Learning SO(3)-Invariant Semantic Correspondence via Local Shape Transform

Chunghyun Park[1*] Seungwook Kim[1*] Jaesik Park[2] Minsu Cho[1]

[1]POSTECH [2]Seoul National University

http://cvlab.postech.ac.kr/research/RIST

Abstract

*Establishing accurate 3D correspondences between shapes stands as a pivotal challenge with profound implications for computer vision and robotics. However, existing self-supervised methods for this problem assume perfect input shape alignment, restricting their real-world applicability. In this work, we introduce a novel self-supervised **R**otation-**I**nvariant 3D correspondence learner with local **S**hape **T**ransform, dubbed RIST, that learns to establish dense correspondences between shapes even under challenging intra-class variations and arbitrary orientations. Specifically, RIST learns to dynamically formulate an SO(3)-invariant local shape transform for each point, which maps the SO(3)-equivariant global shape descriptor of the input shape to a local shape descriptor. These local shape descriptors are provided as inputs to our decoder to facilitate point cloud self- and cross-reconstruction. Our proposed self-supervised training pipeline encourages semantically corresponding points from different shapes to be mapped to similar local shape descriptors, enabling RIST to establish dense point-wise correspondences. RIST demonstrates state-of-the-art performances on 3D part label transfer and semantic keypoint transfer given arbitrarily rotated point cloud pairs of the same category, outperforming existing methods by significant margins.*

1. Introduction

Establishing dense 3D correspondences between different shapes is foundational to numerous applications across computer vision, graphics, and robotics [9, 22, 28, 41]. One of the primary challenges hindering advancements in this domain is the difficulty of annotating dense inter-shape correspondences, which limits the leverage of strongly-supervised learning paradigms.

Recently, self-supervised learning methods have been proposed to address this issue [3, 21], showing promising directions for 3D correspondence estimation. Nonetheless, a significant limitation in existing approaches is their stringent assumption about the alignment of input shape pairs; these methods strongly assume that the input point cloud pair to establish correspondences between is precisely aligned. This assumption is rarely met in practice, where object scans and shape instances can be arbitrarily oriented. We find that the performance of existing methods degrades significantly when confronted with rotated input shapes, restricting their real-world applicability (Figure 1).

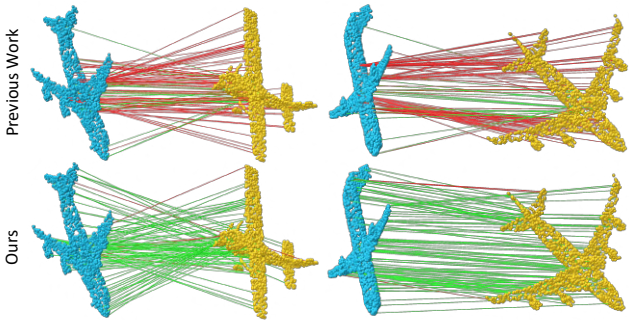

Figure 1. **Semantic correspondence between rotated shapes.** We visualize the semantic correspondence results of the previous SOTA [3] and ours, given two randomly rotated airplanes from the ShapeNetPart dataset [39]. Green and red lines indicate the correct and incorrect matches, respectively. For each method, 100 source points are randomly selected from the source (yellow) for correspondence visualization[1]. Ours predicts SO(3)-invariant correspondences, showing superior accuracy and robustness in comparison to the previous SOTA [3].

To address this challenge, we introduce a novel self-supervised learning approach, dubbed RIST, designed to reliably determine dense *SO(3)-invariant* correspondences between shapes via local shape transform. In essence, RIST learns to formulate SO(3)-invariant local shape transform for each point in a dynamic and input-dependent manner. Each point-wise local shape transform maps the SO(3)-equivariant *global* shape descriptor of the input shape to a *local* shape descriptor, which is passed to the decoder to re-

*Equal contribution

[1]For the details of the inference algorithm, please refer to Appendix A.

construct the shape and pose of the input shape. By training RIST via self- and cross-reconstruction of input shapes, true semantically corresponding points are trained to yield similar local shape descriptors, enabling us to determine dense shape correspondences.

RIST demonstrates state-of-the-art performance on part segmentation label transfer when evaluated on the ShapeNetPart [39] and ScanObjectNN [35] datasets. In particular, significant improvements over existing baselines are observed when our method is applied to randomly oriented shape pair inputs. Furthermore, our approach also proves to be more effective compared to existing methods at semantic keypoint transfer when evaluated on the KeypointNet dataset [40]. This showcases not only the applicability of RIST across a diverse range of tasks, but also its potential to be utilized for efficient dense annotation of 3D shapes. These results highlight the efficacy of RIST in addressing the challenges posed by real-world scenarios where existing methods fail to perform effectively.

The main contributions of our work are as follows:

- We introduce RIST, a novel self-supervised approach for determining dense SO(3)-invariant correspondences between arbitrarily aligned 3D objects.
- We propose to formulate the local shape information of each point as a novel function called *local shape transform* with dynamic input-dependent parameters, which effectively maps the global shape descriptor of input shapes to local shape descriptors.
- RIST achieves state-of-the-art performance on 3D part segmentation label transfer and 3D keypoint transfer under arbitrary rotations, indicating its potential for application in a wide range of practical tasks in computer vision.

2. Related Work

Point cloud understanding via self-supervised learning. While traditional methods for point cloud processing involving hand-crafted features [27, 33] have shown impressive performance, with the advent of deep learning, substantial research efforts have been directed towards developing learning-based algorithms capable of effectively processing and understanding point clouds [5, 24–26, 42]. Due to limited large-scale datasets with rich annotations, self-supervised learning approaches emerged as a viable alternative. One of the most prominent directions to learn point cloud representations in a self-supervised manner is learning through self-reconstruction [23, 38, 43] of the point cloud. Primarily inspired by the efficacy of point cloud reconstruction as a self-supervised representation learning scheme, we train RIST to establish 3D correspondences in a self-supervised manner via self- and cross-reconstruction of point clouds by leveraging SO(3)-invariant dynamic local shape transform.

Equivariance and invariance to rotation. The conventional method to improve a neural network's robustness to rotation is by employing rotation augmentations during training or inference. However, this tends to increase the resources required for training and still shows unsatisfactory results when confronted with an unseen rotation [14, 18]. In recent years, various methods have been proposed to yield point cloud representations, which are equivariant [2, 6, 29, 32] or invariant [14, 18, 19, 31, 37] to the rotation of the input, demonstrating enhanced performances under arbitrary input rotations. To facilitate the rotation-robust establishment of 3D dense correspondences, we utilize SO(3)-equivariant networks in building RIST, leveraging SO(3)-equivariant and -invariant representations to guarantee robustness to rotation by design.

