Synergizing Contrastive Learning and Optimal Transport for 3D Point Cloud Domain Adaptation

Siddharth Katageri[1*] Arkadipta De[2*] Chaitanya Devaguptapu[2*] VSSV Prasad[2]
Charu Sharma[1] Manohar Kaul[2]
[1]IIIT Hyderabad, India [2]Fujitsu Research India
siddharth.katageri@research.iiit.ac.in, charu.sharma@iiit.ac.in, email@chaitanya.one,
{Arkadipta.De, venkatasivasaivaraprasad.kasu, manohar.kaul}@fujitsu.com

Abstract

Recently, the fundamental problem of unsupervised domain adaptation (UDA) on 3D point clouds has been motivated by a wide variety of applications in robotics, virtual reality, and scene understanding, to name a few. The point cloud data acquisition procedures manifest themselves as significant domain discrepancies and geometric variations among both similar and dissimilar classes. The standard domain adaptation methods developed for images do not directly translate to point cloud data because of their complex geometric nature. To address this challenge, we leverage the idea of multimodality and alignment between distributions. We propose a new UDA architecture for point cloud classification that benefits from multimodal contrastive learning to get better class separation in both domains individually. Further, the use of optimal transport (OT) aims at learning source and target data distributions jointly to reduce the cross-domain shift and provide a better alignment. We conduct a comprehensive empirical study on PointDA-10 and GraspNetPC-10 and show that our method achieves state-of-the-art performance on GraspNetPC-10 (with $\approx$ 4-12% margin) and best average performance on PointDA-10. Our ablation studies and decision boundary analysis also validate the significance of our contrastive learning module and OT alignment. https://siddharthkatageri.github.io/COT.

1. Introduction

Representation learning on 3D point clouds is rife with challenges, due to point clouds being irregular, unstructured, and unordered. Despite these hindrances posed by the nature of this complex dataset, learning representations on point clouds have achieved success in a gamut of computer vision areas, such as robotics [19], self-driving vehicles [18], and scene understanding [31], to name a few.

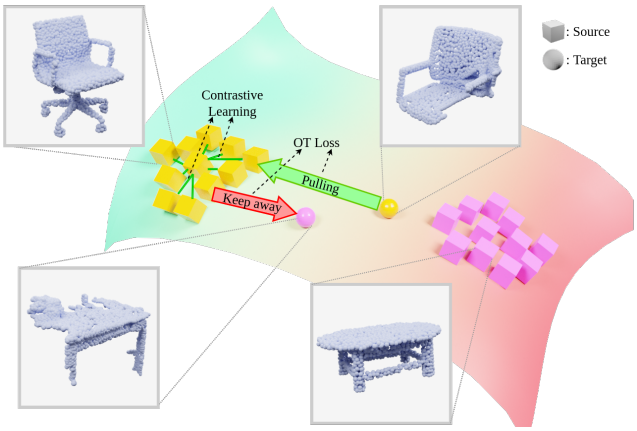

Figure 1. Overview of our method for UDA. Contrastive learning (CL) and optimal transport (OT) are designed to complement each other synergistically. CL establishes class clusters, while OT aligns objects across domains. The colors of data points denote different classes.

While a majority of the point cloud representation learning works have focused on improving performance in supervised and unsupervised tasks [20, 25, 28], very few have focused on the task of *domain adaptation* (DA) between disparate point cloud datasets. This is in part due to the significant differences in underlying structures (i.e., different backgrounds, orientations, illuminations etc. obtained from a variety of data acquisition methods and devices), which in turn manifest themselves as geometric variations and discrepancies between the source and target point cloud domains. An important aspect of achieving cross-domain generalization is to leverage the trained model on simulated data (easy-to-get annotations) and generalize it to real-world data for which obtaining labels is a cumbersome task. The problem persists even in controlled simulated environ-

*These authors contributed equally to this work.

ments. For example, in VR environments, a chair's visual representation can vary significantly between a game and architectural design software. In the more demanding setting of *unsupervised domain adaptation* (UDA) for classification, the source domain consists of labeled point clouds, while the target domain is completely unlabeled.

Recent works focus on incorporating self-supervised learning (SSL) approaches to learn similar features for both domains, along with a regular source domain supervision [1, 24, 32]. The point clouds belonging to the same class must not only be closer in each individual domain, but also achieve cross-domain alignment. However, our analysis reveals that explicit cross-domain alignment is underexplored, given the significant margins between classification accuracies on source and target domains.

Based on our aforementioned observations, we draw inspiration from recent SSL contrastive learning research [2, 5, 17], which has enjoyed major success in other domains such as image and text. We propose a **C**ontrastive SSL method on point clouds to improve class separation individually in both source and target domains that share a common label space. In addition, optimal transport (OT) based methods [9] have also shown promising results as they jointly learn the embeddings between both domains by comparing their underlying probability distributions and exploiting the geometry of the feature space. Thus, we employ **OT** to achieve better cross-domain alignment for domain adaptation. Figure 1 provides a visual overview of our method (COT).

To reduce the domain shift and learn high quality transferable point cloud embeddings, we leverage the idea of multi-modality within the source and target domains and alignment between both their underlying data distributions. We design an end-to-end framework which consists of a multimodal self-supervised contrastive learning setup (shown in Fig. 2) for both source and target domains individually and OT loss for domain alignment. We also incorporate a regular supervised branch that considers labels from the source domain for training. The aim of our setup is to exploit the multimodality of the input data to learn quality embeddings in their respective domains, while reducing the cross-domain shift with the OT alignment.

