

Domain Adaptation on Point Clouds via Geometry-Aware Implicits

Yuefan Shen^{1*} Yanchao Yang^{2†*} Mi Yan³ He Wang³ Youyi Zheng^{1†} Leonidas Guibas²

¹Zhejiang Univerisy[‡] ²Stanford Univerisy ³Peking Univerisy

Abstract

As a popular geometric representation, point clouds have attracted much attention in 3D vision, leading to many applications in autonomous driving and robotics. One important yet unsolved issue for learning on point cloud is that point clouds of the same object can have significant geometric variations if generated using different procedures or captured using different sensors. These inconsistencies induce domain gaps such that neural networks trained on one domain may fail to generalize on others. A typical technique to reduce the domain gap is to perform adversarial training so that point clouds in the feature space can align. However, adversarial training is easy to fall into degenerated local minima, resulting in negative adaptation gains. Here we propose a simple yet effective method for unsupervised domain adaptation on point clouds by employing a self-supervised task of learning geometry-aware implicits, which plays two critical roles in one shot. First, the geometric information in the point clouds is preserved through the implicit representations for downstream tasks. More importantly, the domain-specific variations can be effectively learned away in the implicit space. We also propose an adaptive strategy to compute unsigned distance fields for arbitrary point clouds due to the lack of shape models in practice. When combined with a task loss, the proposed outperforms state-of-the-art unsupervised domain adaptation methods that rely on adversarial domain alignment and more complicated self-supervised tasks. Our method is evaluated on both PointDA-10 and GraspNet datasets. Code and data are available at: <https://github.com/Jhonve/ImplicitPCDA>.

1. Introduction

Point clouds captured under different settings can exhibit prominent variations that cause performance drop when neural networks are tested on a domain that is different from the training ones. This can be troublesome if the net-

[‡]The authors from Zhejiang University are affiliated with the State Key Lab of CAD&CG. *Equal Contributions, [†]Corresponding Authors.

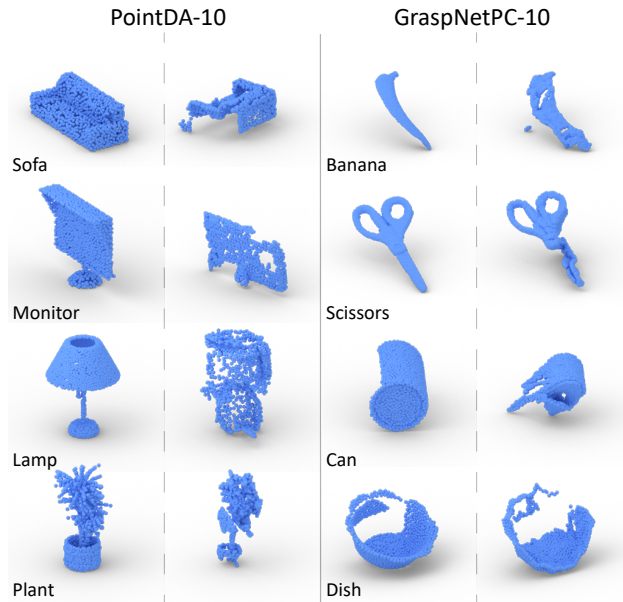


Figure 1. Point clouds in the real world exhibit diverse geometric variations caused by differences in the data capture pipeline. Given these variations, networks trained on one collection of point clouds may incur a performance drop when tested on different ones. Thus adaptation is needed to alleviate generalization issues, especially for domains where the annotation is scarce.

work can *not* be fine-tuned due to time constraints or limited computational budget. More often, labels needed for fine-tuning on the test domain are simply unavailable due to high annotation cost, which is the situation we are interested in and is always formulated as unsupervised domain adaptation (UDA) problems. In UDA, the source domain comes with rich annotations, while the target domain has no annotation at all. The *key* to a successful domain adaptation lies in two folds. First, the two domains have to be (statistically) aligned, either in the point cloud space or in a feature space, so that the shared mapping to the output space can now operate on the same ground across domains. Moreover, the alignment between domains has to be semantically meaningful, e.g., chairs in the source should be aligned with chairs in the target. Otherwise, the shared mapping can still fail in predicting the labels even if the two domains are aligned.

Existing UDA methods on point clouds mainly rely on two mechanisms to align the domains. One is to perform domain adversarial training and enforce the features of point clouds from both domains to be indistinguishable by domain discriminators. Since adversarial training is unstable and easy to get stuck at degenerated local minimas, there is little guarantee that the alignment would be semantically meaningful. For example, adversarial training could distort the geometric information in the point clouds by eliminating too many variations while aligning the domains. In this case, the alignment can result in negative adaptation gains. An extra layer of difficulty is that the alignment process could be highly sensitive to the architecture of the discriminators for point clouds as shown in [32], thus making the alignment more uncontrollable.

The other mechanism is to perform domain alignment through learning self-supervised tasks. The underlying motivation is that a well-designed self-supervised task can facilitate learning domain invariant features since the task itself is shared across domains. A diverse set of carefully designed self-supervised tasks are proposed, which focus on predictive tasks where the self-supervised labels are generated by augmenting or modifying the original point clouds. For instance, rotation angle classification [40] and deformation regression [1]. Compared to domain adversarial training, self-supervised learning enables explicit control over the invariants been learned by adjusting the self-supervised tasks. Consequently, one can also regularize the alignment process through this knob.

