

Mapping in a Cycle: Sinkhorn Regularized Unsupervised Learning for Point Cloud Shapes

Lei Yang¹, Wenxi Liu^{2,1}, Zhiming Cui¹, Nenglun Chen¹, and Wenping Wang¹

¹ Department of Computer Science, The University of Hong Kong, China
{lyang, zmcui, nlchen, wenping}@cs.hku.hk

² College of Mathematics and Computer Science, Fuzhou University, China
wenxi.liu@hotmail.com

Abstract. We propose an unsupervised learning framework with the pretext task of finding dense correspondences between point cloud shapes from the same category based on the cycle-consistency formulation. In order to learn discriminative pointwise features from point cloud data, we incorporate in the formulation a regularization term based on Sinkhorn normalization to enhance the learned pointwise mappings to be as bijective as possible. Besides, a random rigid transform of the source shape is introduced to form a triplet cycle to improve the model’s robustness against perturbations. Comprehensive experiments demonstrate that the learned pointwise features through our framework benefits various point cloud analysis tasks, e.g. partial shape registration and keypoint transfer. We also show that the learned pointwise features can be leveraged by supervised methods to improve the part segmentation performance with either the full training dataset or just a small portion of it.

Keywords: Point cloud, unsupervised learning, dense correspondence, cycle-consistency

1 Introduction

Point clouds are unordered sets of interacting points sampled from surface of objects for 3D shape representation, and have been widely used in computer vision, graphics, robotics, etc. for their accessibility and flexibility. With the recent advancement of deep learning techniques, a spectrum of networks have been proposed to process point cloud data and to learn to perform various tasks, e.g. [32, 33, 26, 25, 47], which have achieved tremendous progress. However, a major limitation of deep networks is their data hunger nature that requires a large amount of supervisory signals to learn a satisfactory task-specific model. Therefore, many attempts have been made to alleviate this issue, and among others training deep networks in an unsupervised manner (without manually labeled data) shows its potential in many scenarios [10, 5, 29]. In the case of 3D point clouds, these techniques are in demand as it is prohibitive to attain accurate, densely labeled ground-truth on point clouds for various shape analysis tasks.

As one of unsupervised learning approaches, algorithms based on cycle consistency have attracted interests in many vision-based applications, e.g., video

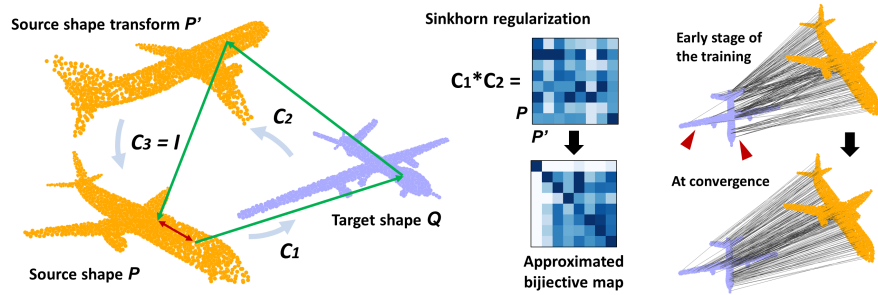


Fig. 1. We train a neural network in an unsupervised manner to derive dense correspondences between point cloud shapes based on the cycle-consistency formulation. Given a source shape \mathcal{P} , its rigid transform \mathcal{P}' and a target one \mathcal{Q} , the cycle is completed by mapping every point in \mathcal{P} , via C_1 , C_2 , C_3 , and finally back to \mathcal{P} (left). The red segment indicates the measure of the cycle deviation which is minimized during the unsupervised training. Within the cycle, the correspondence map formed by the mappings C_1 and C_2 is constrained to approximate a bijective map by Sinkhorn regularization, significantly reducing the number of many-to-one correspondences (right).

object segmentation [44] and facial landmark detection [39], as well as some recent 3D shape analysis works [19, 55, 46, 13]. Intuitively, a cycle consistent transformation between a pair of data instances shall map one data instance to the other, and then transform it backward with little deviation. With a pretext defined, such as registration [46] or deformation [13], one can leverage the cycle consistency formulation between a pair of unlabeled data, model the transformation with a neural network, and thus optimize the network parameters by minimizing the cycle deviation.

In this work, we leverage such a formulation to pre-train the neural network in an unsupervised manner, and aim to learn pointwise features for subsequent 3D point cloud applications. Specifically, the pretext in our setting is to find dense correspondences between two point cloud shapes using the learned pointwise features. In particular, given a pair of source and target shapes, we intend to find, for each point in the source, its corresponding point in the target. Then, starting from the target ones, we search reversely their corresponding points in the source. During this process we minimize the cycle deviation of each reversely corresponded point from its original location in the source. In this way, the network parameters can be optimized and expected to encode each point to a high-dimensional descriptor for correspondence query.

While dense correspondences between contents have been exploited in many image-based applications (e.g. [44, 38]), this self-supervised framework encounters two major challenges when dealing with point cloud data. First, since the point clouds are usually sampled from smooth 3D surfaces, each point embeds very limited information as opposite to image pixels with rich textures. This precludes the network training based on cycle consistency as the obtained correspondence may map a point to a wrong but spatially proximate location, forming

a many-to-one mapping while yielding a small loss. Thus, the learned representation may suffer from this sub-optimality and fail to attain sufficient distinctiveness for correspondence query and potential applications. Second, many point-based networks based on cycle consistency assume the shapes are well aligned, and thus are sensitive to rotations. This makes the extracted features unrobust to small perturbations, and may become less applicable in many applications.

To address the first and primary concern, we propose a novel regularization technique to strengthen the pointwise correspondences to be as bijective as possible. We thus impose a bijective constraint on the cycle-back correspondence by adopting the Sinkhorn normalization technique [37, 24]. We term this constraint as Sinkhorn regularization in our paper.