Main Contributions:

- To the best of our knowledge, we are the first to propose the use of multimodal contrastive learning within individual domains along with OT for domain alignment for 3D point cloud domain adaptation.
- We build an end-to-end framework with two contrastive losses between 3D point cloud augmentations and between a point cloud and its 2D image projections. We also include OT loss for domain alignment.
- We perform an exhaustive empirical study on two popular benchmarks called PointDA-10 and GraspNetPC-10. Our method achieves state-of-the-art performance on GraspNetPC-10 (with $\approx$ 4-12% margin) and the best average performance on PointDA-10. Our method outperforms existing methods in the majority of cases with significant margins on challenging real-world datasets. We also conduct an ablation study and explore decision boundaries for our self-supervised contrastive and OT losses to elucidate the individual contributions of each component in our method.

2. Related Work

Domain Adaptation on Point Clouds Very few works [1, 21, 24, 32] focus on the problem of domain adaptation on point clouds. [21] introduces a benchmark, *PointDA-10* and an approach based on local and global alignment. [1] introduces a self-supervised approach based on deformation reconstruction and leverages *PointMixup* [6]. [32] learns a domain-shared representation of semantic categories by leveraging two self supervised geometric learning tasks as feature regularizers. [24] proposes a self-supervised task of learning geometry-aware implicits for domain-specific variations and additionally propose a new dataset called *GraspNetPC-10* that is developed from *GraspNet* [10]. These works mainly rely on the self-supervision task to improve adaptation, whereas we additionally propose to explicitly align classes across domains.

Optimal Transport for Domain Adaptation Optimal transport based approaches [9, 11, 13, 23, 29] are commonly used in image domain adaptation by aligning the source and target representations. [23] uses Wasserstein distance as a core loss in promoting similarities between embedded representations and proposes *Wasserstein Distance Guided Representation Learning* (WDGRL). [9] proposed *DeepJDOT*, which computes a coupling matrix to transport the source samples to the target domain. [13] presents a new feature selection method that leverages the shift between the domains. [29] proposed *reliable weighted optimal transport* (RWOT) that exploits the spatial prototypical information and the intra-domain structure to dynamically measure the sample-level domain discrepancy across domains to obtain a precise-pair-wise optimal transport plan. [11] proposes an unbalanced optimal transport coupled with a mini-batch strategy to deal with large-scale datasets.

3. Methodology

This section describes our method for UDA of point clouds for classification task. Our method is endowed by *multimodal self-supervised contrastive learning and OT for domain alignment*. The self-supervised multi-modal contrastive learning module leverages both, the 3D information and their corresponding 2D image projections of point clouds. It produces initial class clusters in the source and

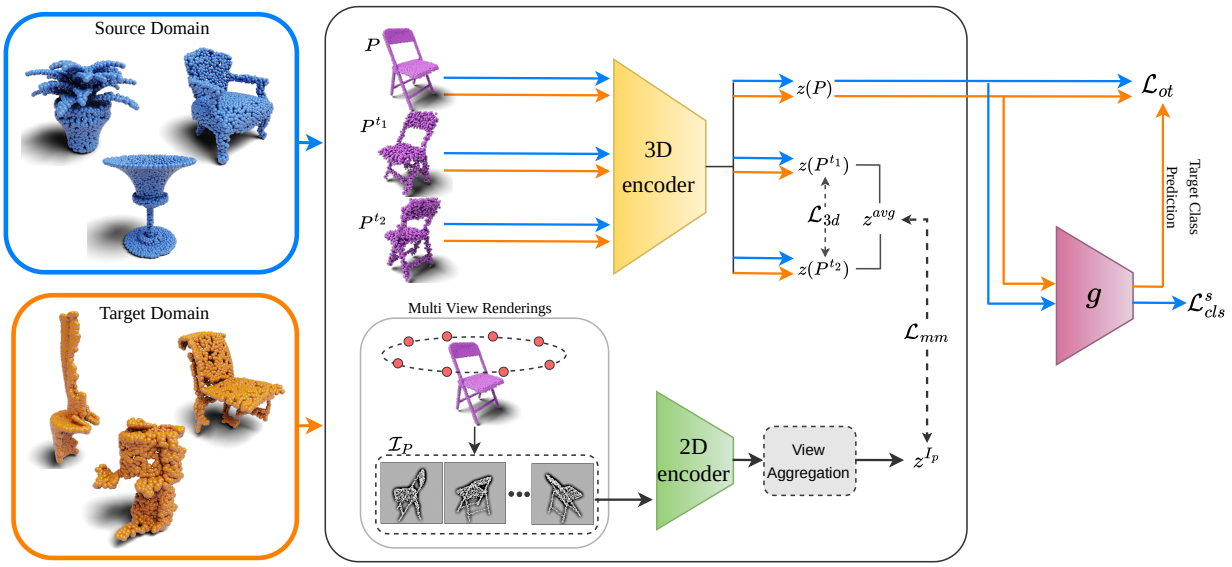

Figure 2. Overview of our framework. Three main components: self-supervised contrastive training ($\mathcal{L}_{3d}$, $\mathcal{L}_{mm}$), self-supervised OT training between both domains ($\mathcal{L}_{ot}$) and a supervised training on source domain ($\mathcal{L}_{cls}$). Contrastive loss uses features from shared Point Cloud and Image encoders with point cloud augmentations and 2D image projections. OT and classifier losses takes features of original point cloud samples from shared Point Cloud encoder.

target domains individually. Subsequently, our OT module better aligns the same class clusters across domains. We additionally also train a classifier on the source domain to improve the class separation, which in turn lessens the burden on our adaptation module.