Semantic correspondences under intra-class variations. Finding correspondences between images or shapes under intra-class variations - manifesting as differences in shape, size, and orientation within the same category of objects - poses significant challenges over photometric or viewpoint variations. This task has been widely studied in the domain of images, where existing methods make use of sparsely annotated image pair datasets to train their method in a strongly- or a weakly-supervised manner [4, 11, 13, 15, 34]. However, learning to establish dense yet reliable 3D correspondences between 3D shapes remains challenging, as it is infeasible to label dense correspondence annotations across point cloud pairs with intra-class variations. Self-supervised methods have been proposed to address this issue [3, 21], but they strongly assume that the input point clouds are aligned, leading to considerable significant degradation when confronted with arbitrarily rotated point clouds. Additionally, the functional map-based approach introduced by Huang *et al*. [10] for non-rigid registration struggles with topological changes and efficiency. To this end, we propose RIST to establish reliable 3D dense correspondences irrespective of the input point clouds' poses.

3. RIST for 3D Semantic Correspondence

In this section, we detail the components of RIST, which come together to facilitate the end-to-end self-supervised training for 3D semantic correspondence establishment. The objective of 3D semantic correspondence is as follows; given two different point clouds instances $\mathbf{P}_1 \in \mathbb{R}^{N\times 3}$ and $\mathbf{P}_2 \in \mathbb{R}^{N\times 3}$ belonging to the same semantic category, we aim to find all semantically corresponding point pairs $\{\mathbf{p}_i, \mathbf{q}_i\}_{i=1}^{N'}$ [2] such that $\mathbf{p}_i \in \mathbf{P}_1$ and $\mathbf{q}_i \in \mathbf{P}_2$. To achieve this, we claim it is crucial to identify the local shape information *i.e*. local semantics and geometry, which is generalizable across different instances within the same category.

[2] $N' \leq N$; there could be points with no matches.

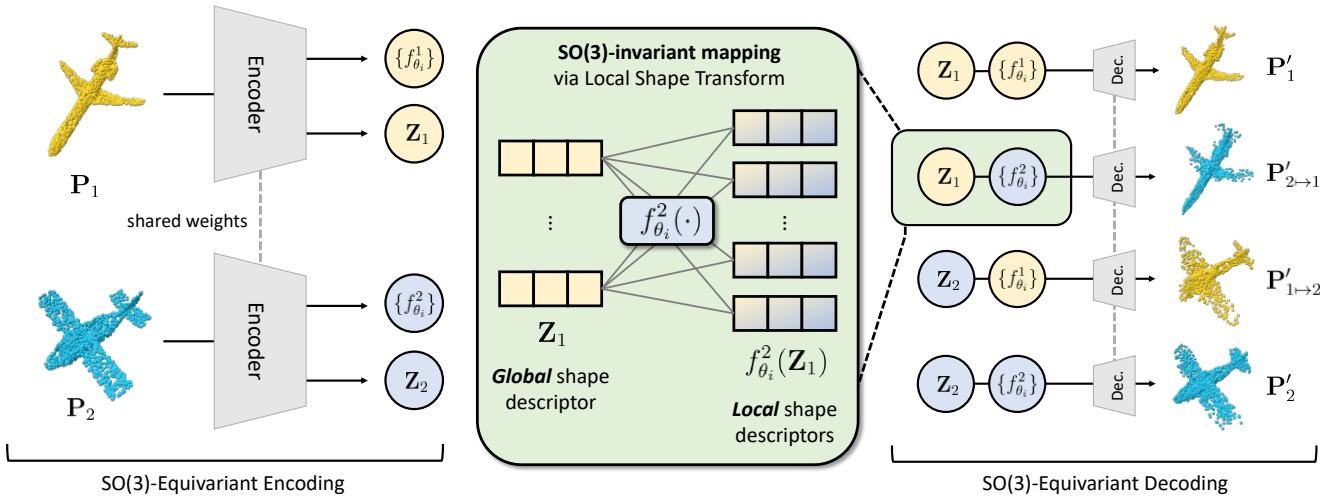

Figure 2. **Overview: Self-supervised training of RIST**. The input point clouds are independently encoded to SO(3)-equivariant global shape descriptor $\mathbf{Z}$ and dynamic SO(3)-invariant point-wise local shape transforms $\{f_{\theta_i}\}$. The local shape transforms map the global shape descriptor to local shape descriptors by infusing local semantics and geometry, which are used as inputs to the decoder for self-reconstruction. For cross-reconstruction, we apply the local shape transforms formulated from *another* point cloud to reconstruct the point cloud, ensuring that the local shape descriptors successfully capture generalizable local semantics and geometries. We supervise RIST via penalizing errors in self- and cross-reconstructions. At inference, we can leverage the local shape transforms for obtaining local shape descriptors, to identify the dense correspondences.

Therefore, the main idea of RIST is to dynamically generate a SO(3)-invariant local shape transform as a function for each point, such that each local shape transform can map the SO(3)-equivariant global shape descriptor of the input point cloud to its respective local shape descriptor. In the following, we elaborate on the network architecture of RIST, in particular how we leverage SO(3)-equivariant and invariant representations to facilitate the dynamic formulation of pointwise SO(3)-invariant local shape transforms and the reconstruction of pose-preserved point clouds (Sec. 3.1). Subsequently, we introduce our self-supervisory objective function, which trains RIST to self- and cross-reconstruct the input point clouds in a rotation-equivariant manner (Sec. 3.2), finally enabling the establishment of 3D dense correspondences (Sec. 3.3) via corresponding local shape descriptors. Figure 2 illustrates the outline of the training scheme of RIST.

