achieved by fiercer augmentation, as shown in [9].

Similarly, when our source and target data come from the same dataset, *i.e.* the diagonal in Table 1, we report an insignificant change in the results. We performed this additional experiment mainly as a sanity check.

To further demonstrate the generalization capabilities of our method, we perform additional experiments on the Lyft [11] dataset. Our findings, which we summarize in Table 2, confirm our initial experiments. In cases where the source and target anchors are similar, *e.g.* Waymo → Lyft,

Task	Method	AP _{3D} @R11
Waymo → KITTI	SN [†]	24.11 / 63.51 / 52.83
	ROS [†]	43.00 / 62.39 / 51.09
	SN w/ SAILOR	22.31 / 55.68 / 38.96
	ROS w/ SAILOR	39.82 / 60.40 / 48.61
nuScenes → KITTI	SN [†]	33.67 / 31.67 / 14.29
	ROS [†]	56.28 / 26.88 / 19.76
	SN w/ SAILOR	39.44 / 0.51 / 0.00
	ROS w/ SAILOR	57.98 / 34.59 / 24.15

Table 3: SN and ROS become entirely unsupervised in combination with SAILOR, otherwise they are weakly-supervised and denoted with [†]. The results are the moderate case for Car / Pedestrian / Cyclist classes.

we report slight performance gains. Additionally, when the object size gap is large, *e.g.* KITTI → Lyft for class car, we report increase of around 6 AP points. Similarly to KITTI → nuScenes, we found that further improvement is limited by an ineffective source model caused by a small source dataset. However, due to sufficient density of the Lyft point clouds (LiDAR with 64 beams), we do not observe any precision drop.

Moreover, we demonstrate that combining our method with existing weakly-supervised approaches leads to competitive results in a completely unsupervised manner. For this, we replace the target statistics required by SN and ROS by the anchor sizes calibrated with SAILOR. The results in Table 3 show that this is well suited to make ROS fully unsupervised, whereas for SN it works well for vehicles, but is not suitable for pedestrians and cyclists. This issue is due to SN, which is tailored explicitly to knowing the target statistics and requires additional manual fine-tuning (which we omitted due to the unsupervised setting of this experiment) to achieve the best results. Note that we are not interested in estimating the actual target domain object sizes, but in estimating the optimal target domain anchors for the given source model. Using these calibrated anchors to guide ROS, our unsupervised ROS variant performs on par with the weakly-supervised one, *e.g.* Waymo → KITTI, or even better, *e.g.* nuScenes → KITTI.

4.2. Ablation Studies

We conduct our ablation experiments to investigate the interaction between the components of our system, as well as to study the vital points of our pipeline. We also indicate the known caveats and demonstrate how we overcome them.

Joint Optimization We demonstrate the benefits of our joint optimization in Table 4, which shows that linear search already provides a decent performance. However, it opti-

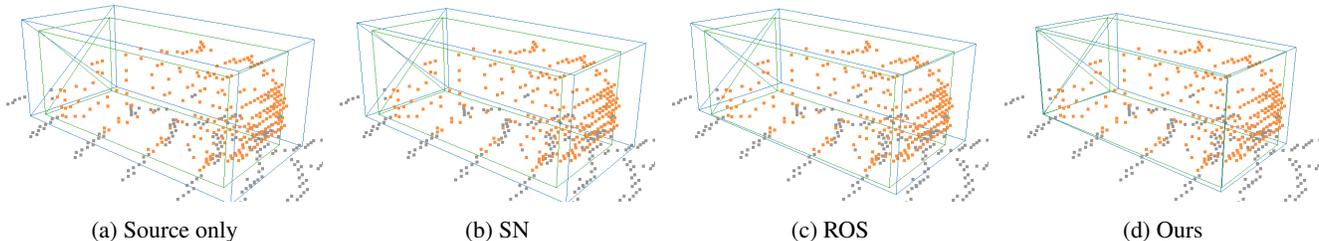


Figure 4: Qualitative comparison of source only, Statistical Normalization (SN) [23], Random Object Scaling (ROS) [31] and our method on Waymo \rightarrow KITTI case. We indicate the ground truth box in green and the predicted boxes in blue. The object points, according to the ground truth annotation, are shown in orange. Best viewed on screen.

Task	Method	$AP_{3D}@R11$
Waymo \rightarrow KITTI	Source	23.94 / 59.99 / 52.32
	LS	50.57 / 58.27 / 51.49
	LS + DE	58.02 / 61.59 / 53.04
nuScenes \rightarrow KITTI	Source	26.37 / 10.37 / 22.90
	LS	36.85 / 10.61 / 25.35
	LS + DE	55.10 / 4.64 / 26.37

Table 4: Detection performance w.r.t. the optimization strategy for the classes Car / Pedestrian / Cyclist. **LS** stands for Linear Search and **DE** is Differential Evolution.

mizes each component individually, disregarding their entanglement. We leverage this even further to boost the performance using joint optimization from Section 3.3.