Our setup aims at learning high quality embeddings, jointly for source and target domains, by exploiting both *contrastive learning with augmentations* and the *multi-modal information of the input point clouds*, while simultaneously *reducing the domain shift* across the domains. Our architecture is illustrated in Figure 2. To this end, we begin by describing self-supervised contrastive learning in Section 3.1. Next, Section 3.2 briefly presents background concepts pertaining to OT and the Wasserstein distance, followed by an explanation of the domain alignment between source and target domains using OT in Section 3.3. Finally, the overall training objective is presented in Section 3.4.

Let a point cloud $P = \{x_1, \ldots, x_n\}$, where $x_i \in \mathbb{R}^3$, be a set of 3D points of cardinality n. Let $\mathcal{D}^s = \{P_i^s, y_i\}_{i=1}^{n_s}$ denote the *labeled source domain dataset*, where P_i^s denotes the i-th source point cloud and y_i its associated class label that takes values in $\mathcal{Y} = \{1, \ldots, K\}$. Note that $\mathcal{Y}$ is a set of shared class labels that is *common* to both the source and target domains. The *target domain dataset* $\mathcal{D}^t = \{P_i^t\}_{i=1}^{n_t}$ contains unlabeled point clouds. The cardinality of $\mathcal{D}^s$ and $\mathcal{D}^t$ are n_s and n_t respectively. Then, the task of UDA for point cloud classification boils down to learning a *domain invariant function* $f : \mathcal{P} \rightarrow \mathcal{Y}$, where $\mathcal{P}$ is a union of unlabeled point clouds from both $\mathcal{D}^s$ and $\mathcal{D}^t$.

3.1. Self-Supervised Contrastive Learning

Motivated by the advancement of contrastive learning [5, 17], where the goal is to pull samples from common classes closer in the embedding space, we build a method to extract 3D and 2D features of point clouds and fuse this information to form initial domain class clusters.

We employ a contrastive loss between augmented versions of a point cloud, which we term as a *3D-modal association loss*, to learn similar features for samples from the same class. This loss forces the point cloud learning to be invariant to geometric transformations. Additionally, we introduce a contrastive loss between the 3D point cloud features and their corresponding projected 2D image features, termed as *multi-modal association loss*. The intuition behind this multi-modal loss is to take advantage of the rich multi-view latent 2D information inherent in the 3D point clouds. Next, we explain these components in detail.

3D-modal association loss Let P_b be a point cloud from a randomly drawn batch B of size k from either $\mathcal{D}^s$ or $\mathcal{D}^t$. Given a set of affine transformations T, we generate two augmented point clouds $P_b^{t_1}$ and $P_b^{t_2}$, where t_1 and t_2 are compositions of transformations picked randomly from T. Additionally, we use random point dropout and add random noise to each point in a point cloud individually to introduce object surface distortions. These transformations introduce geometric variations, which are then used to curate samples

that serve as positive pairs. The augmented point clouds $P_b^{t_1}$ and $P_b^{t_2}$ are then mapped to a d-dimensional feature space using a 3D encoder function producing embeddings $z(P_b^{t_1})$ and $z(P_b^{t_2})$, respectively. These embeddings serve as *positive pairs* and therefore our objective is to place them closer to one another in the feature space.

We define the similarity between the i-th embedding transformed by t_x and the j-th embedding transformed by t_x, with $x \in \{1, 2\}$, as

$$\langle (i, t_x), (j, t_x) \rangle_{ST} = \exp\left(s(z(P_i^{t_x}), z(P_j^{t_x}))/\tau\right) \quad (1)$$

where s denotes the cosine-similarity function and τ is the temperature hyperparameter.

Our 3D-modal association loss is then given by

$$\mathcal{L}_{3d} = -\log\left\{\frac{\langle (i,t_1),(i,t_2)\rangle_{ST}}{\sum_{j=1}^{k}\langle (i,t_1),(j,t_1)\rangle_{ST} + \sum_{j=1}^{k}\langle (i,t_1),(j,t_2)\rangle_{ST}}\right\} \quad (2)$$

For both source and target, we randomly draw respective batches and perform 3D-modal association separately. This method of self-supervised contrastive learning generates class clusters in both domains individually and has been shown to be useful especially for the target domain, as its supervision signal is missing. We further guide the feature learning by introducing image modality in the optimization. We explain our multi-modal association loss next.

Multi-modal association loss We consider using point cloud projections in our method, as the image features can provide another level of discriminative information. 2D projections from various viewpoints allow capturing *silhouette* and *surface boundary* information for shape understanding that is harder to derive from just point-wise distances. Breaking away from the common way of fusing multimodal information [3, 16] where the embeddings of two modalities are fused by simply concatenating or averaging them, we instead compute associative losses between 3D features and image features to establish 2D-3D correspondence understanding helping to provide informative global representation.