We take the latter approach, but we resort to a self-supervised task where the supervision comes from the point clouds themselves, instead of manually designed classification labels. Specifically, we ask for a latent space that encodes the underlying geometry of the point clouds through implicit functions. As the geometry is explicitly modeled and preserved, these latents or implicits should maintain sufficient information for the main task and help prevent mismatch in semantics caused by distortions during the alignment. Due to the lack of shape models, we propose an adaptive unsigned distance field that enables training the implicits for arbitrary point clouds, especially for the ones that are sparse and irregularly sampled. After the initial round of adaptation, we follow the literature and apply self-training with pseudo labels in the target domain to further close the gap. We experiment on two major point cloud datasets, PointDA-10 [22] and GraspNet [9], to report the performance of the proposed method and evaluate the effectiveness of each component. Our contributions are:

- The first method leverages implicit function learning as a self-supervised task for unsupervised domain adaptation on point clouds.
- Effective training strategies to make our method robust to diverse artifacts exhibited in the point clouds.

- State-of-the-art performance on two major datasets, PointDA-10 [22] and GraspNet [9]. Moreover, we are the first to report results on GraspNet.

2. Related Work

2.1. Deep Learning on Point Clouds

To handle the irregularity and permutation-invariance of point clouds, various methods have been proposed. PointNet [20] and PointNet++ [21] use max-pooling as a permutation-invariant local feature extractor and the latter gathers local features in a hierarchical way. DGCNN [33] considers a point cloud as a graph and dynamically updates the graph to aggregate features. Recently, Point Transformer [37] adopts transformer for point cloud processing which achieves state-of-the-art performance in several benchmarks.

2.2. Unsupervised Domain Adaptation

Extensive works have been proposed to perform UDA on 2D images, which can be classified into two categories, i.e., the methods based on domain-invariant feature learning and methods for learning domain mapping. The former ones [10, 12, 14, 23, 25, 30] minimize the discrepancy between two distributions in the feature space, while the latter ones [3, 11, 29] directly learn the translation from the source domain to the target domain using neural networks, e.g., CycleGAN [39]. [28] expand 2D translation to depth images by proposing a differential contrastive learning strategy for preserving underlying geometries. Despite their differences, domain adversarial training is widely exploited in these methods. Several useful techniques are also proposed, for example, pseudo-labeling [24], and batch normalization tailored for domain adaptation [16].

Though lots of efforts have been made on 2D images or depth, UDA on 3D point cloud is still in its early stage. As discussed in Sec. 1, UDA on point clouds can be roughly divided into two categories. The first category [22] directly extends domain adversarial training used in 2D images to 3D point clouds to align features on both local and global levels. However, unlike previous works on the 2D domain, adversarial methods on 3D point clouds can not balance well between local geometry alignment and global semantic alignment. Most recent works in UDA on point clouds fall in the second category, i.e., focusing on designing suitable self-supervised tasks on point clouds to facilitate learning domain invariant features, which we discuss in detail in the following subsection.

Apart from UDA on object point clouds, several methods are proposed to address specific domain gaps on LiDAR point clouds, where the common factors are depth missing and sampling difference between sensors. Both [38]

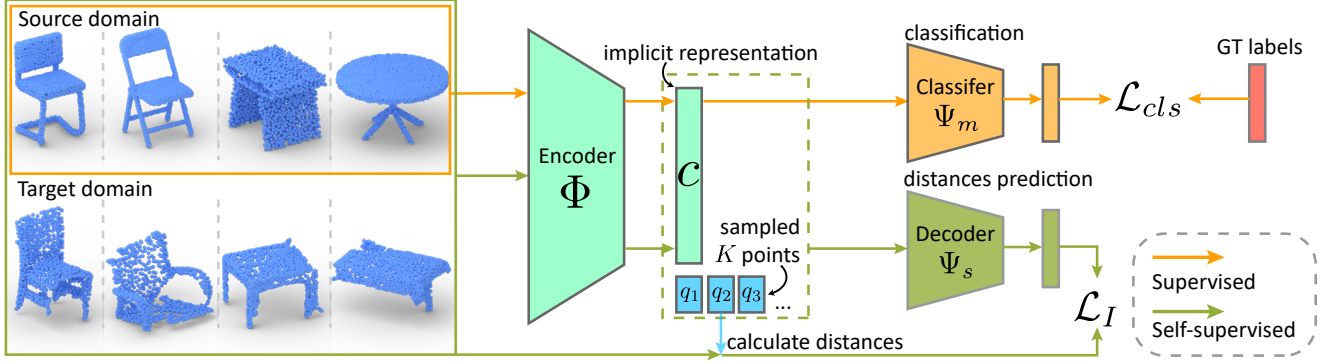


Figure 2. Overview of the proposed framework for unsupervised domain adaptation on point clouds. The two pathways (supervised and self-supervised) in our framework are marked with different colors. The supervised pathway takes as input the point clouds from the source domain and calculates the cross-entropy loss with ground-truth labels. The self-supervised pathway takes point clouds from the source and target domains and calculates the self-supervised loss with the proposed adaptive unsigned distances between sampled points and the input point clouds. Note, in the self-paced self-training stage, the classifier is also trained with pseudo labels.

and [26] use CycleGAN [39] to generate more realistic LiDAR point clouds from synthetic data, i.e., Sim2Real for minimizing feature distances between the source and target domains. Complete & Label [36] leverages segmentation on completed surface reconstructed from sparse point cloud for better adaptation. ST3D [35] presents a task-specific self-training pipeline with curriculum data augmentation.