3.1. Network Design of RIST

3.1.1 Preliminary: SO(3)-Equivariant Representation

One of the main motivations of RIST is to establish reliable and accurate 3D dense correspondences given *arbitrarily rotated* shapes, a setting where existing work shows to be brittle. This requires our encoder to formulate the point-wise local shape transforms not only effectively to capture the local shape semantics and geometry, but also robustly against transformations in the SO(3) space. To this end, we integrate SO(3)-equivariant networks into RIST to facilitate robustness to SO(3) transformations of the input. In this work, we choose VNNs [7] to build our SO(3)-equivariant layers for their simplicity and efficacy. In VNNs, a single neuron, which is represented by a scalar-list of values, is lifted to a *vector-list* feature $\mathbf{V} \in \mathbb{R}^{C\times 3}$, which is essentially a sequence of 3D vectors. The layers of VNNs handle batches of such vector-list features such that equivariance with respect to rotation $R \in \mathrm{SO}(3)$ is satisfied *i.e.* $f(\mathbf{V}R) = f(\mathbf{V})R$[3]. Notably, we can yield an SO(3)-*invariant* output by performing a product of an equivariant vector-list feature $\mathbf{V}R \in \mathbb{R}^{C\times 3}$ with the transpose of another consistently equivariant vector-list feature $\mathbf{U}R \in \mathbb{R}^{C'\times 3}$ as follows: $(\mathbf{V}R)(\mathbf{U}R)^\top = \mathbf{V}RR^\top\mathbf{U}^\top = \mathbf{V}\mathbf{U}^\top$. This serves as a critical functionality when constructing our SO(3)-invariant local shape transform of our encoder (Sec. 3.1.2).

3.1.2 SO(3)-Equivariant Encoder

We design our encoder architecture to take as input a point cloud $\mathbf{P} \in \mathbb{R}^{N\times 3}$, and simultaneously output an SO(3)-*equivariant* global shape descriptor and formulate point-wise SO(3)-*invariant* local shape transforms.

SO(3)-equivariant global shape descriptor. Given a point cloud, we first aim to obtain the SO(3)-equivariant *global* shape descriptor $\mathbf{Z} \in \mathbb{R}^{C\times 3}$, which captures the pose and the global shape characteristics of the input point cloud. We leverage VN-DGCNN [7] as our encoder architecture,

[3]We refer the readers to the original paper [7] for further information and detailed formulations of VNNs.

which consists of 4 edge convolutional VN-layers to capture local semantics at a progressively larger receptive field, and a FPN [20] to aggregate the multi-level features. Then, we apply the global average pooling to the aggregated SO(3)-equivariant point-wise features $\mathbf{V}^{\text{equi}} \in \mathbb{R}^{C\times 3\times N}$ to encode SO(3)-equivariant global shape descriptor $\mathbf{Z}$ of the input point cloud. The global shape descriptor can be used subsequently as the input for our SO(3)-invariant point-wise local shape transform, to be mapped to their respective local shape descriptors, as shown in Figure 2.

SO(3)-invariant local shape transform. Alongside the extraction of SO(3)-equivariant global shape descriptors, we also formulate the SO(3)-*invariant* local shape transform $f_{\theta_i} : \mathbb{R}^{C\times 3} \mapsto \mathbb{R}^{C'\times 3}$ for each point $\mathbf{p}_i \in \mathbb{R}^3$ of the input point cloud $\mathbf{P} \in \mathbb{R}^{N\times 3}$. The parameters of each local shape transform $\theta_i \in \mathbb{R}^{C'\times C}$ are input-dependent - thus, dynamic since they are predicted by our encoder for the i-th point of the point cloud. To predict θ_i, we first obtain SO(3)-*invariant* point-wise features $\mathbf{V}^{\text{in}} \in \mathbb{R}^{C'\times 3\times N}$ as described in Sec. 3.1.1. Then, we transform each vectorized SO(3)-invariant point-wise feature $\text{vec}(\mathbf{v}_i^{\text{in}}) \in \mathbb{R}^{3C'}$ to the vectorized parameter of the local shape transform $\text{vec}(\theta_i) \in \mathbb{R}^{C'C}$ by using a multi-layer perceptron. By reshaping $\text{vec}(\theta_i)$ to $\theta_i \in \mathbb{R}^{C'\times C}$, we finally obtain the dynamic and SO(3)-invariant local shape transform f_{θ_i} for the point $\mathbf{p}_i$.

The role of these local shape transforms is to map the SO(3)-equivariant global shape descriptor $\mathbf{Z} \in \mathbb{R}^{C\times 3}$ to their respective local shape descriptors $\mathbf{v}'_i := f_{\theta_i}(\mathbf{Z}) \in \mathbb{R}^{C'\times 3}$, which is provided as the input to our decoder for reconstruction. Our self-supervised training scheme encourages the point-wise dynamic local shape transform to encapsulate the local shape information *e.g.* semantics and geometry, to enhance the reconstruction performance.

3.1.3 SO(3)-Equivariant Decoder

Our decoder aims to reconstruct the initial input shapes using the obtained SO(3)-equivariant global shape descriptors $\mathbf{Z}$ and the SO(3)-invariant local shape transforms $\{f_{\theta_i}\}_{i=1}^N$. To reconstruct the point clouds aligned to their initial poses, we leverage SO(3)-equivariant layers as the building blocks of our decoder architecture. We first train our decoder to perform self-reconstruction, using the local shape descriptors $\mathbf{V}'$, *i.e.* $\mathbf{P} \leftrightarrow \mathbf{P}' := \text{Decoder}(\mathbf{V}') = \text{Decoder}(\{f_{\theta_i}(\mathbf{Z})\}_{i=1}^N)$. We also train our decoder to perform cross-reconstruction, where we use the local shape descriptors obtained using global shape descriptors and local shape transforms from *different* point clouds. Specifically, assume we are given two point clouds $\mathbf{P}_1, \mathbf{P}_2 \in \mathbb{R}^{N\times 3}$, with SO(3)-equivariant global shape descriptors $\mathbf{Z}_1, \mathbf{Z}_2 \in \mathbb{R}^{C\times 3}$ and SO(3)-invariant local shape transforms $\{f^1_{\theta_i}\}_{i=1}^N, \{f^2_{\theta_i}\}_{i=1}^N$. As shown in Figure 2, we then can perform cross-reconstruction from $\mathbf{P}_2$ to $\mathbf{P}_1$ as follows: $\mathbf{P}_1 \leftrightarrow \mathbf{P}'_{2\mapsto 1} := \text{Decoder}(\{f^2_{\theta_i}(\mathbf{Z_1})\}_{i=1}^N)$. Intuitively, for the above cross-reconstruction to be carried out successfully, the local shape transforms for points of a *true* correspondence should hold similar dynamic parameters, mapping global shape descriptors to similar local shape descriptors. By training RIST to cross-reconstruct point clouds, we are supervising local shape transforms to map corresponding points between shapes to similar local shape descriptors, which encode local semantics and geometry that are generalizable across different instances within a category.