As contrastive learning is known to be good for alignment tasks, we advocate using a contrastive objective to fuse multimodal (3D and 2D) information. Let $\mathcal{I}_P = \{I_n\}_{n=1}^m$ be the set of m 2D image projections of point cloud P. To generate these images, we set virtual cameras around the object in a circular fashion to obtain views of the object from all directions. For a point cloud P, each of its corresponding 2D images is passed to a 2D encoder, generating a d-dimensional embedding. Following [14, 25], we use a simple max-pooling operation to aggregate feature information from all views and get a d-dimensional vector z^{I_P}. In order to fuse the 3D augmented point cloud embeddings (i.e., $z(P^{t_1})$ and $z(P^{t_2})$) with the 2D point cloud embedding z^{I_P}, we compute the average of the 3D augmented point cloud embeddings to get z^{avg}. We then use the z^{avg} and z^{I_P} that contain summarized information from 3D and 2D modalities respectively in a self-supervised contrastive loss to maximize their similarity in the embedding space. We define the similarity between the i-th embedding z_i and the j-th embedding z'_j as $\langle z_i, z'_j \rangle_S = \exp\left(s(z_i, z'_j)/\tau\right)$. Then, our multi-modal association loss is given by

$$\mathcal{L}_{mm} = -\log\left\{\frac{\langle z_i^{avg}, z_i^{I_P}\rangle_S}{\sum_{j=1}^{k}\langle z_i^{avg}, z_j^{avg}\rangle_S + \sum_{j=1}^{k}\langle z_i^{avg}, z_j^{I_P}\rangle_S}\right\} \quad (3)$$

The total self-supervised contrastive loss is given by adding the 3D-modal association loss ($\mathcal{L}_{3d}$) that maximizes the similarity between augmentations of a point cloud and the multi-modal association loss ($\mathcal{L}_{mm}$) that maximizes the similarity between 3D and 2D features of a point cloud.

3.2. Optimal Transport and Wasserstein Distance

Optimal transport offers a way to compare two probability distributions irrespective of whether the measures have common support. It aims to find the most efficient way of transferring mass between two probability distributions, considering the underlying geometry of the probability space. Formally, given two probability distributions μ and ν on a metric space $\mathcal{X}$, for $p \geq 1$, the p-Wasserstein distance [26] is given by $W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x,y)^p d\pi(x,y)\right)^{1/p}$ where π is a transport plan that defines a flow between mass from μ to locations in ν, $\Pi(\mu, \nu)$ is the joint probability distribution with the marginals μ and ν and $c(x, y)$ is the ground metric which assigns a cost of moving a unit of mass $x \in \mathcal{X}$ from μ to some location $y \in \mathcal{X}$ in ν.

For the discrete case, given two discrete distributions $\hat{\mu} = \sum_{i=1}^m a_i \delta(x_i)$ and $\hat{\nu} = \sum_{j=1}^n b_j \delta(y_j)$, where $\{a_i\}_{i=1}^m$ and $\{b_j\}_{j=1}^n$ are the probability masses that should sum to 1, $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$ are the support points in $\mathbb{R}^d$ with m and n being the number of points in each measure. The discrete form of the above equation can be given as $W_p(\hat{\mu}, \hat{\nu}) = \left(\min_{\psi \in U(a,b)} \langle C^p, \psi \rangle_F\right)^{1/p}$, where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius dot-product, $C^p \in \mathbb{R}_+^{m \times n}$ is the pairwise ground metric distance, ψ is the coupling matrix and U is the set of all possible valid coupling matrices, i.e. $U(a, b) = \{\psi \in \mathbb{R}^{m \times n} : \psi \mathbb{1}_n = a, \psi^\top \mathbb{1}_m = b\}$.

3.3. Domain Alignment via Optimal Transport

As explained in Section 3.1, contrastive learning generates class clusters in source and target domains individually. The underlying idea is to further achieve alignment of point clouds belonging to the same class across two domains. We leverage an OT based loss that uses point cloud

features and source labels for domain alignment. The classifier $g : \mathbb{R}^d \rightarrow \mathcal{Y}$ that maps the point cloud embedding from feature space to label space also needs to work well for the target domain. The OT flow is greatly dependent on the choice of the cost function as shown by [7]. Here, as we want to jointly optimize the feature and the classifier decision boundary learning, we define our cost function as

$$\sum_{i=1}^{k}\sum_{j=1}^{k} c(z_i^s, z_j^t) = \alpha||z_i^s - z_j^t||_2^2 + \beta||y_i^s - g(z_j^t)||_2^2 \quad (4)$$

where superscripts s and t denote the source and target domains, respectively. α, β are the weight coefficients. Here, the first term computes the squared-$L2$ distance between the embeddings of source and target samples. The second term computes squared-$L2$ distance between the classifier's target class prediction and the source ground truth label. Jointly, these two terms play an important role in pulling or keeping apart source and target samples for achieving domain alignment. For example, if a target sample lies far from a source sample having the same class, the first term would give a high cost. However, for a decently trained classifier, the distance between its target class prediction and source ground truth label would be less, thus making the second term low. This indicates that these source and target samples must be pulled closer. Conversely, if a target sample lies close to a source sample having a different class, the first term would be low, and the second term would be high, indicating this sample should be kept apart. As evident from the example, the second term is a guiding entity for inter-domain class alignment. It penalizes source-target samples based on their classes and triggers a pulling mechanism. The problem of finding optimal matching can be formulated as $\psi^* = \min_{\psi \in U(a_s, b_t)} \langle C^p, \psi \rangle_F$, where ψ^* is the ideal coupling matrix, a_s and b_t are the uniform marginal distributions of source and target samples from a batch. The optimal coupling matrix ψ^* is computed by freezing the weights of the 3D encoder function and the classifier function g. The OT loss for domain alignment is given by

$$\mathcal{L}_{ot} = \sum_{i=1}^{k}\sum_{j=1}^{k} \psi_{ij}^*(\alpha||z_i^s - z_j^t||_2^2 + \beta\mathcal{L}_{ce}(y_i^s, g(z_j^t))) \quad (5)$$

where $\mathcal{L}_{ce}$ is the cross-entropy loss.