In some instances, *e.g.* nuScenes \rightarrow KITTI for the class Pedestrian, DE degrades the result. Our observation shows that this happens for two reasons. When the number of instances in the source domain is too low for the proper estimation of the GMM parameters, *e.g.* KITTI \rightarrow Waymo for Pedestrians and Cyclists, our objective function becomes harder to optimize due to the appearing local minima. A similar effect can be observed by inadequate latent representations of the source model, caused by the point cloud sparsity, *e.g.* nuScenes \rightarrow KITTI for Pedestrians.

Smoothness of the Optimization Objective As previously hinted, the number of samples used for the GMM fitting plays an important role in our method. Our objective function from Equation (5) gets smoother with the number of instances we use. If we build the source feature database from underrepresented samples, it becomes hard to optimize many appearing local minima. This implicitly affects the performance on the target dataset, as depicted in Figure 5. On the other hand, SAILOR is not sensitive to the number of components in the GMM at all. We found that any reasonable choice of $K \geq 4$ works similarly well.

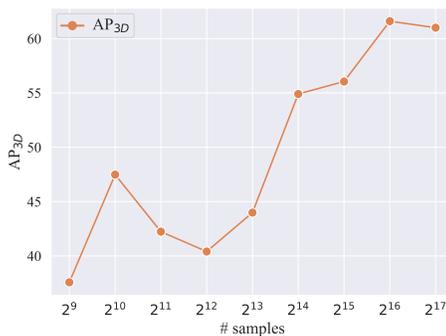


Figure 5: Applying SAILOR on Waymo \rightarrow KITTI while varying the number of instances used for GMM fitting. We report Average Precision for the moderate case of cars.

5. Conclusion

We presented SAILOR, an unsupervised approach for anchor calibration on the target domain. We estimate an optimal anchor configuration under the source model without prior knowledge. We compare our approach with weakly-supervised methods, which are widely used for unsupervised 3D domain adaptation. Moreover, SAILOR can be used as a stand-alone method or can make these weakly-supervised approaches completely unsupervised.

In this work, we focus explicitly on optimizing anchor sizes due to the tremendous domain gap across datasets, particularly KITTI \leftrightarrow Waymo. Note, however, that any model hyperparameter is optimizable with our SAILOR schema, even the checkpoint selection procedure, which we want to investigate in future experiments.

Acknowledgments We gratefully acknowledge the financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association. The presented experiments have been achieved (in part) using the Vienna Scientific Cluster.