3.2. Self-Supervised Objective

Due to the lack of annotated datasets for dense 3D inter-shape correspondences, we train RIST in a self-supervised manner by penalizing inaccurate shape reconstructions. First, we supervise RIST for self-reconstruction using the following loss:

$$\mathcal{L}_{\text{SR}} = \lambda_{\text{MSE}}\,\text{MSE}(\mathbf{P}, \mathbf{P}') + \lambda_{\text{EMD}}\,\text{EMD}(\mathbf{P}, \mathbf{P}'), \quad (1)$$

where MSE is the Mean Squared Error, EMD stands for the Earth Mover's Distance, and both λ_{MSE} and λ_{EMD} are weight coefficients. In essence, we are trying to minimize the difference between the input and reconstructed point cloud. We also supervise RIST for cross-reconstruction as follows: $\mathcal{L}_{\text{CR}} = \lambda_{\text{CD}}\,\text{CD}(\mathbf{P}_1, \mathbf{P}'_{2\mapsto 1})$, where CD stands for the Chamfer distance, and λ_{CD} is a weight coefficient. Finally, our total loss $\mathcal{L}_{\text{total}}$ is defined as: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SR}} + \mathcal{L}_{\text{CR}}$. We omit the CD loss from self-reconstruction, as we can directly use the input point cloud to provide supervision using the MSE loss. We also omit the EMD loss from cross-reconstruction, as EMD tends to overlook the fidelity of detailed structures [36], which is crucial in cross-reconstruction of shapes under intra-class variations.

3.3. SO(3)-Invariant Correspondence

In this section, given two randomly rotated point clouds $\mathbf{P}_1$ and $\mathbf{P}_2$, we elaborate on how our RIST establishes the 3D dense correspondence from $\mathbf{P}_1$ to $\mathbf{P}_2$. As shown in Figure 2, we first encode the SO(3)-equivariant global shape descriptor of $\mathbf{P}_1$, $\mathbf{Z}_1 \in \mathbb{R}^{C\times 3}$, and the SO(3)-invariant local shape transform functions of $\mathbf{P}_2$, $\{f^2_{\theta_i}\}_{i=1}^N$. Then, we cross-reconstruct $\mathbf{P}_1$ as follows: $\mathbf{P}'_{2\mapsto 1} := \text{Decoder}(\{f^2_{\theta_i}(\mathbf{Z_1})\}_{i=1}^N)$. Finally, we define the 3D dense correspondence from $\mathbf{P}_2$ to $\mathbf{P}_1$ as the nearest point pairs among all possible pairs between $\mathbf{P}_1$ and $\mathbf{P}'_{2\mapsto 1}$. Since both encoder and decoder are SO(3)-equivariant, the cross-reconstructed point cloud $\mathbf{P}'_{2\mapsto 1}$ is aligned to $\mathbf{P}_1$. As a result, our RIST can predict 3D dense correspondences between randomly rotated point clouds, while previous approaches [3, 21] experience a high rate of failure.

Training	Method	Airplane	Cap	Chair	Guitar	Laptop	Motorcycle	Mug	Table	Average
w/o Rotations	FoldingNet [38]	17.8	34.7	22.5	22.1	36.2	12.6	50.0	34.6	28.8
	AtlasNetV2 [8]	19.7	31.4	23.6	22.7	36.0	13.1	49.7	35.2	28.9
	DPC [17]	22.7	37.1	25.6	31.9	35.0	17.5	51.3	36.8	32.2
	CPAE [3]	21.0	38.0	26.0	22.7	34.9	14.7	51.4	35.5	30.5
	RIST (ours)	**52.1**	**54.5**	**58.3**	**74.1**	**56.5**	**48.6**	**75.0**	**41.3**	**57.6**
w/ Rotations	FoldingNet [38]	22.5	33.2	24.0	31.0	35.9	13.5	49.9	37.0	30.9
	AtlasNetV2 [8]	21.1	32.7	25.2	28.8	35.5	14.5	49.9	41.0	31.1
	DPC [17]	24.6	38.5	25.6	40.2	34.9	19.3	51.8	37.3	34.0
	CPAE [3]	17.0	36.6	24.5	39.4	37.4	15.8	51.9	36.7	32.4
	RIST (ours)	**51.2**	**57.0**	**55.0**	**73.5**	**60.6**	**48.5**	**72.2**	**44.4**	**57.8**

Table 1. **Average IoU (%) of part label transfer for eight categories in the ShapeNetPart dataset [39].** Ours consistently outperforms previous approaches [3, 8, 17, 38] both with and without rotation augmentations during training, achieving the state-of-the-art IoU. We also provide results of the other classes in Appendix I.

4. Experiments

We present evaluations of RIST on the tasks of 3D part segmentation label transfer and 3D semantic keypoint transfer, following prior work [3, 21]. For both tasks, each method is trained with and without rotation augmentations, and tested on arbitrarily rotated inputs to validate its rotation robustness in predicting semantic correspondences between the rotated inputs. Note that previous approaches [3, 8, 21, 38] used aligned point clouds or slightly rotated point clouds with a subset of SO(3), not the full SO(3), for testing - which is unrealistic in practice, as described in Appendix H.

Datasets. We use the ShapeNetPart [39] and ScanObjectNN [35] datasets to evaluate RIST on the task of 3D part segmentation label transfer, which requires 3D dense semantic correspondence. The ShapeNetPart dataset [39] consists of 16,880 synthetic 3D data from 16 categories. The ScanObjectNN dataset [35] contains *real-world* scanned data, and provides part label annotations for the chair category. Following the previous work [3], we use the same pre-processed KeypointNet [40] dataset for the 3D semantic keypoint transfer task. Since both KeypointNet and ShapeNetPart are based on the ShapeNet dataset [1], we use the eight overlapping categories between the ShapeNetPart and KeypointNet [40] datasets to evaluate each method on both tasks without fine-tuning the method on each dataset. For all tasks, we follow the experiment setting of the previous work [3].

Baseline methods. Throughout the evaluation section, we mainly compare RIST against CPAE [3], the state-of-the-art self-supervised method to establish 3D dense correspondence by exploiting an intermediate UV canonical space. When open-sourced pre-trained models or codes are applicable, we also compare RIST with AtlasNetV2 [8], FoldingNet [38] and DPC [17]. AtlasNetV2 proposes to represent shapes as the deformation and combination of learnable elementary 3D structures, which can be extended to 3D correspondence establishment. FoldingNet introduces a folding-based decoder to 'fold' a canonical 2D grid into the 3D object surface, where the canonical 2D grid can be applied to identify cross-shape correspondences. DPC predicts 3D dense correspondence between non-rigidly deformed 3D humans, meaning that it can be a powerful baseline for both 3D part segmentation label transfer and 3D keypoint transfer tasks as well.