3.4. Overall Training Loss

The overall pipeline of our unsupervised DA method is trained with the combination of the following objective functions $\mathcal{L}_{total} = \mathcal{L}_{3d} + \mathcal{L}_{mm} + \mathcal{L}_{ot} + \mathcal{L}_{cls}^s$. The loss consists of three self-supervised losses (i.e., $\mathcal{L}_{3d}$, $\mathcal{L}_{mm}$ and $\mathcal{L}_{ot}$) and a supervised loss $\mathcal{L}_{cls}^s$. Besides three SSL tasks, supervised learning is performed based on source samples and labels. For this purpose, a regular cross-entropy loss or a *mixup* variant can be applied [30]. We use a supervised loss ($\mathcal{L}_{cls}^s$) inspired by the PointMixup method (PCM) [6]. PCM is a data augmentation method for point clouds by computing interpolation between samples. Augmentation strategies have proven to be effective and enhance the representation capabilities of the model. Similarly, PCM has shown its potential to generalize across domains and robustness to noise and geometric transformations.

We also employ the *self-paced self-training* (SPST) strategy introduced by [32] to improve the alignment between domains. In SPST, pseudo-labels for the target samples are generated using the classifier's prediction and confidence threshold. The first step computes the pseudo labels for the target samples depending on the confidence of their class predictions, while the next step updates the point cloud encoder and classifier with the computed pseudo labels for target and ground truth labels of source. In our method, we use SPST strategy as a fine-tuning step for our models.

4. Experiments

We conduct an exhaustive experimental study to show the effectiveness of the learned representations and the significance of our COT. Our model is evaluated on two benchmark datasets with and without the SPST strategy for the classification task. We consider recent state-of-the-art self-supervised methods such as DANN [12], PointDAN [21], RS [22], DefRec+PCM [1], GAST [32] and ImplicitPCDA [24] for comparison. Additionally, we report results for the baseline without adaptation (unsupervised) which trains the model using labels from the source domain and tests on the target domain. The supervised method is the upper bound which takes labels from the target domain into consideration during training. We will release our code upon acceptance.

4.1. Datasets

PointDA-10 introduced by [21] is a combination of ten common classes from ModelNet [28], ShapeNet [4] and ScanNet [8]. ModelNet and ShapeNet are synthetic datasets sampled from 3D CAD models, containing $4,183$ training, 856 test samples and $17,378$ training, $2,492$ test samples, respectively. On the other hand, ScanNet consists of point clouds from scanned and reconstructed real-world scenes and consists of $6,110$ training and $1,769$ test samples. Point clouds in ScanNet are usually incomplete because of occlusion by surrounding objects in the scene or self-occlusion in addition to realistic sensor noises. We follow the standard data preparation procedure used in [1, 21, 24, 32].

GraspNetPC-10 [24] consists of synthetic and real-world point clouds for ten object classes. It is developed from GraspNet [10] by re-projecting raw depth scans to 3D space and applying object segmentation masks to crop out the corresponding point clouds. Raw depth scans are captured by

Methods	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 [12]		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 [21]		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 [22]		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 [32]		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
ImplicitPCDA [24]		**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
COT		83.2 ± 0.3	54.6 ± 0.1	78.5 ± 0.4	53.3 ± 1.1	**79.4** ± 0.4	**77.4** ± 0.5	**71.0**
	✓	84.7 ± 0.2	57.6 ± 0.2	**89.6** ± 1.4	51.6 ± 0.8	**85.5** ± 2.2	**77.6** ± 0.5	**74.4**

Table 1. Classification accuracy (%) on the PointDA-10. M: ModelNet, S: ShapNet, S*: ScanNet; → indicates the adaptation direction. OT: Optimal transport, SPST: self-paced self-training. Results in black and blue represent accuracy without and with SPST strategy, respectively. Bold represents the best result and underlined represents the second best for both the colors.