2.3. Self-Supervised Learning on Point Clouds

Previous works design various kinds of self-supervised tasks to align the two domains. DefRec [1] proposes deformation-reconstruction and [15] extends it into a learnable deformation task to further improve the performance. [2, 27] shuffle and restore the input point cloud to improve discrimination. [4, 8, 40] further combine self-learning strategies and their proposed self-supervised tasks. Besides, [8, 40] present self-supervised tasks to align features at both local and global levels.

However, there are two main issues with these methods. Some of them can not be applied to more challenging datasets where object point clouds are not aligned and are heavily occluded, resulting in ambiguity in the rotation prediction [8, 19, 40] and restoring [2, 27] tasks. Besides, by aligning high-level features [1, 15, 27, 40], i.e., in semantic space, they could lose valuable information of the underlying geometry, which limits their applicability to more general geometric processing tasks. Motivated by these two observations, we design a task where the point cloud itself generates the self-supervision on the two domains and features are aligned to preserve geometric primitives. The aligned features can further be used for high-level semantic extraction, making our method more general for various main tasks.

3. Method

We tackle unsupervised domain adaptation (UDA) on point clouds for classification. Let $P \in \mathbb{R}^{N \times 3}$ be a point cloud consisting of the spatial coordinates of N points in the 3D space. Accordingly, let $D^s = \{P_i^s; Y_i^s\}$ be the point clouds and their ground-truth labels from the source domain. Similarly, $D^t = \{P_i^t\}$ is the collection of target domain point clouds whose labels are missing. Our goal is to train a network Ψ , i.e., $Y = \Psi(P)$ using the labeled point clouds from the source domain so that it can work well on the target point clouds without further labeling.

The *key* is to *align* the point clouds from both domains, and at the same time, ensure that the correspondence is *semantically meaningful*, i.e., the point clouds of the same category are expected to be aligned after the adaptation. One can apply domain adversaries for aligning domains, however, the alignment is hard to control and may result in negative adaptation gains due to difficulties in adversarial training. We resort to the strategy of utilizing self-supervised tasks that are shared across domains for alignment in a multi-task fashion. This enables an explicit control of the meaningfulness of the alignment by selecting an appropriate self-supervised task. There are two pathways in our framework, as shown in Fig. 2. The *main task* is performed by Ψ_m , i.e., $Y = \Psi_m(P)$, with an encoder that extracts features from the point clouds and Ψ_m the main task head (classifier). Likewise, the *self-supervised task* is performed by Ψ_s (shared with the main task pathway) and \mathcal{L}_I , which can be trained on both domains. Next, we detail each of the proposed components and their training.

3.1. Self-Supervised Geometry-Aware Implicit

Implicit representations are capable of preserving complex details for given shapes [6, 18]. Instead of high-quality shape reconstruction, we leverage the implicit representation space for aligning point clouds from different domains

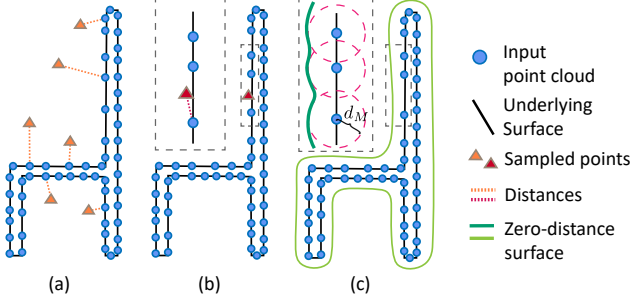


Figure 3. Adaptive unsigned distance field. (a): examples of calculating distances from sampled points (triangle) to their nearest points in the input point cloud. (b): when a sampled point is close to the surface, its nearest neighbor distance is still large due to sparsity. (c): the adaptive unsigned distance field and the zero-surface, with d_M the adaptive clamping value.

by performing the following self-supervised task.

Given a point cloud P , either complete or partial, the shared encoder first maps it to a feature vector $c = \mathcal{E}(P)$ as the implicit representation of the unknown underlying shape from where P is observed. Suppose $Q \subset \mathbb{R}^K$ are K randomly sampled points in the unit cube. By definition, the implicit value (e.g., distance to the surface) for each point $q \in Q$ can be decoded as:

$$f_P(q) = \mathcal{D}_s(q; c) \quad (1)$$

where f_P is the implicit function of the underlying geometry conditioned on the input point cloud P . Following the literature [6, 17], the decoder \mathcal{D}_s takes as input the concatenation of the query point and the encoded implicit representation. Since the point clouds can be partial, we set the implicit values as unsigned distances to the underlying surface. The computation of these values is described in the following.

3.1.1 Adaptive Unsigned Distance of Point Cloud

Different from reconstruction where the known meshes can be used to compute the ground-truth for the distance values, we only have access to the point clouds. However, as our goal is to leverage the implicit representation to align domains and reduce performance drop, we do not need the implicits to perfectly represent the underlying geometry and reconstruct the point cloud. To this end, we can compute approximates of the unsigned distance fields to supervise the training of the implicit space.

An intuitive method is to approximate the unsigned distance from a query point to the underlying surface by the distance between the same query point and its nearest neighbor from the point cloud (Fig. 3 (a)). This could work if the point clouds are densely and uniformly sampled. Nevertheless, in practice, point clouds are usually sparse and irregularly sampled due to sensor noise and complex geometry in the scene. These peculiarities can cause problems for

the nearest neighbor approximations. For example, when the query point is very close to the underlying surface, the distance could still be large, as shown in Fig. 3 (b). Thus the learned implicit space may not faithfully represent the geometry of the point clouds and can induce performance drop across domains.