Method	w/o Rotations	w/ Rotations
FoldingNet [38]	23.2	23.3
AtlasNetV2 [8]	23.6	24.1
DPC [17]	23.9	23.9
CPAE [3]	24.4	23.9
RIST (ours)	**39.6**	**37.9**

Table 2. **Average IoU (%) of part label transfer for the chair category in the ScanObjectNN dataset [35].** Ours shows the best IoU both with and without rotation augmentations during training.

Implementation details. We use VN-DGCNN [7] as our SO(3)-equivariant encoder, and VN-based multi-layer perception as our SO(3)-equivariant decoder. For a fair comparison, we set the dimension, C, of SO(3)-equivariant global shape descriptor $\mathbf{Z} \in \mathbb{R}^{C\times 3}$ as 170 ($\approx 512/3$) since CPAE [3] uses 512-dimensional global shape descriptors. Following the training setup of CPAE [3], we use λ_{MSE}, λ_{EMD}, and λ_{CD} as 1000, 1, and 10, respectively. RIST is implemented in PyTorch, and is optimized with the Adam [16] optimizer at a constant learning rate of $1e^{-3}$.

4.1. Part Segmentation Label Transfer

We compare RIST with the state of the art in 3D part segmentation label transfer on ShapeNetPart [39] and ScanObjectNN [35]. For both datasets, we use the average of instance-wise IoU scores as the evaluation metric.

ShapeNetPart [39]. The quantitative results are presented in Table 1, where RIST outperforms CPAE [3], DPC [17], AtlasNetV2 [8], and FoldingNet [38] on all classes by a large margin with and without rotation augmentations dur-

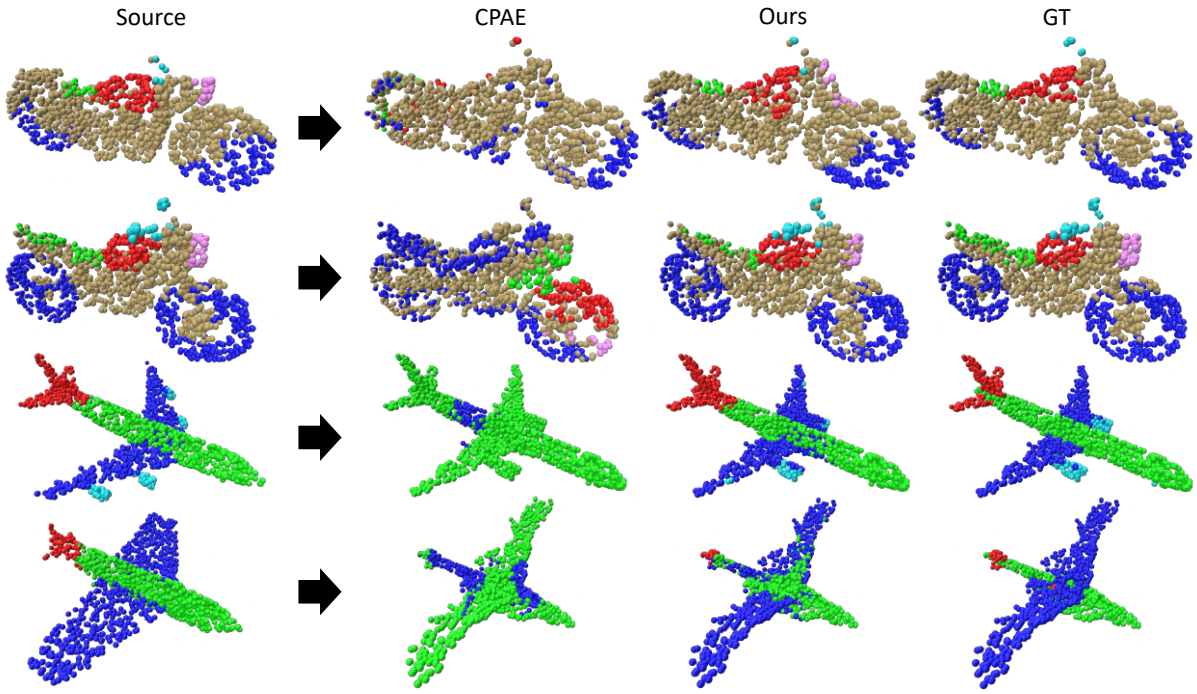

Figure 3. **Qualitative results of part label transfer on the ShapeNetPart dataset [39].** We visualize the label transfer results via learned correspondences of each method with the ground truth labels of targets. Note that the input shapes were arbitrarily rotated at evaluation, differently for both the source and targets of each row, but have been aligned in the above figure for better visibility of part label transfer results. RIST shows to outperform CPAE [3] consistently, showing a high resemblance to ground truth results.

ing training. We also provide the qualitative results of the part segmentation label transfer experiments on ShapeNetPart in Figure 3. Attributing to the SO(3)-invariant nature of correspondences established by RIST, we are able to transfer part labels significantly more accurately given randomly rotated shape pairs.

ScanObjectNN [35]. We evaluate each method trained on synthetic chair data of ShapeNet [1] on the real chair data of ScanObjectNN [35], which is partial and more challenging than ShapeNetPart [39], without fine-tuning. As shown in Table 2, RIST consistently outperforms previous approaches [3, 8, 17, 38] both with and without rotation augmentations during training. The qualitative results presented in Figure 4 show that RIST can predict rotation-robust 3D semantic correspondence between real and partial chair shapes, while CPAE [3] fails.