We also introduce, into the cycle, an additional shape which is a random rigid transform of the source shape, forming a 3-edge cycle as shown in Fig. 1. In this particular setting, each point starting from the source first finds its corresponding point in the target (i.e. C_1 in Fig. 1), and then arrives at the source transform (i.e. C_2). Since the last transport edge (i.e. C_3) from the source transform to its origin provides us the ground-truth dense correspondences that form a one-to-one map, we can safely impose the bijective constraint by Sinkhorn regularization on this particular edge without assuming any shape pair should meet the bijective constraint. Further, unlike traditional cycle consistency methods on shapes, the introduction of a transformed shape allows the network and the learned pointwise features to be less sensitive against non-aligned point cloud data. This partially addresses the second challenge as mentioned before, thus making the learned pointwise features appealing to many downstream tasks.

To demonstrate the effectiveness of our proposed framework, we conduct comprehensive experiments, in which we leverage the pointwise features learned from our model to perform partial shape registration, keypoint transfer, and as an additional pre-trained feature for supervised part segmentation. In these experiments, it is demonstrated that our approach can surpass the state-of-the-art methods or be comparable to their performances. Contributions of this paper are summarized as follows:

- 1) A novel unsupervised learning framework based on cycle-consistent dense correspondences between point cloud shapes of the same category;
- 2) The Sinkhorn regularization that brings a large improvement on learning discriminative pointwise features;
- 3) Extensive evaluation showing the effectiveness of the proposed unsupervised learning strategy in multiple point cloud analysis tasks.

2 Related Work

Deep unsupervised methodology and applications. Unsupervised learning methodology has emerged to address one of the major concerns for data-driven algorithms – the need for a large set of labeled data, and has achieved state-of-the-art performances in many applications in language [10] and images [5, 17]. To this end, a pretext is often required for network pretraining and representation

learning, such as by contrastive learning [31, 50, 40, 29, 17], mutual information maximization [41], reconstruction [36], and correspondence [39, 11, 44].

Deep unsupervised point cloud analysis. Point-based networks have demonstrated their capability in many shape analysis tasks, such as classification, segmentation, reconstruction and registration [32, 33, 48, 26, 47, 9, 53]. To enforce the network to learn semantically meaningful and consistent representations, many pretext tasks have been designed such as part re-arranging [36], half-to-half prediction [15], deformation [12, 13], and self-supervised classification and clustering [16]. Many of these works rely on the reconstruction metric as an indicator for the unsupervised training process [12, 54, 9]. In this paper, we provide an alternative viewpoint that is complementary to the prior works. Instead of geometric reconstruction of the content, we consider the pretext of finding dense correspondences between shapes, and solve it as a *soft permutation recovery problem* for the indices of points in point cloud.

Learning from 3D shape correspondences. 3D shape correspondence has long been an exciting topic in 3D shape modeling [42, 22, 21, 20, 35, 14]. Many state-of-the-art works have leveraged dense correspondences to learn and perform various shape analysis tasks. [7, 8, 53, 6] design network architectures to learn local contextual descriptors via the task of 3D scan registration. This is amiable especially in the case of scene understanding and processing. As for the analysis of man-made shapes, [45, 46] and [2] instill classical methodologies (e.g. iterative closest point) in the neural network design and achieve state-of-the-art performance. In our case, we are more focused on learning pointwise features that are consistent across man-made shapes and thus differ from these studies.

In pursuit of such pointwise latent representations of man-made shapes, [18, 4, 30] make use of rough dense shape correspondence as supervision and demonstrate promising performance in shape correspondence. Alternatively, cycle consistency, initially proposed in [19], has been widely employed to establish correspondences between diverse data in an unsupervised manner, on images [56, 55], videos [44, 11], and more recently on point cloud data [13].

In line of these prior works, we build our unsupervised learning framework based on cycle consistency to process point cloud shapes. Different from prior arts [13][46] that evaluate cycle consistency by measuring shape discrepancy, we innovatively cast the problem of finding dense correspondences as solving permutation on point clouds. This particular design provides an alternative view to existing works and allows the network to learn a pointwise representation. While the unsupervised learning works [13, 46] focus on their pretexts such as deformation and registration, we show a variety of applications with the proposed network as well as the learned pointwise representation. We further propose a novel Sinkhorn regularization in the cycle consistency framework to enforce the learned pointwise features to be sparsely corresponded for different instances.

3 Methodology

Our overarching goal is to learn, without manual labels, a category-specific point-wise encoding that benefits the downstream applications such as shape registration, keypoint detection and part segmentation. To this end, we train the network via a pretext task of finding dense correspondences between two point cloud shape instances based on the cycle-consistent pointwise encoding features learned through the proposed framework.

3.1 Unsupervised loss based on cycle consistency

The pretext of finding cycle-consistent dense correspondences is depicted in Fig. 1. We denote the source shape \mathcal{P} , its random rigid transform \mathcal{P}' , and the target \mathcal{Q} , forming a 3-edge cycle from \mathcal{P} to \mathcal{Q} (mapping C_1), then from \mathcal{Q} to \mathcal{P}' (mapping C_2), and finally return to \mathcal{P} from \mathcal{P}' (mapping C_3). In order to formulate it as an optimization problem, with the dense correspondences between shapes, our goal is to minimize the deviation between each cycle-back point and its origin (i.e. the red segment in Fig. 1), thus enforcing the cycle consistency.