To prevent unexpected distortions of the geometry in the approximation, we propose an adaptive clamping technique based on a global average over statistics of the local geometry. For a point p_j in the input point cloud P , we first calculate the mean of the distances between p_j and its M nearest neighbors within P :

$$d_j = \frac{1}{M} \sum_m k p_m - p_j k \quad (2)$$

where p_m is from the M nearest neighbours and we name d_j the *local affinity* of point p_j . Then, we compute the average of the local affinity of all the points in the point cloud, i.e., $d_M = \frac{1}{N} \sum_{j=0}^{N-1} d_j$, which is the *adaptive clamping threshold* and is used in the following to compute the adaptive approximate of the unsigned distance field from the point cloud:

$$d_P(q) = \begin{cases} kq - p(q)k & \text{if } kq - p(q)k > d_M \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where $p(q)$ is the nearest neighbor of the query q in the point cloud P . Also, note that d_M depends on P and can vary between point clouds to accommodate different sparsity levels. An example of the adaptive unsigned distance field can be found in Fig. 3 (c). As observed, the unsigned distance field approximated via Eq. (3) captures the underlying geometry of the point cloud and is more robust to sampling issues. With the adaptive unsigned distance field d_P , the self-supervised loss for learning the implicit space is:

$$L_I = \frac{1}{|Q|} \sum_{q \in Q} |f_P(q) - d_P(q)| \quad (4)$$

here, $|Q|$ is the cardinality of the sampled query points. Next, we discuss a few issues encountered during the articulation of the whole pipeline and our solutions.

3.1.2 Point Cloud Augmentation

Jittering. The point cloud backbone usually assumes a fixed number of points during training, for example, 1024 points for a single point cloud. However, in practice, the number of points in a single point cloud may not be the same due to irregular sampling or different shape sizes. For example, in the unsupervised domain adaptation benchmark PointDA-10 [22], point clouds from ModelNet and ScanNet can have very different numbers of points. A commonly

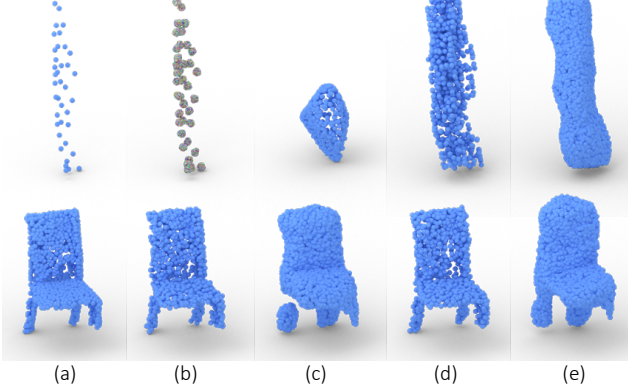


Figure 4. (a): Input point clouds with duplication. The point cloud in the top row has only 38 unique points, but it contains 1024 points in total with duplication to account for the fixed input size of the backbone. (b): Point clouds after random jittering in the range $[0.03; 0.06]$ (instead of duplicating points), which simply dilates each point. The points in the top row are colored randomly for better discrimination. (c): Point clouds sampled from the implicits learned with the random jittering scheme in (b). We can observe much heavier geometric distortion for sparser point clouds (top vs. bottom). (d): Point clouds jittered using the proposed scheme based on the spatially varying local affinity measure d_j in Eq. (2). (e) Sampled point clouds from implicits learned with the jittering scheme in (d), which preserves the geometry for both sparse and dense point clouds.

used technique is to pad the point clouds to the same number of points through jittering, which may also improve the training if the jittering is properly designed. The simplest jittering method is to add duplicate points to the original point cloud (Fig. 4 (a)). Another method is to add uniform random perturbations (Fig. 4 (b)). However, both methods will generate points that make the local affinity measure uninformative, so that the proposed adaptive unsigned distance field may not be effective in preventing geometric distortions for sparse and irregular point clouds. As shown in Fig. 4 (c), the resampled point clouds from the learned implicits with random perturbations exhibit significant geometric distortions for sparse point clouds.

In order to avoid such degenerated cases in the padding procedure, we propose to perform the point jittering in an affinity-aware manner. Similar to calculating the adaptive unsigned distance field, for each point p_j in the raw point cloud, we first obtain its local affinity d_j using Eq. (2). A random offset in the range of $[\frac{d_j}{2}, \frac{d_j}{2}]$ is then added to p_j to generate jittered points. The point clouds generated with this jittering scheme have little deviation from the raw point clouds, as shown in Fig. 4 (d). Moreover, the resampled point clouds from the implicits learned with the affinity-aware jittering maintain the underlying geometry for both sparse and dense point clouds as observed in Fig. 4 (e).

Random masking. Point clouds may come in a partial form due to self-occlusions. To improve the robustness of the implicits concerning the partiality and further reduce the

domain variations, we choose to mask out a local neighborhood of a randomly selected point as an additional data augmentation.

Let P be a point cloud before random masking, and \hat{P} be the point cloud obtained by dropping out a neighborhood of radius of r_m , i.e., the masked point cloud. We ask the implicits of both point clouds to be similar as they are sampled from the same geometry. During training, we add a loss term between the implicit representations of the input point cloud and its masked version (k is the L-2 distance):

$$L_M = k(P) - k(\hat{P}) \quad (5)$$

3.2. Self-Paced Self-Training

The main task we tackle here is point cloud classification. Before adaptation, we only have labeled data in the source domain, i.e., $fP_i^s; Y_i^s g$, which allows us to train the main task branch with a cross-entropy loss:

$$L_{cls}^s = \frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{j=1}^J Y_{i,j}^s \log(m(P_i^s)_j) \quad (6)$$

where $Y_{i,j}^s$ represents the ground-truth one-hot labels and $m(P_i^s)_j$ is the predicted probability for the j^{th} class.