4.2. 3D Semantic Keypoint Transfer

Following the previous work [3], we compute the distances from the transferred M keypoints to the ground truth keypoints, and report PCK (Percentage of Correct Keypoints) of our transferred keypoints, which is computed by:

$$\text{PCK} = \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}[\|\mathbf{k}_m - \hat{\mathbf{k}}_m\| \leq \tau], \qquad (2)$$

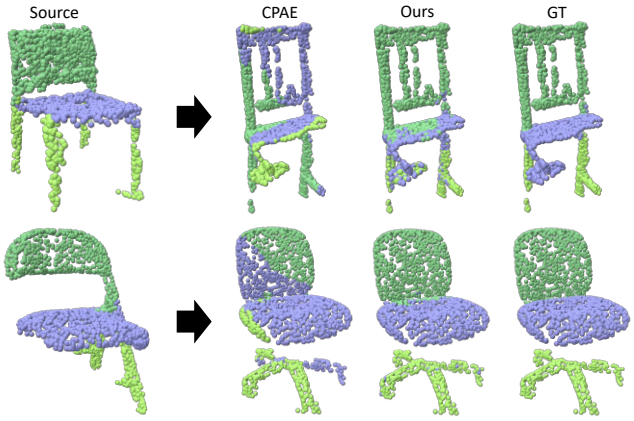

Figure 4. **Qualitative results of part label transfer on ScanObjectNN [35].** Note that both source and target point clouds were arbitrarily rotated at evaluation, but have been aligned in the figure for better visibility of part label transfer results. The results show that RIST reasonably predicts the semantic correspondences between arbitrarily rotated and partial real point clouds.

where τ, $\mathbf{k}_m$, and $\hat{\mathbf{k}}_m$ are a distance threshold, m-th ground truth keypoint, and m-th transferred keypoint, respectively. The results on the KeypointNet dataset [40] are illustrated in Figure 5 for varying distance thresholds τ. It can be seen that RIST consistently outperforms baseline methods with

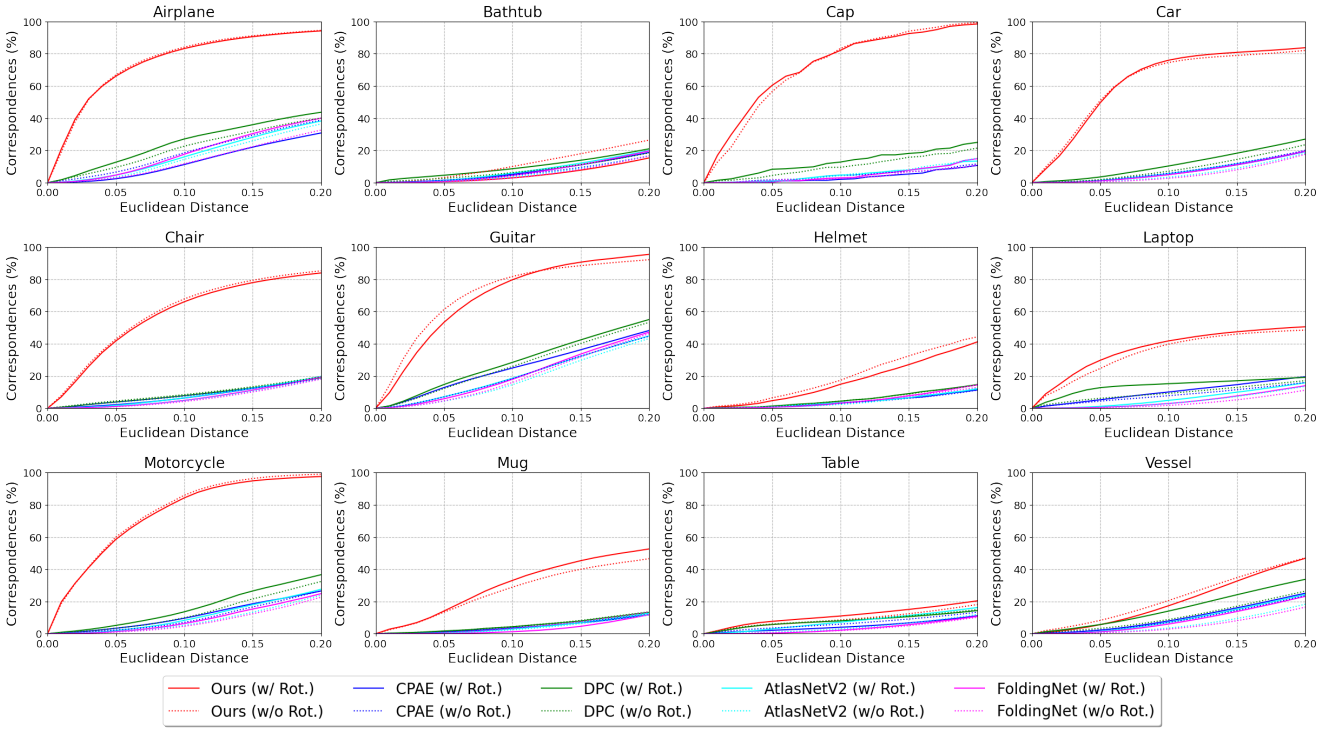

Figure 5. **Percentage of Correct Keypoints (PCK) for the 12 categories of the KeypointNet dataset [40] with and without rotation augmentations during training**. RIST consistently outperforms previous approaches on all classes and thresholds in both settings.

and without rotation augmentations during training for all classes, by up to 10× on certain classes and thresholds. This substantiates RIST's superior efficacy at establishing dense 3D correspondences between varying shapes. However, for certain classes such as Bathtub or Table, the performance is noticeably low, outperforming baseline methods only by a tight margin. We speculate this to be due to the prevalent rotational symmetry of those classes, making it especially challenging to establish accurate 3D correspondences under arbitrary rotations. The qualitative results of RIST in comparison to baseline methods are presented in Figure 6. It can also be seen that RIST can identify more accurate keypoint correspondences compared to CPAE [3] under arbitrary rotations, confirming the results of Figure 5.

4.3. Ablation Study and Analyses

We perform an ablation study to justify the design choice of RIST, and evidence the efficacy of each component.

Self- and cross-reconstruction. We train RIST in a self-supervised manner via penalizing errors in self- and cross-reconstruction of input point clouds. We conduct an ablation study on RIST's reconstruction, providing comparative results for scenarios with and without its use. The results are illustrated in the first graph of Figure 7. It can be seen that with and without rotation augmentations during training, incorporating *both* self- and cross-reconstruction yields the best results. Removing self-reconstruction results in a much dramatic drop in performance; we conjecture this is because without self-reconstruction, the dynamic local shape transform (Sec. 3.1.2) fails to capture the required locality of its own point cloud in the first place, being unsuitable to establish correspondences.