Correspondence query between point cloud shapes. We use a point-based neural network, denoted as f_θ with trainable parameters θ , to learn the pointwise features, which will be employed for the correspondence query. We denote a point using its index as p_k where $p_k \in \{0, 1\}^{|\mathcal{P}|}$ is a one-hot vector with the k -th entry equal to 1 and the rest to 0. The corresponding 3D coordinate of p_k is denoted by \mathbf{p}_k . If a particular point q_i from shape \mathcal{Q} is said to correspond to p_k from \mathcal{P} , then the associated learned representation $f_\theta(q_i)$ is more similar with $f_\theta(p_k)$ than all other points in \mathcal{Q} as below:

$$i = \arg \max_{q_j \in \mathcal{Q}} S(f_\theta(p_k), f_\theta(q_j)), \quad (1)$$

where $S(\cdot, \cdot)$ measures the similarity between any two pointwise representations and is defined as their inner product. Since the operator $\arg \max$ is not differentiable, we approximate the solution to the above equation by a scaled softmax function:

$$q_i \approx C(\mathcal{Q}, p_k; f_\theta) = \frac{\exp(f_\theta(p_k)^T f_\theta(q_j)/\tau)}{\sum_j \exp(f_\theta(p_k)^T f_\theta(q_j)/\tau)}, \quad (2)$$

where $C(\mathcal{Q}, p_k; f_\theta)$ is a vector that represents the probability that p_k corresponds to all points in \mathcal{Q} . Thus, the dense correspondences from \mathcal{P} to \mathcal{Q} can be approximated as follow:

$$\mathbf{Q} \approx C(\mathcal{Q}, \mathcal{P}; f_\theta). \quad (3)$$

where, ideally, \mathbf{Q} is expected to be a permutation matrix, establishing a one-to-one mapping between two given shapes \mathcal{P} and \mathcal{Q} .

Cycle-consistency loss. In this paper, we use three shapes to form a 3-edge cycle $\{\mathcal{P} \rightarrow \mathcal{Q} \rightarrow \mathcal{P}' \rightarrow \mathcal{P}\}$, where \mathcal{P} and \mathcal{Q} are termed the source and the target shapes, respectively, and \mathcal{P}' is a random rigid transform of \mathcal{P} that helps

increase the robustness of the model. Thus, with the cycle-consistency condition met, this closed cycle should finally bring every point (in terms of index and not spatial coordinates) back to its origin index via the following mappings,

$$C_{cycle}(\mathcal{P}) = C_3(\mathcal{P}, \mathcal{P}')C_2(\mathcal{P}', \mathcal{Q})C_1(\mathcal{Q}, \mathcal{P}) = C_3(\mathcal{P}, \mathcal{P}')C_{1,2}(\mathcal{P}', \mathcal{P}), \quad (4)$$

where $C_{cycle}(\mathcal{P})$ shall be the identity matrix that brings points in \mathcal{P} back to its origin index via C_1 , C_2 , and C_3 . Similarly, $C_{1,2}(\mathcal{P}', \mathcal{P})$ forms the mapping from the source shape \mathcal{P} to the transformed shape \mathcal{P}' via \mathcal{Q} . To measure the cycle deviation from the above formulation, a loss should be defined

$$d_{cycle} = D(\mathbf{I}_{|\mathcal{P}|}, C_{cycle}(\mathcal{P})), \quad (5)$$

where $\mathbf{I}_{|\mathcal{P}|}$ is the identity matrix of size $|\mathcal{P}|$.

As we introduce a rigid transform to the end of the cycle list, the cycle mapping mainly depends on two parts, i.e., $C_{1,2}(\mathcal{P}', \mathcal{P})$ and $C_3(\mathcal{P}, \mathcal{P}')$, in Eq. 4. First, as rigid transformations in \mathbb{R}^3 do not alter the permutation of the point cloud data. So, when it is perfectly estimated, $C_3(\mathcal{P}, \mathcal{P}')$ should be the identity matrix that maintains the original permutation of \mathcal{P} . On the other hand, the mapping $C_{1,2}$ from \mathcal{P} to \mathcal{P}' (via \mathcal{Q}), in an ideal situation, should be the identity matrix as well. Hence, the cycle loss minimization can be reduced to minimize two terms, $D(\mathbf{I}_{|\mathcal{P}|}, C_{1,2}(\mathcal{P}', \mathcal{P}))$ and $D(\mathbf{I}_{|\mathcal{P}|}, C_3(\mathcal{P}, \mathcal{P}'))$.

One way to concretely define $D(\cdot, \cdot)$ is to use KL-divergence or cross-entropy losses to formulate the problem as classification. However, minimizing such *thousand-way* classification losses may be difficult at the beginning and overlook, during the course of optimization, the underlying geometric relationship of the points cloud shapes. Therefore, we cast the cycle consistency loss in a regression form similar to the losses used in [11, 38]. This way, we impose a soft penalty on the wrong cycles relying on the distances from their correct correspondences,

$$\mathcal{L}_C = \|D(\mathcal{P}) \otimes C_{1,2}(\mathcal{P}', \mathcal{P})\|_1, \quad (6)$$

where $D(\mathcal{P}) = \{d_{p,p'} = d_{Euclid}(\mathbf{p}, \mathbf{p}'), \forall \mathbf{p}, \mathbf{p}' \in \mathcal{P}\}$ measures the Euclidean distance between a pair of points in \mathcal{P} and \otimes is the element-wise product. Here the Euclidean distance is adopted for simplicity and computational efficiency, but one may employ more accurate geodesics distance for training. Note that the diagonals of $D(\mathcal{P})$ are zeros, which makes the loss of Eq. 6 to be zero when $C_{1,2}(\mathcal{P}', \mathcal{P})$ converges to be an identity matrix. This loss is thus equivalent to the classification-based formulation at convergence, while additionally taking the spatial proximity of points into consideration. Similarly, we formulate $D(\mathbf{I}_{|\mathcal{P}|}, C_3(\mathcal{P}, \mathcal{P}'))$ as follow:

$$\mathcal{L}_R = \|D(\mathcal{P}) \otimes C_3(\mathcal{P}, \mathcal{P}')\|_1. \quad (7)$$

3.2 Sinkhorn regularization for bijective constraint

Optimizing the regression-based cycle loss can converge to the correct correspondences as demonstrated in many image-based applications [11, 38]. However, the

convergence will be slowed down or even get stuck in the case of 3D point cloud data. This is because the decaying distance-based penalty imposed in Eq. 6 cannot provide a sufficient magnitude of loss that encourages the network to distinguish nearby points as the optimization proceeds. Thus, it may still result in many-to-one mappings and thus wrong cycles, leading to undesirable results.