Methods	SPST	Syn. → Kin.	Syn → RS.	Kin. → RS.	RS. → Kin.	Avg.
Supervised		97.2 ± 0.8	95.6 ± 0.4	95.6 ± 0.3	97.2 ± 0.4	96.4
Baseline(w/o adap.)		61.3 ± 1.0	54.4 ± 0.9	53.4 ± 1.3	68.5 ± 0.5	59.4
DANN [12]		78.6 ± 0.3	70.3 ± 0.5	46.1 ± 2.2	67.9 ± 0.3	65.7
PointDAN [21]		77.0 ± 0.2	72.5 ± 0.3	65.9 ± 1.2	82.3 ± 0.5	74.4
RS [22]		67.3 ± 0.4	58.6 ± 0.8	55.7 ± 1.5	69.6 ± 0.4	62.8
Defrec+PCM [1]		80.7 ± 0.1	70.5 ± 0.4	65.1 ± 0.3	77.7 ± 1.2	73.5
GAST [32]		69.8 ± 0.4	61.3 ± 0.3	58.7 ± 1.0	70.6 ± 0.3	65.1
	✓	81.3± 1.8	72.3 ± 0.8	61.3 ± 0.9	80.1 ± 0.5	73.8
ImplicitPCDA [24]		81.2 ± 0.3	73.1 ± 0.2	66.4 ± 0.5	82.6 ± 0.4	75.8
	✓	94.6 ± 0.4	80.5 ± 0.2	76.8 ± 0.4	85.9 ± 0.3	84.4
COT		**87.7** ± 0.7	**80.2** ± 2.1	**69.3** ± 5.2	**85.8** ± 4.3	**80.0**
	✓	**98.2** ± 0.5	**83.7** ± 0.2	**81.9** ± 2.1	**98.0** ± 0.1	**91.0**

Table 2. Classification accuracy (%) on the GraspNet-10 dataset. Sys.: Synthetic domain, Kin.: Kinect domain, RS.: Real domain; → indicates the adaptation direction. OT: Optimal transport, and SPST: self-paced self-training. Results in black and blue represent accuracy without and with SPST strategy, respectively. Bold represents the best result and underlined represents the second best for both the colors.

two different depth cameras, Kinect2 and Intel Realsense to generate real-world point clouds. In the Synthetic, Kinect, and RealSense domains, there are $12,000$ training, $10,973$ training, $2,560$ testing, and $10,698$ training, $2,560$ testing point clouds, respectively. There exist different levels of geometric distortions and missing parts. Unlike PointDA-10, point clouds in GraspNetPC-10 are not aligned and all domains have almost uniform class distribution.

Implementation Details We use DGCNN [27] as the point cloud feature extractor and pre-trained ResNet-50 [15] as the feature extractor for images to get 1024-dimensional embedding vectors. For the contrastive losses ($\mathcal{L}_{3d}$, $\mathcal{L}_{mm}$) we convert these 1024-dimensional embeddings to 256 dimensions using projection layers. The classifier network consists of three fully connected layers with dropout and batch normalization. We use rendered point cloud images of size 224×224 and set the number of views to 12. In total, we train our models for 150 epochs for PointDA-10 and 120 epochs for GraspNetPC-10 with a batchsize of 32 on NVIDIA RTX-2080Ti GPUs and perform three runs with different seeds. We report results from the model with the best classification accuracy on source validation set, as target labels are unavailable. We provide more details about the implementation setup in our supplementary material.

4.2. Unsupervised DA: Classification

In Tables (1, 2), we compare the results of our COT with the existing point cloud domain adaptation methods [1, 21, 24, 32] on PointDA-10 and GraspNetPC-10 datasets respectively. Similar to [24] and [32], we also test our methodology with SPST strategy. As shown in Table 1, COT achieves SoTA performance in terms of the overall average performance on PointDA-10 dataset. We observe that COT beats existing methods by a huge margin when the target dataset is synthetic. This is because target point clouds have well-defined geometry, and the classifier can make accurate predictions with high confidence, thus majorly help-

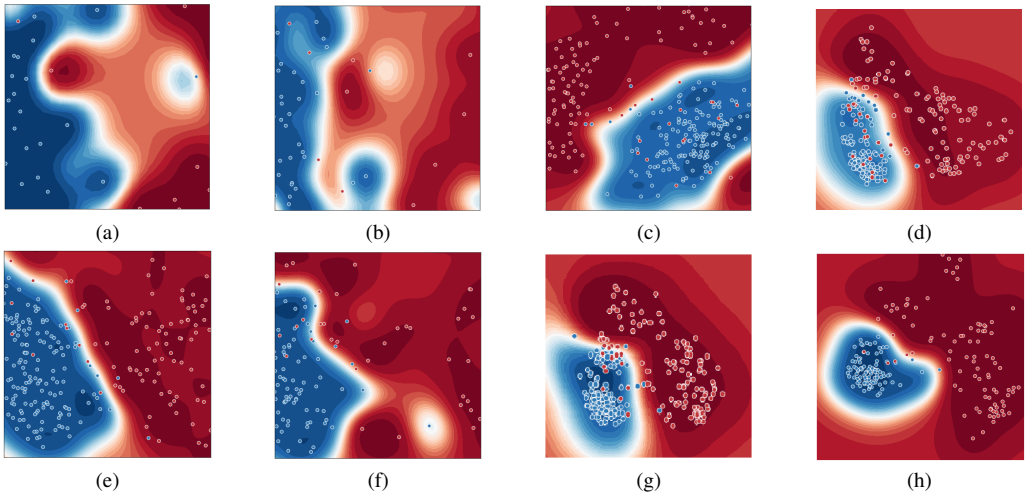

Figure 3. Early (top-row) and final (bottom-row) epochs decision boundaries on target samples for One-vs-Rest (Monitor class) for $S \rightarrow M$. (a), (e) Only PCM (without adaptation), (b), (f) Contrastive learning with PCM, (c), (g) Optimal transport and contrastive learning with PCM (Our COT) and (d), (h) Our COT fine-tuned with SPST.

ing alignment. As existing methods only propose to use self-learning tasks, their performance is very low compared to our self-learning task with explicit domain alignment endowed by OT. For the settings where the target dataset is real, it becomes harder for the classifier to provide good predictions, making the alignment process noisy. In these settings, we achieve on-par results compared to the existing methods. In $S \rightarrow M$, our method with SPST strategy outperforms existing methods $\approx 8\%$, and in $M \rightarrow S$, we achieve on-par results compared to the existing methods. We also use t-SNE to visualize the learned features of both domains (shown in supplementary). For PointDA-10, we observe that when the target domain is synthetic, the learned features are distinctive; however, when the target domain is real, the features lack distinctive power. This portrays the challenging setting of synthetic to real adaptation. Overall we achieve the highest average accuracy on PointDA-10 dataset showing effectiveness of COT.