When the initial adaptation is made, point clouds from the source and target domains should be aligned to some extent. In this case, techniques used in semi-supervised learning are now in their functioning state. For example, GAST [40] employs the strategy of self-paced self-training (SPST) [13, 41] to further align the two domains by generating pseudo labels in the target domain using highly confident predictions. We follow this strategy to squeeze more juice out of the source labels. Let \hat{Y}_i^t be the predicted pseudo labels, the loss function to perform the self-training is:

$$L_{cls}^t = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(\sum_{j=1}^J \hat{Y}_{i,j}^t \log(m(P_i^t)_j) + j \hat{Y}_{i,j}^t \right) \quad (7)$$

Similarly, the first term in Eq. (7) is a cross-entropy loss between the target pseudo labels and the predictions, and the second term is used to avoid degenerate solutions that assign all \hat{Y}^t as 0. We follow [40, 41] to apply a two-stage optimization using Eq. (7), where the pseudo labels are first computed using nonlinear integer programming. Then the branch m is updated using the pseudo labels. These two steps are performed iteratively to adapt between the source and target domains progressively. The hyperparameter α controls the number of selected target samples.

3.3. Overall Loss

The overall training loss of our method is:

$$L = L_I + L_M + L_{cls}^s + L_{cls}^t \quad (8)$$

Note, the self-supervised implicit representation learning can be pre-trained on point clouds to encourage faster convergence during adaptation, i.e., set $\lambda = 0$. After pre-training the networks \mathcal{E}_s for learning geometry-aware implicits, together with loss terms of the classification task L_{cls}^s and L_{cls}^t can be added back to perform the joint domain adaptation.

4. Experiments

To show that the implicits effectively encode geometries of point clouds and verify the importance of the proposed adaptive unsigned distance field, we examine the implicit reconstructions in Sec. 4.3. To have a comprehensive understanding of both the effectiveness and limitations of the implicits learned from unconstrained point clouds for aligning the domains, we evaluate the whole pipeline for point cloud UDA on the classification task with two major datasets. We report our results with and without self-paced self-training.

We compare to a list of recent state-of-the-art methods on unsupervised point cloud domain adaptation: DANN [10], PointDAN [22], RS [27], DefRec+PCM [1] and GAST [40]. In addition, we report the results obtained from the same network trained in a supervised manner on the target domain (“Supervised”, upper-bound). For reference, the network trained in the source domain but tested on the target domain without any adaptation is also included (“Baseline”, lower-bound).

4.1. Datasets

PointDA-10 [22] consists of three widely-used datasets: ModelNet [34], ShapeNet [5] and ScanNet [7]. All three datasets share the same ten categories (bed, table, sofa, chair, etc.). ModelNet contains 4183 training and 856 test samples, while ShapeNet contains 17378 training and 2492 test samples. ModelNet and ShapeNet are both sampled from 3D CAD models. Unlike these synthetic point cloud datasets, ScanNet consists of point clouds from scanned and reconstructed real-world scenes. There are 6110 training samples and 2048 test samples in ScanNet, and point clouds therein are usually incomplete because of occlusion by surrounding objects in the scene or self-occlusion in addition to realistic sensor noises. We follow the data preparation procedure used in [1, 22, 40]. Specifically, all object point clouds in all datasets are aligned along the direction of gravity, while arbitrary rotations along the z axis are allowed. Moreover, the input point cloud with batching is a list of 1024 points, which are sampled with duplicative padding from the original point clouds and are normalized to a unit scale.

GraspNetPC-10 In order to test the domain adaptation on sim-to-real and real-to-real and check how the adaptation copes with different types of sensor noise, we introduce GraspNetPC-10. It is created from GraspNet [9]

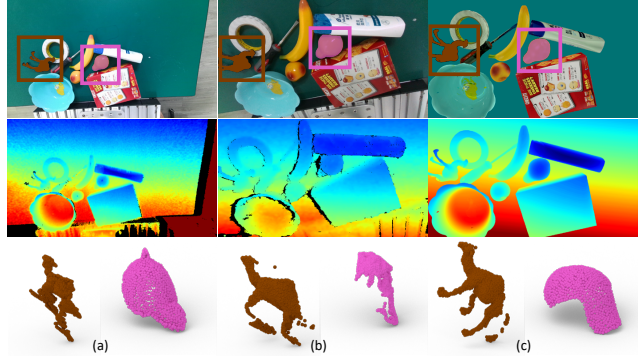


Figure 5. Point clouds from GraspNetPC-10 created with GraspNet [9]. (a-b): RGBD and raw point cloud captured by Kinect and Realsense devices, respectively, and (c): Synthetic RGBD and point cloud. Segmentation masks are provided, as shown in the first row. The corresponding re-projected and cropped point clouds are visualized in the same color at the bottom.

proposed for training robotic grasping on raw depth scans and corresponding reconstructed 3D CAD models of various objects. As shown in Fig. 5, we create GraspNetPC-10 by re-projecting raw depth scans to 3D space and applying object segmentation masks to crop out the corresponding point clouds. Meanwhile, we synthesize similar senses with the same objects and render the synthetic depth scans to re-project synthetic point clouds. Different from point clouds in PointDA-10, point clouds in GraspNetPC-10 are *not aligned*.