To address this issue, we introduce a so-called Sinkhorn regularization term, L_S , in addition to the previous ones. This design relies on the fact that $C_{1,2}(\mathcal{P}', \mathcal{P})$ in our setting should ideally form a bijective map. Instead of directly enforcing $C_{1,2}$ to be the identity matrix, we *relax* this constraint to any permutation matrices, retaining the bijective property. The reason of using a relaxed bijective map instead of the identity is that while this relaxation penalizes the deviation of $C_{1,2}$ from a permutation, the synergistic effect of L_S and L_C gradually makes $C_{1,2}$ converge to the identity as the training proceeds. This novel relaxation brings the performance gain by a large margin in terms of the percentage of correct cycle-consistent correspondences, as shown in the ablation study (Sec. 4.2).

We follow the methods proposed in [1, 28] to enforce this constraint and describe it to make our paper self-contained. One may compute the optimal approximant to $C_{1,2}$ from the permutation set \mathbb{P} with dimension $|\mathcal{P}|$,

$$X^* = \arg \max_{X \in \mathbb{P}_{|\mathcal{P}|}} \langle X, C_{1,2} \rangle_F, \quad (8)$$

where $\langle X, C_{1,2} \rangle_F$ denotes the Frobenius inner product of the two matrices. As solving the above linear assignment problem (Eq. 8) is generally NP-hard, the constraint can be further relaxed to solve the best approximation of $C_{1,2}$ from the set of doubly stochastic matrices $\mathbb{B}_{|\mathcal{P}|}$,

$$\tilde{X} = \arg \max_{X \in \mathbb{B}_{|\mathcal{P}|}} \langle X, C_{1,2} \rangle_F. \quad (9)$$

Solving the maximization problem of Eq. 9 has been shown to be exceptionally simple by taking row-wise and column-wise softmax normalizations in an alternating fashion. This is known as the Sinkhorn normalization [37] where \tilde{X} shall meet the following conditions at convergence:

$$\tilde{X}\mathbf{1} = \mathbf{1}, \quad \tilde{X}^T\mathbf{1} = \mathbf{1}, \quad \text{and } \tilde{X} \in \mathbb{B}_{|\mathcal{P}|}.$$

While the solution to Eq. 9 can be reached in a limit sense, practically a truncated Sinkhorn normalization [1] is used to obtain a fair approximation,

$$\tilde{X} \approx SH(C_{1,2}; t, l), \quad (10)$$

where two hyper-parameters, i.e., the number of iterations l for the column-wise and row-wise normalization and the temperature t for softmax normalization are to be furnished. We adopt this truncated Sinkhorn normalization and set t and l to be 0.3 and 30 across all our experiments.

Sinkhorn regularization. Accordingly, during the network optimization we add the following Sinkhorn regularization to the loss function:

$$\mathcal{L}_S = \|C_{1,2}(\mathcal{P}', \mathcal{P}) - SH(C_{1,2}; t, l)\|_1. \quad (11)$$

This loss term enforces $C_{1,2}$ to be a bijective map, and thus encourages the neural network f to learn discriminative pointwise features by reducing many-to-one mappings between point clouds \mathcal{P} and \mathcal{Q} which $C_{1,2}$ traverses. As it is derived based on $C_{1,2}$, $SH(C_{1,2}; t, l)$ keeps its closeness to $C_{1,2}$ at every iteration step. Thus, this formulation provides a gradual guidance for $C_{1,2}$ to become a permutation matrix ensuring the bijective property.

3.3 Loss function

To sum up, the loss function of our unsupervised framework consists of three terms, as below:

$$\mathcal{L} = \lambda_C \mathcal{L}_C + \lambda_R \mathcal{L}_R + \lambda_S \mathcal{L}_S, \quad (12)$$

where λ_C , λ_R , and λ_S are predefined coefficients for balancing these loss terms. By the loss term \mathcal{L}_C , we can constrain the chained mapping $C_{1,2}(\mathcal{P}', \mathcal{P})$ to be an identity matrix. In addition, as we explicitly require the last shape to be some random rigid transform \mathcal{P}' of the source \mathcal{P} , \mathcal{L}_R enforces the learned representation to be robust to mild rotations. Moreover, \mathcal{L}_S encourages the correspondence to be as bijective as possible, benefiting the pointwise representation learning.

3.4 Network architecture

Our network takes 1024 points as input from a point cloud shape sampled by the farthest sampling technique, and outputs pointwise representations of dimension 64. PointNet++ [33] is adopted as the backbone for unsupervised learning, and trained with Eq. 12. Other backbones (e.g. DGCNN [47]) may be applicable but we limit our discussion here to PointNet++.

In addition, we incorporate the multi-head attention module (see [43, 25]) in between adjacent layers of the original PointNet++. The motivation of such design is to combine both local information gathered by convolution-like operations and non-local information by self-attention modules to benefit the unsupervised learning process. For details of the network architecture, please refer to our supplementary materials.

4 Experimental Results

4.1 Implementation and results of the unsupervised pretraining

We pre-train the proposed networks on the pretext of finding dense correspondences on *ShapeNet part segmentation* dataset [52]. *ShapeNet part segmentation* dataset contains 16 categories of shapes, each of which is represented by a point set and associated with normals and the part labels. The number of parts per shape in a category varies from 2 to 6, totaling 50 parts in the dataset. All the point clouds are pre-aligned.