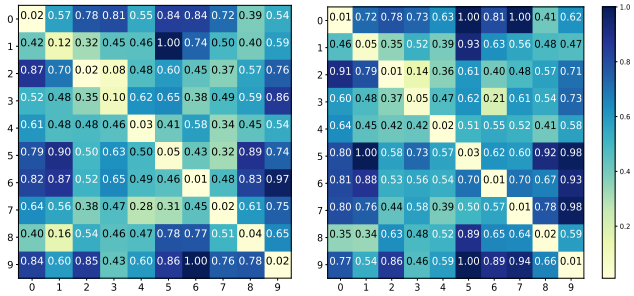

Figure 4. Class-wise MMD for $S \rightarrow M$ for (a) baseline (only PCM w/o adaptation), and (b) our COT with SPST.

Our method outperforms all the existing methods with a significant margin on all the combinations of GraspNetPC-10 dataset, as shown in Table 2. COT beats existing methods in both with and without SPST strategy; also, in some cases, it beats the supervised method (upper bound). It is interesting to note the difference in behaviour of COT and other methods on real-world data in PointDA-10 and GraspNetPC-10. PointDA-10, in general, has a very skewed class-wise sample distribution and has a small set of real-world samples. Whereas, GraspNetPC-10 has almost uniform class-wise sample distribution with approximately double the size of ScanNet. COT performs significantly better with larger datasets and almost equal class-wise sample distribution. Existing methods that propose classification-based [32] or geometry-aware implicit learning-based [24] tasks fall short in terms of performance boost compared to COT when real-world datasets are large and have uniform class distribution. This shows the effectiveness of COT for unsupervised domain adaptation achieving SoTA performance on real-world data from GraspNet-10 dataset.

4.3. Domain Alignment

In this section, we discuss our used sampling strategy for creating a batch and explain its working in our $\mathcal{L}_{ot}$ loss for domain alignment. For every iteration, we use random sampling to draw source and target batches independently. Note that it does not ensure the coherence of source and target classes in a batch. Using these batches, the OT flow finds the best one-to-one matching amongst both domains using the defined cost function and updates both network's (encoder and classifier) weights to minimize the $\mathcal{L}_{ot}$ loss. Even though we use random sampling we find that repeat-

$\mathcal{L}_{3d}$	$\mathcal{L}_{ot}$	$\mathcal{L}_{mm}$	SPST	M → S	M → S*	S → M	S → S*	S* → M	S* → S	Avg.
✓	✓			82.50	53.82	74.65	47.26	75.35	71.39	67.5
✓		✓		82.66	46.64	**78.50**	**53.82**	**82.24**	75.40	69.9
✓	✓	✓		**83.20**	**54.61**	**78.50**	53.30	79.44	**77.41**	**71.0**
✓	✓		✓	**84.91**	56.76	84.93	47.26	77.22	73.07	70.7
✓		✓	✓	**84.91**	54.32	85.51	**53.31**	**86.0**	75.92	73.3
✓	✓	✓	✓	84.71	**57.66**	**89.60**	51.61	85.50	**77.69**	**74.4**

Table 3. Ablation Study: Target classification accuracy for UDA task on PointDA-10 dataset. Bold represents best results.

ing this process for multiple iterations eventually converges the overall alignment loss ($\mathcal{L}_{ot}$) giving discriminative features for classes with aligned source and target distributions. For examining the distance between class clusters from the source and target, we compute the maximum mean discrepancy (MMD) between learned point cloud features. In Figure 4, we show class-wise MMD, where Figures 4a, 4b are for baseline (without adaptation) and our COT respectively on ShapeNet to ModelNet. The diagonal of the matrix represents MMD between the same classes from source and target, and the upper and lower triangular matrices represent MMD between different classes for source and target. It is clearly evident that the MMD matrix for our COT has higher distances in the upper and lower triangular regions than the baseline. This shows that classes within the source and target individually are well separated. Further, the diagonal values for our COT are lower than the baseline without adaptation, indicating that the same classes in source and target are closer for features obtained from our method. Overall, we can see that point cloud embeddings generated by COT have better inter-class distances and source and target class alignment.

4.4. Discussion: Decision Boundary

We also examine the decision boundaries of our learned models. Figure 3 illustrates the decision boundaries from early (top-row) and final (bottom-row) epochs for four variants of our model. For this experiment, we select target samples from the hidden space of our trained models. We consider four variants of our model, i.e., i) only PCM (no adaptation), ii) contrastive learning with PCM, iii) contrastive learning and OT with PCM (our COT method), and iv) our COT fine-tuned with SPST strategy. All the representations are retrieved with the labels predicted by our trained model. Next, we fit the SVM and consider a "one-vs-rest strategy" to visualize the decision boundaries.