Raw depth scans in GraspNet [9] are captured by two different depth cameras, Kinect2 and Intel Realsense, so we have two domains of real-world point clouds. Following PointDA-10, we collect synthetic and real-world point clouds for ten object classes. In the synthetic domain, there are 12,000 training point clouds. In the Kinect domain, there are 10,973 training and 2,560 testing point clouds. Similarly, in the Realsense domain, there are 10,698 training and 2,560 testing point clouds. The real-world point clouds from the two devices are always corrupted by different noises, and there exist different levels of geometric distortions and missing parts, as observed in Fig. 1 and Fig. 5.

4.2. Implementation Details

Our experiments are conducted on servers with four GeForce RTX 3090 GPUs, and the networks are implemented within the PyTorch framework. For training, we use an Adam optimizer, with an initial learning rate 0.001, weight decay 0.00005, and an epoch-wise cosine annealing learning rate scheduler. We train all models for 200 epochs on PointDA-10 and 120 epochs on GraspNetPC-10 with a batch size of 32.

Network architecture Following [41], we choose the commonly used point cloud processing network DGCNN [33] as the backbone for the encoder \mathcal{E} . The implicit decoder \mathcal{D}_s and the category classifier \mathcal{C}_m are multi-

Methods	Adv.	SSL	SPST	M/ S	M/ S*	S/ M	S/ S*	S*/ M	S*/ S	Avg.						
Supervised				93.9	0.2	78.4	0.6	96.2	0.1	78.4	0.6	96.2	0.1	93.9	0.2	89.5
Baseline (w/o adap.)				83.3	0.7	43.8	2.3	75.5	1.8	42.5	1.4	63.8	3.9	64.2	0.8	62.2
DANN [10]	×			74.8	2.8	42.1	0.6	57.5	0.4	50.9	1.0	43.7	2.9	71.6	1.0	56.8
PointDAN [22]	×			83.9	0.3	44.8	1.4	63.3	1.1	45.7	0.7	43.6	2.0	56.4	1.5	56.3
RS [27]		×		79.9	0.8	46.7	4.8	75.2	2.0	51.4	3.9	71.8	2.3	71.2	2.8	66.0
DefRec+PCM [1]		×		81.7	0.6	51.8	0.3	78.6	0.7	54.5	0.3	73.7	1.6	71.1	1.4	68.6
GAST [40]		×		83.9	0.2	56.7	0.3	76.4	0.2	55.0	0.2	73.4	0.3	72.2	0.2	69.5
		×	×	84.8	0.1	59.8	0.2	80.8	0.6	56.7	0.2	81.1	0.8	74.9	0.5	73.0
Ours		×		85.8	0.3	55.3	0.3	77.2	0.4	55.4	0.5	73.8	0.6	72.4	1.0	70.0
		×	×	86.2	0.2	58.6	0.1	81.4	0.4	56.9	0.2	81.5	0.5	74.4	0.6	73.2

Table 1. Classification accuracy (%) averaged over 3 seeds (SEM) on the PointDA-10. M: ModelNet, S: ShapNet, S*: ScanNet; ! indicates the adaptation direction. Adv.: adversarial domain alignment, SSL: self-supervised learning, and SPST: self-paced self-training.

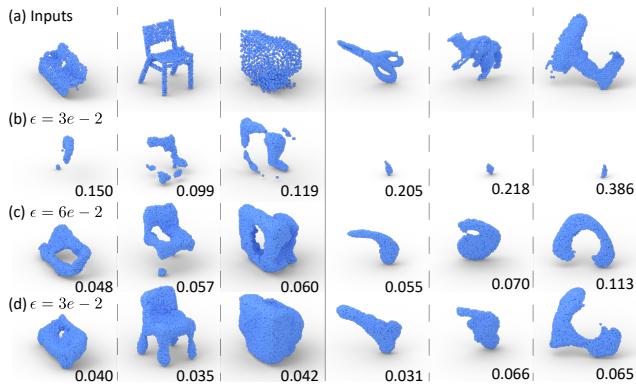


Figure 6. Visualization of resampled point clouds from the learned implicits. The left half shows test samples from the PointDA-10 dataset, and the right half shows test samples from GraspNetPC-10. (a): input point clouds, (b-c): resampled point clouds without adaptive unsigned distance field (AUD) at $\epsilon = 3e-2$ and $\epsilon = 6e-2$, (d): resampled point clouds with AUD at $\epsilon = 3e-2$. The inserted numbers are the Chamfer distances between the resampled and the input point clouds (a).

layer perceptrons (MLP) with fully connected layers. Decoder \mathcal{D}_s is a four-layers MLP $f_{512;256;128;1g}$ followed by ReLU activation function (to make the output distance always positive) and classifier \mathcal{M} is a three-layers MLP $f_{512;256;10g}$ in view of 10 semantic classes.

Hyper parameters We set $M = 3$ for searching nearest neighbor points when calculating our adaptive unsigned distance field (AUD). The radius for the random masking r_m is sampled from a uniform distribution in the range of $[0.1, 0.3]$. The weights of loss terms are set to $w_1 = 100$; $w_2 = 1.0$; $w_3 = 1.0$ and we adjust them slightly for better convergence on different datasets.

4.3. Implicit Reconstruction

We show the resampled point clouds from learned implicit representations for analyzing the quality of the self-supervised geometry-aware implicits. Once the implicit encoder-decoder (\mathcal{D}_s) is trained, given an input point cloud P , we randomly sample $200,000$ points $q_s \in \mathbb{R}^{200,000 \times 3}$ in the unit cube and calculate their unsigned

distances conditioned on the original point cloud with our trained networks $d^{q_s} = \mathcal{D}_s(q_s; (P))$. By setting a distance threshold ϵ , we can choose the subset $q_s \in \mathcal{Q}_s$ s.t. $d^{q_s} < \epsilon$ for visualization. If the distance field f_P is a good approximation of the underlying geometry, then \mathcal{Q}_s will be similar to the input point clouds when ϵ varies.