Training data are augmented with random rotations, translations and scalings sampled from a uniform distribution. The rotations and translations along

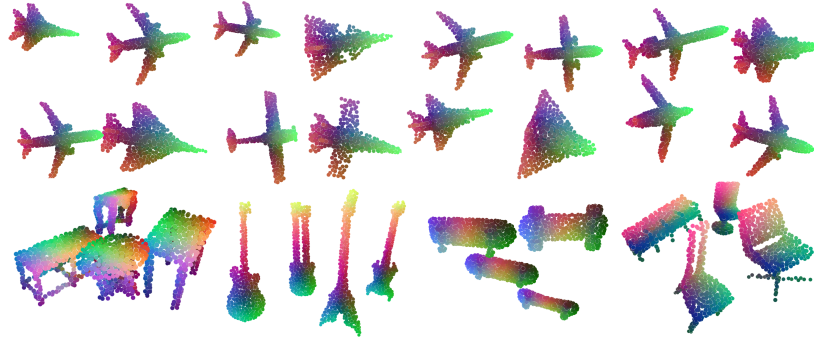


Fig. 2. Visualization of the learned pointwise representations on rotated shapes from different categories, i.e., *Airplane*, *Table*, *Guitar*, *Skateboard*, and *Chair*. Color-codes reflect the consistency of the learned pointwise representations across a variety of shape instances, even if the shapes are under perturbation of rigid transformations

each coordinate axis are sampled from the range $[-15deg, +15deg]$ and $[-0.2, +0.2]$, respectively; and the uniform scaling factor ranges from $[0.8, 1.25]$. These random transformations are applied to each of the training triplet, i.e., source shape, its rigid transform, and the target. We first sample a source shape and a target one from the same category, and then apply two sets of random transformations described above to the source and one set of random transformations to the target, respectively. Thus, three transformed shapes are generated, forming a triplet, i.e. the source, its rigid transform, and the target, for training.

We use a variant [34] of the Adam optimizer [23] with $\beta_1 = 0.900$ and $\beta_2 = 0.999$ across all experiments presented in the paper. The learning rates for bias and the rest of parameters are set to 0.0001 and 0.0005, respectively, without decay during the training process. Balancing coefficients $\lambda_{C,R,S}$ are set to 1.0, 1.0, 0.06 to weight each loss term for the network pre-training. All network models are trained with an NVIDIA GeForce GTX 1080Ti graphical card.

We randomly sample a set of perturbed shapes with random rigid transformations as described above, and visualize their pointwise features in Fig. 2. The features are dimension-reduced via t-SNE [27] and color-coded to reflect their similarity. Although the shapes are randomly transformed, the visualization shows that our learned representations on various non-aligned shape instances are consistent. More qualitative results can be found in the supplementary.

4.2 Ablation study

We validate our network design and our proposed training strategy in this subsection. For evaluation, we employ the ratio of correct cycle matches ($CC\%$) during training and validation as the metrics, which indicates the success rate of completing cycle-consistent correspondences. Point-cloud shapes from three

categories, i.e. Airplane, Chair, and Table, from the *ShapeNet part segmentation* dataset are adopted for evaluation.

Different settings are compared to justify our proposed framework with the designed loss function (Eq. 12) and the self-attention module. We first evaluate two variants of the loss functions: 1) removing the bijective constraint enabled by the Sinkhorn regularization (*w/o* L_S); and 2) enforcing the correspondence matrix to be the identity matrix instead of permutation matrices (i.e. replacing Eq. 11 by $L_I = \|C_{1,2} - \mathbf{I}\|_1$, denoted L_I). The comparison results are depicted in Tab. 1. As can be seen from the second row block of Tab. 1, our complete loss function (*Ours*) produces the best result among the three settings. The performances of the other two settings, i.e. *w/o* L_S and L_I , are similar.

The performance gain by Sinkhorn regularization is mainly due to the penalty it imposes on the many-to-one correspondences, which smoothly increases as the training proceeds and thus drives the resulting mapping as much close to a bijective map as possible. On the contrary, the setting without such a constraint (*w/o* L_S) would indulge many-to-one mappings; and the setting (L_I) that enforces the mapping to identity would be too difficult for training at the very beginning stage, thus impeding the convergence.

As shown in Fig. 3, we compare the $CC\%$ of testing using the models obtained at different training iterations, in which the higher results are better. As observed, after training for more than 2000 iterations, the results w/ L_S (in solid curves) perform significantly better than the ones w/o L_S (in dashed curves), which shows the advantage of our proposed Sinkhorn regularization.

In addition, we compare our network structure with self-attention modules against the vanilla PointNet++. As observed from Tab. 1, our results are higher than those produced by the vanilla PointNet++. This is primarily because the self-attention modules will attend to long-range dependency that convolution-like operations overlook at the entrancing levels. Note that, although using the network structure equipped with self-attention modules, the two settings (L_I and *w/o* L_S) are generally inferior to the vanilla PointNet++ (*w/o self-attention*) trained with the Sinkhorn regularization, revealing that it is the primary contributor to the performance gain. Besides, we also evaluate the input of our model. As shown in Tab. 1, the performance will be degraded without normals as inputs. But such a decrease in performance is relatively small, comparing to settings of L_I and *w/o* L_S that use normals.

4.3 Applications to shape analysis tasks

As the pointwise features learned by the proposed unsupervised framework method are independent of the subsequent tasks, we demonstrate their applicability to the following point cloud analysis tasks: partial-to-partial shape registration, keypoint transfer via correspondence, and supervised part segmentation.

Partial-to-partial shape registration. To perform partial shape registration between a shape and its rigid transform, we leverage the obtained pointwise

Table 1. Ablation study on different loss terms, network structures, and the input of our model. $CC\%$ denotes the percentage of the correct cycle matches. We compare the metrics on three categories of data: *Airplane*, *Chair* and *Table*

Category	Airplane		Chair		Table		Mean	
$CC\%$	Train	Val.	Train	Val.	Train	Val.	Train	Val.
Ours	69.4	67.9	68.7	69.4	67.1	65.1	68.4	67.5
w/o L_S	44.8	44.9	40.4	41.0	40.5	39.1	41.9	41.5
L_I (replacing L_S)	40.6	42.0	61.4	62.0	43.4	42.2	48.8	48.7
w/o self-attention	51.5	49.9	51.0	51.4	50.6	48.8	51.0	50.0
w/o normals	57.6	56.6	63.6	64.9	48.3	47.3	56.5	56.3