References

- [1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proc. CVPR*, 2020.
- [2] Benjamin Caine, Rebecca Roelofs, Vijay Vasudevan, Jiquan Ngiam, Yuning Chai, Zhifeng Chen, and Jonathon Shlens. Pseudo-labeling for Scalable 3D Object Detection. *arXiv CoRR*, abs/2103.02093, 2021.
- [3] Yukang Chen, Yanwei Li, Xiangyu Zhang, Jian Sun, and Jiaya Jia. Focal Sparse Convolutional Networks for 3D Object Detection. In *Proc. CVPR*, 2022.
- [4] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *J. R. Stat. Soc.*, 39(1):1–22, 1977.
- [5] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel R-CNN: Towards High Performance Voxel-based 3D Object Detection. In *Proc. AAAI*, 2021.
- [6] OpenPCDet Development Team. OpenPCDet: An Open-source Toolbox for 3D Object Detection from Point Clouds. <https://github.com/open-mmlab/OpenPCDet>, 2020.
- [7] Christian Fruhwirth-Reisinger, Michael Opitz, Horst Possegger, and Horst Bischof. FAST3D: Flow-Aware Self-Training for 3D Object Detectors. In *Proc. BMVC*, 2021.
- [8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proc. CVPR*, 2012.
- [9] Martin Hahner, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Quantifying Data Augmentation for LiDAR based 3D Object Detection. *arXiv CoRR*, abs/2004.01643, 2020.
- [10] Deepti Hegde, Vishwanath Sindagi, Velat Kilic, A Brinton Cooper, Mark Foster, and Vishal Patel. Uncertainty-aware Mean Teacher for Source-free Unsupervised Domain Adaptive 3D Object Detection. *arXiv CoRR*, abs/2109.14651, 2021.
- [11] 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, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. Level 5 Perception Dataset 2020. <https://level-5.global/level5/data/>, 2019.
- [12] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast Encoders for Object Detection from Point Clouds. In *Proc. CVPR*, 2019.
- [13] Alexander Lehner, Stefano Gasperini, Alvaro Marcos-Ramiro, Michael Schmidt, Mohammad-Ali Nikouei Mahani, Nassir Navab, Benjamin Busam, and Federico Tombari. 3D-VField: Adversarial Augmentation of Point Clouds for Domain Generalization in 3D Object Detection. In *Proc. CVPR*, 2022.
- [14] Zhipeng Luo, Zhongang Cai, Changqing Zhou, Gongjie Zhang, Haiyu Zhao, Shuai Yi, Shijian Lu, Hongsheng Li, Shanghang Zhang, and Ziwei Liu. Unsupervised Domain Adaptive 3D Detection with Multi-Level Consistency. In *Proc. ICCV*, 2021.
- [15] Cristiano Saltori, Stéphane Lathuilière, Nicu Sebe, Elisa Ricci, and Fabio Galasso. SF-UDA^{3D}: Source-Free Unsupervised Domain Adaptation for LiDAR-based 3D Object Detection. In *Proc. 3DV*, 2020.
- [16] 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 *Proc. CVPR*, 2020.
- [17] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud. In *Proc. CVPR*, 2019.
- [18] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From Points to Parts: 3D Object Detection from Point Cloud with Part-aware and Part-aggregation Network. *TPAMI*, 43(8), 2021.
- [19] Rainer Storm and Kenneth Price. Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *J. Glob. Optim.*, 11(4):341–359, 1997.
- [20] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Etinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *Proc. CVPR*, 2020.
- [21] Darren Tsai, Julie Stephany Berrio, Mao Shan, Stewart Worral, and Eduardo Nebot. See Eye to Eye: A Lidar-Agnostic 3D Detection Framework for Unsupervised Multi-Target Domain Adaptation. *arXiv CoRR*, abs/2111.09450, 2021.
- [22] Sean Walsh, Jason Ku, Alex D. Pon, and Steven L. Waslander. Leveraging Temporal Data for Automatic Labelling of Static Vehicles. In *Proc. CVPR*, 2020.
- [23] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Weiliun Chao. Train in Germany, Test in The USA: Making 3D Object Detectors Generalize. In *Proc. CVPR*, 2020.
- [24] Yi Wei, Zibu Wei, Yongming Rao, Jiaxin Li, Jie Zhou, and Jiwen Lu. LiDAR Distillation: Bridging the Beam-Induced Domain Gap for 3D Object Detection. *arXiv CoRR*, abs/2203.14956, 2022.
- [25] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. In *Proc. IROS*, 2020.
- [26] Wenxuan Wu, Zhi Yuan Wang, Zhuwen Li, Wei Liu, and Li Fuxin. PointPWC-Net: Cost Volume on Point Clouds for (Self-) Supervised Scene Flow Estimation. In *Proc. ECCV*, 2020.
- [27] Qiangeng Xu, Yin Zhou, Weiyue Wang, Charles R Qi, and Dragomir Anguelov. SPG: Unsupervised Domain Adaptation for 3D Object Detection via Semantic Point Generation. In *Proc. ICCV*, 2021.
- [28] Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely Embedded Convolutional Detection. *Sensors*, 18(10), 2018.

- [29] Bin Yang, Min Bai, Ming Liang, Wenyuan Zeng, and Raquel Urtasun. Auto4D: Learning to Label 4D Objects from Sequential Point Clouds. *arXiv CoRR*, abs/2101.06586, 2021.
- [30] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. ST3D++: Denoised Self-training for Unsupervised Domain Adaptation on 3D Object Detection. *arXiv CoRR*, abs/2108.06682, 2021.
- [31] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. ST3D: Self-training for Unsupervised Domain Adaptation on 3D Object Detection. In *Proc. CVPR*, 2021.
- [32] Zeng Yihan, Chunwei Wang, Yunbo Wang, Hang Xu, Chaoqiang Ye, Zhen Yang, and Chao Ma. Learning Transferable Features for Point Cloud Detection via 3D Contrastive Co-training. In *NeurIPS*, 2021.
- [33] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3D Object Detection and Tracking. In *Proc. CVPR*, 2021.
- [34] Yurong You, Carlos Andres Diaz-Ruiz, Yan Wang, Weilun Chao, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Exploiting Playbacks in Unsupervised Domain Adaptation for 3D Object Detection. *arXiv CoRR*, abs/2103.14198, 2021.
- [35] Weichen Zhang, Wen Li, and Dong Xu. SRDAN: Scale-Aware and Range-Aware Domain Adaptation Network for Cross-Dataset 3D Object Detection. In *Proc. CVPR*, 2021.