In Fig. 6, we compare our resampled point clouds with the input point clouds and the resampled point clouds from implicits learned without using our adaptive unsigned distance field, i.e., directly using the distances to the nearest neighbor but with a fixed clamping value. As observed, the resampled point clouds with AUD can preserve the underlying geometry well. However, using the same ϵ , the resampled point clouds without using AUD (“w/o AUD”) are much inferior, meaning the learned implicits distort the geometry information. Moreover, we report the Chamfer distance between the resampled point clouds and the input. Fig. 6 (c) shows the best resampled one for implicits learned without AUD (“w/o AUD”). One can see that “w/o AUD” needs to apply a much larger ϵ , i.e., two times larger than what is needed by “AUD”, but the resampled point clouds are still severely deformed and exhibit lots of missing. These results demonstrate that our adaptive unsigned distance field is critical and effective in the proposed implicit representation alignment module.

4.4. Unsupervised Domain Adaptation

Table 1 and Table 2 show the comparisons between our method and other state-of-the-art on PointDA-10 and GraspNetPC-10, respectively. We perform hyper-parameter search for all the methods. For PointDA-10, we follow [40] and report performances on six different settings including ModelNet (M) $\&$ ShapeNet (S), M $\&$ ScanNet (S*) and S $\&$ S*. We find that methods utilizing self-supervised tasks generally perform better than methods based on adversarial training, especially on “synthetic to real” settings. Compared to other self-learning-based methods, our method (w/o SPST) excels on four settings and the average performance. After adding self-paced learning, our method competes

Methods	Adv.	SSL	SPST	Syn./ Kin.	Syn./ RS.	Kin./ RS.	RS./ Kin.	Avg.
Supervised				97.2	0.8	95.6	0.4	96.4
Baseline (w/o adap.)				61.3	1.0	54.4	0.9	59.4
DANN [10]	×			78.6	0.3	70.3	0.5	65.7
PointDAN [22]	×			77.0	0.2	72.5	0.3	74.4
RS [27]		×		67.3	0.4	58.6	0.8	62.8
DefRec+PCM [1]		×		80.7	0.1	70.5	0.4	73.5
GAST [40]		×		69.8	0.4	61.3	0.3	65.1
		×	×	81.3	1.8	72.3	0.8	73.8
Ours		×		81.2	0.3	73.1	0.2	75.8
		×	×	94.6	0.4	80.5	0.2	84.4

Table 2. Classification accuracy (%) averaged over 3 seeds (SEM) on GraspNetPC-10. Syn.: Synthetic domain, Kin.: Kinect domain, RS.: Realsense domain. Our models achieve the best performance over all settings.

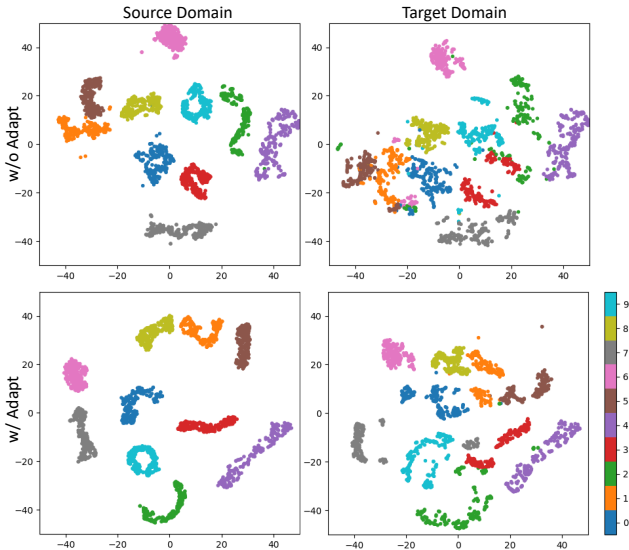


Figure 7. t-SNE [31] visualization of the output from our point cloud backbone on the Kinect domain (source) and the Realsense domain (target), which shows that the alignment through the implicits is effective, i.e., point cloud implicits from the target align better with the source ones after the adaptation (top vs. bottom). Different classes are displayed with different colors.

with the most recent state-of-the-art method GAST [40] on PointDA-10. Compared to RS [27] and DefRec+PCM [1] that both use reconstruction-based self-supervised tasks, our method again achieves better performance.

For GraspNetPC-10, our method outperforms the others with a significant margin before and after adding self-paced learning. One can observe a substantial decline of GAST [40] on GraspNetPC-10. The reasons are that point clouds are not canonicalized in GraspNetPC-10 which will involve ambiguities when using rotation classification and switching off rotation term will lose global alignment for adaptation. The local alignment method proposed in PointDAN [22] now performs better than on the PointDA-10 dataset. DefRec+PCM [1] ranks similarly. Our method achieves the highest score across all settings, whether with or without SPST. It is also worth noting that SPST is effective

on all datasets, both GAST [40], and our method improves with SPST. However, “GAST+SPST” is still worse than ours without SPST, which again shows the effectiveness of the proposed geometry-aware implicits for aligning domains with realistic sensor noise.

We also visualize the 1024-dimension latent codes in the implicit space using t-SNE [31]. As seen in Fig. 7, without domain adaptation, features of different classes in the target domain are mixed-up (e.g., class 1 and 5, class 2 and 3), and the overall distribution is different from that in the source domain. After adaptation, the distribution of the features in the target domain becomes similar to the source one and shows clear clusters. More analyses on domain distances can be found in our supplemental material.