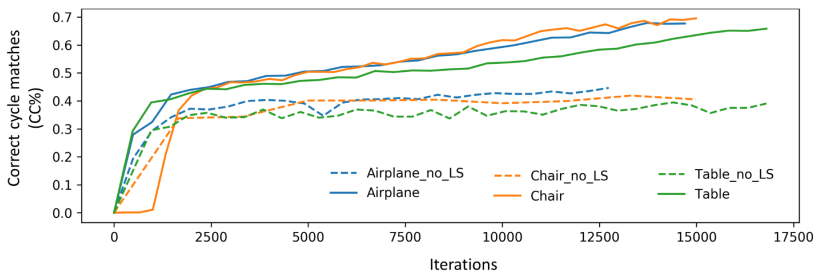


Fig. 3. Performances with the Sinkhorn regularization L_S (solid) or without it (dashed) are compared in terms of the percentages of correct cycle matches ($CC\%$) on three data categories, showing that the Sinkhorn regularization can facilitate the optimization and achieves consistently better results

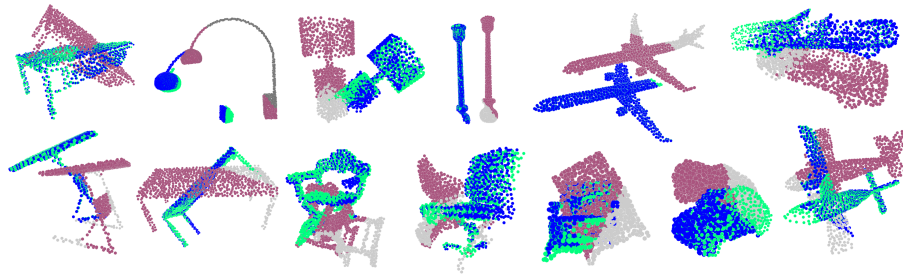
correspondence between these two shapes, and compute a rigid transformation [3] to perform shape registration. To this end, we first pre-train our network on the pretext as described above. We then fine-tune it with more emphasis on the rigid transformation term L_R by setting $\lambda_{C,R,S} = 0.0001, 1.0, 0.06$.

We compare, in a category-specific manner, our results to those produced by PRNet [46] on five categories of shapes from *ModelNet40* [49]. We follow the training settings in [46] by using 1024 points to train our network and PRNet with respect to each category data. It is worthy noting that different from their training strategy where a portion of points are subtracted from input data to mimic partial point clouds, we do not apply this specialized data augmentation to our training data. During test time, a shape with 1024 points is truncated to 768 points to produce a partial point cloud. Both our method and PRNet perform 3 iterative estimations for pose registration. Same random transformations are applied to generating a consistent testing set, ensuring a fair, category-specific comparison between our method and PRNet.

We evaluate the comparison results using the metrics Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). As shown in Tab. 2, our network

Table 2. Category-specific comparison with PRNet [46] for partial-to-partial registration on unseen point clouds. Bold values refer to the better performance.

Category		Aeroplane		Car		Chair		Table		Lamp	
Metric		[46]	Ours	[46]	Ours	[46]	Ours	[46]	Ours	[46]	Ours
Rotation	RMSE	7.206	4.287	15.42	4.678	4.93	6.202	39.6	3.13	37.1	21.85
	MAE	3.78	3.532	7.58	3.876	3.09	5.279	23.7	2.71	23.1	18.22
Translation	RMSE	0.047	0.018	0.127	0.018	0.027	0.015	0.175	0.017	0.174	0.0476
	MAE	0.030	0.016	0.096	0.015	0.019	0.013	0.124	0.015	0.125	0.0418

**Fig. 4.** Visualization of partial-to-partial shape registration results. Purple and green point clouds are the source and target poses of a shape; the blue are the obtained results via three iterative match-and-transform; and the grey parts are those randomly truncated during the testing stage

can achieve results better than or at least comparable to PRNet (trained in a category-specific manner) across the listed categories. An exception is the *Chair* category in terms of the rotation metrics. Some registration results are randomly selected and visualized in Fig. 4 where shapes in purple, green and blue represent the source pose, target pose and our result, respectively.

Keypoint transfer via correspondences. Keypoints are a group of sparsely defined landmarks on a shape, crucial for many shape analysis applications. In this part, we demonstrate the learned pointwise representations can be leveraged to transfer a set of keypoints defined on a shape to other shapes alike. We compare our results to several state-of-the-art unsupervised learning methods for 3D shapes, i.e. [9, 51] based on autoencoder structures and [4, 30] based on pointwise feature learning (similar to ours). All methods are trained on the *Airplane*, *Chair*, and *Bike* data from the *ShapeNet part segmentation* dataset in a category-specific, unsupervised manner. Shapes are pre-aligned for training and testing to conduct fair comparison. Evaluation is made on the test set in [18] where ground-truth keypoints of around 100 shapes per category are provided and each shape may contains 6 to 12 keypoints.

Table 3. Results of keypoint transfer comparing with the state-of-the-art methods. The results are measured by the percentage of the keypoints with the distance error less than 0.05. Bold values refer to the top performance

	LMVCNN[18]	AtlasNet[9]	FoldingNet[51]	EdgeNet[4]	ShapeUnicode[30]	Ours
Airplane	30.3	51.1	26.6	33.5	30.8	57.9
Chair	12.0	37.3	16.2	12.6	25.4	40.4
Bike	17.4	34.2	31.7	27.2	58.3	49.8
Mean	19.9	40.9	24.9	24.4	38.2	49.4

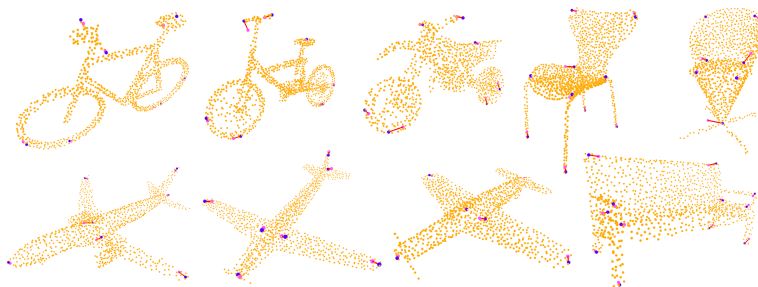


Fig. 5. Visualization of the key point transfer task. Blue points are the ground-truth landmarks while points in magenta are estimated by our network

Given a point cloud shape, \mathcal{P} , with several 3D keypoints, we find their corresponding points in another given shape \mathcal{Q} . As the ground-truth keypoints are not given as a particular point in the point cloud data, we first sample 5 neighboring points from the point cloud and then search for each of the neighbors its correspondence point in \mathcal{Q} via the learned pointwise features. Finally, we simply average the correspondence points to predict the corresponding keypoints in \mathcal{Q} .