Encoder outputs and SO(3)-equivariance. RIST uses VNNs [7] as the SO(3)-equivariant layers to facilitate 3D dense correspondence establishment between arbitrarily rotated point cloud pairs, leveraging local shape transform to map global shape descriptors to local shape descriptors which encode the pointwise semantics and local geometry. We perform an ablation study to demonstrate the efficacy of local shape transforms and SO(3)-equivariant and -invariant representations in RIST on the motorcycle class of the KeypointNet dataset [40]. We start our comparison from the architecture of CPAE [3], given that they also employ an encoder-decoder architecture to self-supervise their network via shape reconstruction. The results are presented in the two rightmost graphs of Figure 7, showing the evaluation results with and without rotation augmentations during training in order. It can be seen that our design choice of using equivariant encoders and decoders shows consistent improvements over using an SO(3)-variant counterpart. Also,

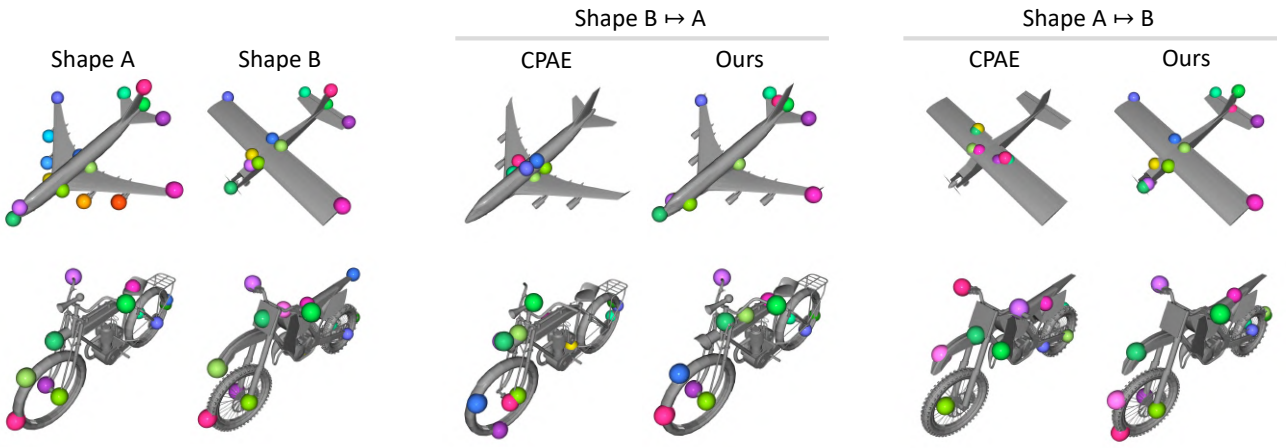

Figure 6. **Keypoint transfer results for airplane and motorcycle categories of KeypointNet [40]**. Each row contains a shape pair, each with ground-truth keypoints and the keypoint transfer results. Note that the input shapes were arbitrarily rotated at evaluation, but have been aligned in the above figure for better visibility of keypoint transfer results. RIST shows to transfer the keypoints more accurately.

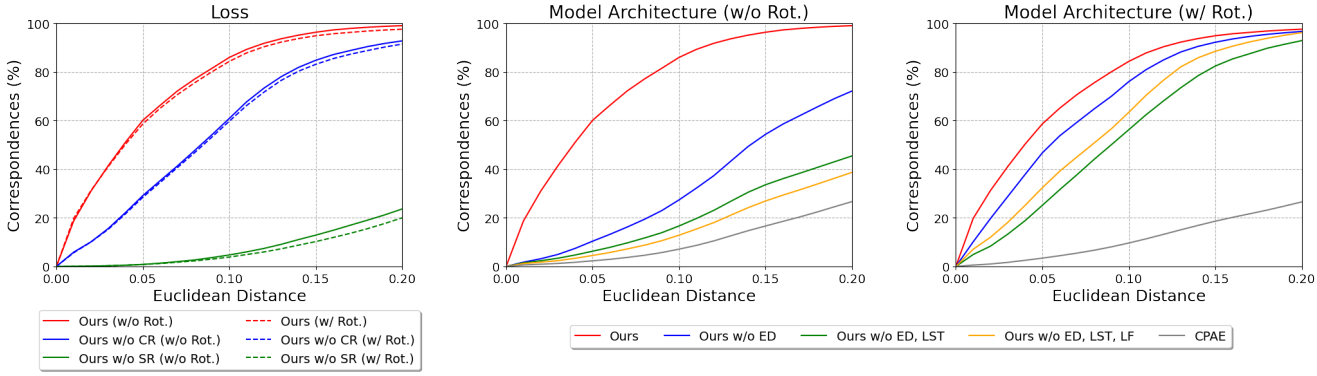

Figure 7. **Ablation study on losses (the leftmost) and the components of model architecture (the others)**; self-reconstruction loss (SR), cross-reconstruction loss (CR), equivariant decoder (ED), local shape transform (LST), and local feature (LF). When excluding the equivariant decoder, local shape transforms, or local features, the default is to use an SO(3)-variant decoder, UV coordinates [3], or global features for encoding, respectively.

using UV coordinates as proposed in CPAE [3] performs worse compared to our dynamic local shape transform, evidencing the comparatively better efficacy of transforming each point to their local shape descriptors via our dynamic SO(3)-invariant shape transform. While using local features as inputs to the encoder shows varied trends across with and without augmentation settings, using the point-wise local feature as input is a key component that facilitates the learning of *point-wise* dynamic local shape transform that is essential in establishing the 3D correspondences in RIST.

5. Conclusion

We've introduced RIST, a novel self-supervised learner for dense 3D semantic matching across shapes of the same category, even with arbitrary rotations. Its robustness stems from our innovative use of SO(3)-equivariant and -invariant representations, enabling dynamic local shape transforms that preserve rotation equivariance. These transforms map global descriptors to local ones, facilitating the establishment of dense correspondences. Our method outperforms existing ones on tasks like part label and keypoint transfer, enhancing applicability in computer vision and robotics, e.g., AR/VR and texture mapping. Future research could focus on improving robustness under common corruption.

Acknowledgement. This work was supported by IITP grants (No.2021-0-02068: AI Innovation Hub (50%), No. 2022-0-00290: Visual Intelligence for Space-Time Understanding and Generation (40%), No.2019-0-01906: Artificial Intelligence Graduate School Program at POSTECH (5%), NO.2021-0-01343, Artificial Intelligence Graduate School Program at Seoul National University (5%)) funded by the Korea government (MSIT). Seungwook was supported by the Hyundai-Motor Chung Mong-koo Foundation.