5. Discussion

It is challenging to align point clouds while maintaining a correct correspondence in terms of semantics without the target labels. However, we show that a simple alignment via the proposed implicit space training can be quite effective for the current unsupervised domain adaptation benchmarks on point clouds. Our method achieves state-of-the-art performance on two benchmarks covering varying factors affecting the point cloud geometry within the data collection pipeline. We hope our method can serve as a ground where low-level geometric distortions or variations are learned away so one can focus on high-level shape variations that are also generative factors for domain gaps. This would require a carefully designed dataset with controllable disentangled elements of geometric variations and is out of the scope of our current work.

Acknowledgment. This work was supported by National Key Research & Development Program of China (Grant No. 2018YFE0100900) in part, a Vannevar Bush faculty fellowship, ARL grant W911NF2120104, NSF grant IIS-1763268, and a gift from the Autodesk Corporation.

References

- [1] Idan Achituve, Haggai Maron, and Gal Chechik. Self-supervised learning for domain adaptation on point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 123–133, 2021. 2, 3, 6, 7, 8
- [2] Antonio Alliegro, Davide Boscaini, and Tatiana Tommasi. Joint supervised and self-supervised learning for 3d real world challenges. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6718–6725. IEEE, 2021. 3
- [3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3722–3731, 2017. 2
- [4] Adriano Cardace, Riccardo Spezialetti, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Refrec: Pseudo-labels refinement via shape reconstruction for unsupervised 3d domain adaptation. In *2021 International Conference on 3D Vision (3DV)*, pages 331–341. IEEE, 2021. 3
- [5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 6
- [6] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 3, 4
- [7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 6
- [8] Hehe Fan, Xiaojun Chang, Wanyue Zhang, Yi Cheng, and Ying Sun. Self-supervised global-local structure modeling for point cloud domain adaptation with reliable voted pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3
- [9] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11444–11453, 2020. 2, 6
- [10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2, 6, 7, 8
- [11] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018. 2
- [12] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019. 2
- [13] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 5
- [14] Mingsheng Long, Yue Cao, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Transferable representation learning with deep adaptation networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):3071–3085, 2018. 2
- [15] Xiaoyuan Luo, Shaolei Liu, Kexue Fu, Manning Wang, and Zhijian Song. A learnable self-supervised task for unsupervised domain adaptation on point clouds. *arXiv preprint arXiv:2104.05164*, 2021. 3
- [16] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Buló. Autodial: Automatic domain alignment layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5067–5075, 2017. 2
- [17] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 4
- [18] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 3
- [19] Omid Poursaeed, Tianxing Jiang, Han Qiao, Nayun Xu, and Vladimir G Kim. Self-supervised learning of point clouds via orientation estimation. In *2020 International Conference on 3D Vision (3DV)*, pages 1018–1028. IEEE, 2020. 3
- [20] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 2
- [21] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017. 2
- [22] Can Qin, Haoxuan You, Lichen Wang, C-C Jay Kuo, and Yun Fu. Pointdan: A multi-scale 3d domain adaption network for point cloud representation. *Advances in Neural Information Processing Systems*, 32:7192–7203, 2019. 2, 4, 6, 7, 8
- [23] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):801–814, 2018. 2
- [24] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 2988–2997. PMLR, 2017. 2

- [25] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. [2](#)
- [26] Khaled Saleh, Ahmed Abobakr, Mohammed Attia, Julie Iskander, Darius Nahavandi, Mohammed Hossny, and Saeid Nahvandi. Domain adaptation for vehicle detection from bird’s eye view lidar point cloud data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. [3](#)
- [27] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. *arXiv preprint arXiv:1901.08396*, 2019. [3](#), [6](#), [7](#), [8](#)
- [28] Yuefan Shen, Yanchao Yang, Youyi Zheng, C. Karen Liu, and Leonidas J. Guibas. Dcl: Differential contrastive learning for geometry-aware depth synthesis. *IEEE Robotics and Automation Letters*, 7(2):4845–4852, 2022. [2](#)
- [29] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2107–2116, 2017. [2](#)
- [30] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017. [2](#)
- [31] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. [8](#)
- [32] He Wang, Zetian Jiang, Li Yi, Kaichun Mo, Hao Su, and Leonidas J Guibas. Rethinking sampling in 3d point cloud generative adversarial networks. *arXiv preprint arXiv:2006.07029*, 2020. [2](#)
- [33] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. [2](#), [6](#)
- [34] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1912–1920, 2015. [6](#)
- [35] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10368–10378, 2021. [3](#)
- [36] Li Yi, Boqing Gong, and Thomas Funkhouser. Complete & label: A domain adaptation approach to semantic segmentation of lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15363–15373, 2021. [3](#)
- [37] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021. [2](#)
- [38] Sicheng Zhao, Yezhen Wang, Bo Li, Bichen Wu, Yang Gao, Pengfei Xu, Trevor Darrell, and Kurt Keutzer. epointda: An end-to-end simulation-to-real domain adaptation framework for lidar point cloud segmentation. *arXiv preprint arXiv:2009.03456*, 2, 2020. [2](#)
- [39] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2223–2232, 2017. [2](#), [3](#)
- [40] Longkun Zou, Hui Tang, Ke Chen, and Kui Jia. Geometry-aware self-training for unsupervised domain adaptation on object point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6403–6412, 2021. [2](#), [3](#), [5](#), [6](#), [7](#), [8](#)
- [41] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. [5](#), [6](#)