From Figures 3a to 3d and 3e to 3h, we can clearly interpret that the baseline model with only PCM and no adaptation leads to irregular boundaries in Figures 3a and 3e. The representations are enhanced, and the boundary becomes smoother by applying contrastive learning to both the domains in Figures 3b and 3f. In contrast, training the model with our COT, which includes the previous two strategies (PCM and contrastive learning) along with OT loss further improves the decision boundaries in Figures 3c and 3g. Finally, with the SPST strategy, which finetunes the COT with pseudo labels of target samples, the region gets even more compact and smoother in Figures 3d and 3h. This shows that contrastive learning separates the two classes which are improved by OT alignment. Also, SPST further makes the classes more compact and achieves the best results.

4.5. Ablation Studies

We perform ablation studies to understand the significance of proposed losses in our method. In Table 3, we compare the results of our COT trained with various components on PointDA-10. $\mathcal{L}_{3d}$ is always used as it is our base self-learning task for 3D point clouds. The significance of $\mathcal{L}_{ot}$ can be seen by comparing row 3 and row 2. When $\mathcal{L}_{ot}$ is removed from COT, the performance drops on almost all settings. Comparing row 3 and row 1, we can see the effect of $\mathcal{L}_{mm}$ as the performance decreases for all settings when it is turned off. In both cases, the average accuracy also drops. This indicates positive contribution of both $\mathcal{L}_{ot}$ and $\mathcal{L}_{mm}$ in the formulation of our COT. A similar trend is also observed with the SPST strategy as well. Comparing row 6 with rows 4 and 5, we see the best performance when both losses are used. Also, note that SPST increases the performance for all three settings shown. Overall, these results suggest that both image modality and OT-based domain alignment are crucial for achieving the best results.

5. Conclusion

In this work, we tackled the domain adaptation problem on 3D point clouds for classification. We introduced a novel methodology to synergize contrastive learning and optimal transport for effective UDA. Our method focuses on reducing the domain shift and learning high-quality transferable point cloud embeddings. Our empirical study reveals the effectiveness of COT as it outperforms existing methods in overall average accuracy on one dataset, and achieves SoTA performance on another. The conducted ablation studies demonstrate the significance of our proposed method. An interesting future direction would be to extend our OT-based approach for UDA of point clouds on more complex tasks like segmentation or object detection in indoor scenes.

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, 5, 6

[2] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, A. Dharmasiri, K. Thilakarathna, and R. Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. *Ieee/cvf Conference On Computer Vision And Pattern Recognition (cvpr)*, 2022. 2

[3] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *Int. J. Comput. Vision*, 123(1):4–31, may 2017. 4

[4] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv: Arxiv-1512.03012*, 2015. 5

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning (ICML)*, pages 1597–1607. PMLR, 2020. 2, 3

[6] Yunlu Chen, Vincent Tao Hu, Efstratios Gavves, Thomas Mensink, Pascal Mettes, Pengwan Yang, and Cees G. M. Snoek. Pointmixup: Augmentation for point clouds. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 330–345, Cham, 2020. Springer International Publishing. 2, 5

[7] Marco Cuturi and David Avis. Ground metric learning. *J. Mach. Learn. Res.*, 15(1):533–564, jan 2014. 5

[8] Angela Dai, Angel X. Chang, M. Savva, Maciej Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *Ieee Conference On Computer Vision And Pattern Recognition (cvpr)*, 2017. 5

[9] B. Damodaran, B. Kellenberger, Rémi Flamary, D. Tuia, and N. Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. *ECCV*, 2018. 2

[10] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11441–11450, 2020. 2, 5

[11] Kilian Fatras, Thibault S'ejourn'e, N. Courty, and Rémi Flamary. Unbalanced minibatch optimal transport; applications to domain adaptation. *ICML*, 2021. 2

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. 5, 6

[13] Léo Gautheron, Ievgen Redko, and Carole Lartizien. Feature selection for unsupervised domain adaptation using optimal transport. *arXiv preprint arXiv: Arxiv-1806.10861*, 2018. 2

[14] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtn: Multi-view transformation network for 3d shape recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, October 2021. 4

[15] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Ieee Conference On Computer Vision And Pattern Recognition (cvpr)*, 2015. 6

[16] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20, 2017. Language in Vision. 4

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18661–18673, 2020. 2, 3

[18] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5667–5675, 2018. 1

[19] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928, 2015. 1

[20] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[21] 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. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2, 5, 6

[22] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 5, 6

[23] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. *AAAI*, 2017. 2

[24] Yuefan Shen, Yanchao Yang, Mi Yan, He Wang, Youyi Zheng, and Leonidas J. Guibas. Domain adaptation on point clouds via geometry-aware implicits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7223–7232, June 2022. 2, 5, 6, 7

[25] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proc. ICCV*, 2015. 1, 4

[26] Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003. 4

[27] 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 Trans. Graph.*, 38(5), oct 2019. 6

[28] Zhirong Wu, S. Song, A. Khosla, F. Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. *Ieee Conference On Computer Vision And Pattern Recognition (cvpr)*, 2014. 1, 5

[29] Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, and Jindong Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[30] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 5

[31] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *IEEE international conference on robotics and automation (ICRA)*, pages 3357–3364, 2017. 1

[32] 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, 5, 6, 7