References

[1] 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*, 2015. 5, 6, 3, 4

[2] Haiwei Chen, Shichen Liu, Weikai Chen, Hao Li, and Randall Hill. Equivariant point network for 3d point cloud analysis. In *CVPR*, 2021. 2

[3] An-Chieh Cheng, Xueting Li, Min Sun, Ming-Hsuan Yang, and Sifei Liu. Learning 3d dense correspondence via canonical point autoencoder. In *NeurIPS*, 2021. 1, 2, 4, 5, 6, 7, 8, 3

[4] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. In *NeurIPS*, 2021. 2

[5] Jaesung Choe, Chunghyun Park, Francois Rameau, Jaesik Park, and In So Kweon. Pointmixer: Mlp-mixer for point cloud understanding. In *ECCV*, 2022. 2

[6] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. In *ICLR*, 2018. 2

[7] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for so(3)-equivariant networks. In *ICCV*, 2021. 3, 5, 7

[8] Theo Deprelle, Thibault Groueix, Matthew Fisher, Vladimir Kim, Bryan Russell, and Mathieu Aubry. Learning elementary structures for 3d shape generation and matching. In *NeurIPS*, 2019. 5, 6, 1, 2

[9] Qiang Hao, Rui Cai, Zhiwei Li, Lei Zhang, Yanwei Pang, Feng Wu, and Yong Rui. Efficient 2d-to-3d correspondence filtering for scalable 3d object recognition. In *CVPR*, 2013. 1

[10] Jiahui Huang, Tolga Birdal, Zan Gojcic, Leonidas J Guibas, and Shi-Min Hu. Multiway non-rigid point cloud registration via learned functional map synchronization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2038–2053, 2022. 2

[11] Shuaiyi Huang, Luyu Yang, Bo He, Songyang Zhang, Xuming He, and Abhinav Shrivastava. Learning semantic correspondence with sparse annotations. In *ECCV*, 2022. 2

[12] Oren Katzir, Dani Lischinski, and Daniel Cohen-Or. Shape-pose disentanglement using se (3)-equivariant vector neurons. In *ECCV*, 2022. 3

[13] Seungwook Kim, Juhong Min, and Minsu Cho. Transformatcher: Match-to-match attention for semantic correspondence. In *CVPR*, 2022. 2

[14] Seungwook Kim, Chunghyun Park, Yoonwoo Jeong, Jaesik Park, and Minsu Cho. Stable and consistent prediction of 3d characteristic orientation via invariant residual learning. In *ICML*, 2023. 2

[15] Seungwook Kim, Juhong Min, and Minsu Cho. Efficient semantic matching with hypercolumn correlation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 139–148, 2024. 2

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[17] Itai Lang, Dvir Ginzburg, Shai Avidan, and Dan Raviv. DPC: Unsupervised Deep Point Correspondence via Cross and Self Construction. In *3DV*, 2021. 5, 6, 1, 2

[18] Feiran Li, Kent Fujiwara, Fumio Okura, and Yasuyuki Matsushita. A closer look at rotation-invariant deep point cloud analysis. In *ICCV*, 2021. 2

[19] Xianzhi Li, Ruihui Li, Guangyong Chen, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. A rotation-invariant framework for deep point cloud analysis. *IEEE TVCG*, 28(12):4503–4514, 2021. 2

[20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 4

[21] Feng Liu and Xiaoming Liu. Learning implicit functions for topology-varying dense 3d shape correspondence. In *NeurIPS*, 2020. 1, 2, 4, 5, 3

[22] Andrew T Miller, Steffen Knoop, Henrik I Christensen, and Peter K Allen. Automatic grasp planning using shape primitives. In *ICRA*, 2003. 1

[23] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *ECCV*, 2022. 2

[24] Chunghyun Park, Yoonwoo Jeong, Minsu Cho, and Jaesik Park. Fast point transformer. In *CVPR*, 2022. 2

[25] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017.

[26] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, 2017. 2

[27] Samuele Salti, Federico Tombari, and Luigi Di Stefano. Shot: Unique signatures of histograms for surface and texture description. *CVIU*, 125:251–264, 2014. 2

[28] Ashutosh Saxena, Justin Driemeyer, Justin Kearns, and Andrew Ng. Robotic grasping of novel objects. In *NIPS*, 2006. 1

[29] Wen Shen, Binbin Zhang, Shikun Huang, Zhihua Wei, and Quanshi Zhang. 3d-rotation-equivariant quaternion neural networks. In *ECCV*, 2020. 2

[30] Ken Shoemake. Uniform random rotations. In *Graphics Gems III (IBM Version)*, pp. 124–132. Elsevier, 1992. 3

[31] Xiao Sun, Zhouhui Lian, and Jianguo Xiao. Srinet: Learning strictly rotation-invariant representations for point cloud classification and segmentation. In *ACM MM*, 2019. 2

[32] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv*, 2018. 2

[33] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique shape context for 3d data description. In *ACM workshop on 3D object retrieval*, 2010. 2

[34] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Probabilistic warp consistency for weakly-supervised semantic correspondences. In *CVPR*, 2022. 2

[35] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *ICCV*, 2019. 2, 5, 6, 1

[36] Tong Wu, Liang Pan, Junzhe Zhang, Tai Wang, Ziwei Liu, and Dahua Lin. Density-aware chamfer distance as a comprehensive metric for point cloud completion. In *NeurIPS*, 2021. 4

[37] Chenxi Xiao and Juan Wachs. Triangle-net: Towards robustness in point cloud learning. In *WACV*, 2021. 2

[38] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *CVPR*, 2018. 2, 5, 6, 1

[39] Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM TOG*, 35(6): 1–12, 2016. 1, 2, 5, 6, 3, 4

[40] Yang You, Yujing Lou, Chengkun Li, Zhoujun Cheng, Liangwei Li, Lizhuang Ma, Cewu Lu, and Weiming Wang. Keypointnet: A large-scale 3d keypoint dataset aggregated from numerous human annotations. In *CVPR*, 2020. 2, 5, 6, 7, 8, 1, 3

[41] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *CVPR*, 2020. 1

[42] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. Point transformer. In *ICCV*, 2021. 2

[43] Yongheng Zhao, Tolga Birdal, Haowen Deng, and Federico Tombari. 3d point capsule networks. In *CVPR*, 2019. 2

[44] Zerong Zheng, Tao Yu, Qionghai Dai, and Yebin Liu. Deep implicit templates for 3d shape representation. In *CVPR*, 2021. 3, 4

[45] Mohammad Zohaib and Alessio Del Bue. Sc3k: Self-supervised and coherent 3d keypoints estimation from rotated, noisy, and decimated point cloud data. In *ICCV*, 2023. 3