We measure the distance error between the ground-truth keypoints and the predicted ones with the Euclidean metric. Noticing that the distance error greater than 0.05 is a relatively large value for a shape whose size is normalized with respect to its shape diameter, we show the percentage of keypoints with the distance error smaller than this threshold in Tab. 3. As shown in the table, our result generally outperforms the other existing methods, except a slight fall behind ShapeUnicode [30] on the category of *Bike*, showing the effectiveness of the learned pointwise representations in correspondence query. Some qualitative results are shown in Fig. 5.

Supervised part segmentation. To demonstrate the ability of the learned pointwise representations as a feature to boost subsequent applications, we also use them as additional features to train a supervised part segmentation network. In particular, we adopt the PointNet++[33] as a baseline for this supervised task.

Table 4. Comparison of the segmentation results on *ShapeNet part segmentation* dataset trained with full and 10% of the dataset

Full train	aero.	bag	cap	car	chair	ear.	guitar	knife	lamp	laptop	motor	mug	pistol	rocket	skate.	table	mean
[32]	83.40	78.70	82.50	74.90	89.60	73.00	91.50	85.90	80.80	95.30	65.20	93.00	81.20	57.90	72.80	80.60	83.7
[33]	82.30	79.00	87.70	77.30	90.80	71.80	91.00	85.90	83.70	95.30	71.60	94.10	81.30	58.70	76.40	82.60	85.1
Ours	82.66	81.97	79.96	78.03	85.77	70.12	91.61	86.53	81.81	96.03	73.55	95.57	83.49	59.10	75.39	88.23	85.5
10% train	aero	bag	cap	car	chair	ear.	guitar	knife	lamp	laptop	motor	mug	pistol	rocket	skate.	table	mean
[32]	76.10	69.80	62.60	61.40	86.00	62.10	86.20	79.70	73.60	93.30	59.10	83.40	75.90	41.80	57.70	74.80	77.3
[33]	76.40	43.40	77.80	75.52	87.50	67.70	87.40	77.40	71.40	94.10	61.30	90.40	72.80	51.40	68.70	75.30	78.6
Ours	77.09	73.24	81.80	74.39	84.71	70.23	88.37	84.23	76.63	94.12	62.98	91.29	80.60	51.25	65.02	77.94	79.8

During this supervised training for part segmentation, the proposed network pre-trained on the dense correspondence pretext is frozen and serves as a feature extractor. We compare the part segmentation results obtained with our additional input features and with the original inputs containing point coordinates and normals only. As can be seen from Tab. 4, the results with our pointwise features are improved in most categories. The part-averaged mean IoU (Intersection-over-Union) reaches 85.5%, higher than the performance obtained by the PointNet++.

On the other hand, when we train the network with 10% of the labeled training data (without fine-tuning the pre-trained network), the performance gains are observed in 12 out of 16 categories, many of which outperforms the original results by a large margin (i.e. up to 7%). In this setting, we achieve a part-averaged mean IoU of 79.8%.

5 Conclusions and Future Work

This paper proposes a pretext of finding dense correspondences between two different shapes for unsupervised learning of pointwise features for point cloud shapes and formulates a cycle-consistency based framework to solve this problem. In order to learn discriminative pointwise features, we force the cycle correspondences to be as bijective as possible using the Sinkhorn regularization. Ablation study validates the design and effectiveness of the proposed unsupervised framework. Furthermore, we demonstrate the applicability of acquired pointwise features in downstream tasks: partial-to-partial shape registration, unsupervised keypoint transfer, and supervised part segmentation.

While geometric correspondences can be effectively learned by the proposed approach, such correspondences learned in this unsupervised manner may fail to capture the semantic meaning of the shapes. As future work, we would like to explore solutions to this problem based on the proposed unsupervised framework.

Acknowledgement

We acknowledge valuable comments from anonymous reviewers. Our work is supported in part by Hong Kong Innovation and Technology Fund ITS/457/17FP.

References

1. Adams, R.P., Zemel, R.S.: Ranking via Sinkhorn propagation. arXiv preprint arXiv:1106.1925 (2011)
2. Aoki, Y., Goforth, H., Srivatsan, R.A., Lucey, S.: PointNetLK: Robust & efficient point cloud registration using pointnet. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7163–7172 (2019)
3. Arun, K.S., Huang, T.S., Blostein, S.D.: Least-squares fitting of two 3-D point sets. IEEE Transactions on Pattern Analysis and Machine Intelligence **9**(5), 698–700 (1987)
4. Chen, M., Zou, Q., Wang, C., Liu, L.: EdgeNet: Deep metric learning for 3d shapes. Computer-Aided Geometric Design **72**, 19–33 (2019)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)
6. Choy, C., Park, J., Koltun, V.: Fully convolutional geometric features. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8958–8966 (2019)
7. Deng, H., Birdal, T., Ilic, S.: Ppf-foldnet: Unsupervised learning of rotation invariant 3D local descriptors. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 602–618 (2018)
8. Deng, H., Birdal, T., Ilic, S.: Ppfnet: Global context aware local features for robust 3d point matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 195–205 (2018)
9. Deprelle, T., Groueix, T., Fisher, M., Kim, V., Russell, B., Aubry, M.: Learning elementary structures for 3D shape generation and matching. In: Advances in Neural Information Processing Systems. pp. 7433–7443 (2019)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
11. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: Temporal cycle-consistency learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1801–1810 (2019)
12. Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: A papier-mâché approach to learning 3D surface generation pp. 216–224 (2018)
13. Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: Unsupervised cycle-consistent deformation for shape matching. In: Computer Graphics Forum. vol. 38, pp. 123–133. Wiley Online Library (2019)
14. Halimi, O., Litany, O., Rodola, E., Bronstein, A.M., Kimmel, R.: Unsupervised learning of dense shape correspondence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4370–4379 (2019)
15. Han, Z., Wang, X., Liu, Y.S., Zwicker, M.: Multi-angle point cloud-VAE: unsupervised feature learning for 3d point clouds from multiple angles by joint self-reconstruction and half-to-half prediction. arXiv preprint arXiv:1907.12704 (2019)
16. Hassani, K., Haley, M.: Unsupervised multi-task feature learning on point clouds. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8160–8171 (2019)
17. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. arXiv preprint arXiv:1911.05722 (2019)
18. Huang, H., Kalogerakis, E., Chaudhuri, S., Ceylan, D., Kim, V.G., Yumer, E.: Learning local shape descriptors from part correspondences with multiview convolutional networks. ACM Transactions on Graphics (TOG) **37**(1), 1–14 (2017)

19. Huang, Q.X., Guibas, L.: Consistent shape maps via semidefinite programming. In: *Computer Graphics Forum*. vol. 32, pp. 177–186. Wiley Online Library (2013)
20. Huang, Q.X., Su, H., Guibas, L.: Fine-grained semi-supervised labeling of large shape collections. *ACM Transactions on Graphics (TOG)* **32**(6), 1–10 (2013)
21. Kim, V.G., Li, W., Mitra, N.J., Chaudhuri, S., DiVerdi, S., Funkhouser, T.: Learning part-based templates from large collections of 3d shapes. *ACM Transactions on Graphics (TOG)* **32**(4), 1–12 (2013)
22. Kim, V.G., Li, W., Mitra, N.J., DiVerdi, S., Funkhouser, T.: Exploring collections of 3D models using fuzzy correspondences. *ACM Transactions on Graphics (TOG)* **31**(4), 1–11 (2012)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
24. Knight, P.A.: The Sinkhorn–Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications* **30**(1), 261–275 (2008)
25. Lee, J., Lee, Y., Kim, J., Kosiolek, A.R., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. *arXiv preprint arXiv:1810.00825* (2018)
26. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B.: PointCNN: Convolution on x-transformed points. In: *Advances in Neural Information Processing Systems*. pp. 820–830 (2018)
27. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. *Journal of Machine Learning Research* **9**(Nov), 2579–2605 (2008)
28. Mena, G., Belanger, D., Linderman, S., Snoek, J.: Learning latent permutations with Gumbel-Sinkhorn networks. *arXiv preprint arXiv:1802.08665* (2018)
29. Misra, I., van der Maaten, L.: Self-supervised learning of pretext-invariant representations. *arXiv preprint arXiv:1912.01991* (2019)
30. Muralikrishnan, S., Kim, V.G., Fisher, M., Chaudhuri, S.: Shape Unicode: A unified shape representation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3790–3799 (2019)
31. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018)
32. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep learning on point sets for 3D classification and segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 652–660 (2017)
33. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: *Advances in Neural Information Processing Systems*. pp. 5099–5108 (2017)
34. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of Adam and beyond. *arXiv preprint arXiv:1904.09237* (2019)
35. Sahillioglu, Y.: Recent advances in shape correspondence. *The Visual Computer* pp. 1–17 (2019)
36. Sauder, J., Sievers, B.: Self-supervised deep learning on point clouds by reconstructing space. In: *Advances in Neural Information Processing Systems*. pp. 12942–12952 (2019)
37. Sinkhorn, R.: A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics* **35**(2), 876–879 (1964)
38. Thewlis, J., Albanie, S., Bilen, H., Vedaldi, A.: Unsupervised learning of landmarks by descriptor vector exchange. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 6361–6371 (2019)

39. Thewlis, J., Bilen, H., Vedaldi, A.: Unsupervised learning of object landmarks by factorized spatial embeddings. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 5916–5925 (2017)
40. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. *arXiv preprint arXiv:1906.05849* (2019)
41. Tschannen, M., Djolonga, J., Rubenstein, P.K., Gelly, S., Lucic, M.: On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625* (2019)
42. Van Kaick, O., Zhang, H., Hamarneh, G., Cohen-Or, D.: A survey on shape correspondence. In: *Computer Graphics Forum*. vol. 30, pp. 1681–1707. Wiley Online Library (2011)
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: *Advances in Neural Information Processing Systems*. pp. 5998–6008 (2017)
44. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2566–2576 (2019)
45. Wang, Y., Solomon, J.M.: Deep closest point: Learning representations for point cloud registration. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 3523–3532 (2019)
46. Wang, Y., Solomon, J.M.: Prnet: Self-supervised learning for partial-to-partial registration. In: *Advances in Neural Information Processing Systems*. pp. 8814–8826 (2019)
47. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)* **38**(5), 1–12 (2019)
48. Wu, W., Qi, Z., Fuxin, L.: Pointconv: Deep convolutional networks on 3d point clouds. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 9621–9630 (2019)
49. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D ShapeNets: A deep representation for volumetric shapes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1912–1920 (2015)
50. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3733–3742 (2018)
51. Yang, Y., Feng, C., Shen, Y., Tian, D.: FoldingNet: Point cloud auto-encoder via deep grid deformation pp. 206–215 (2018)
52. Yi, L., Kim, V.G., Ceylan, D., Shen, I.C., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., Guibas, L.: A scalable active framework for region annotation in 3D shape collections. *ACM Transactions on Graphics (TOG)* **35**(6), 1–12 (2016)
53. Zeng, A., Song, S., Nießner, M., Fisher, M., Xiao, J., Funkhouser, T.: 3DMatch: Learning local geometric descriptors from rgb-d reconstructions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1802–1811 (2017)
54. Zhao, Y., Birdal, T., Deng, H., Tombari, F.: 3D point capsule networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1009–1018 (2019)
55. Zhou, T., Krahenbuhl, P., Aubry, M., Huang, Q., Efros, A.A.: Learning dense correspondence via 3D-guided cycle consistency. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 117–126 (2016)

56. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (2017)