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

Supplementary Material

In this supplementary material, we provide a detailed explanation of RIST and additional experiment results.

A. Inference Algorithm of RIST

In this section, we provide a detailed algorithm for the inference process of RIST for a better understanding of RIST. As shown in Algorithm 1, given a pair of shapes, we cross-reconstruct one shape and find the nearest neighbor for each point of the cross-reconstructed shape from the other shape, following the inference algorithm of the previous work [3].

Algorithm 1 : Inference

```
Input: A pair of shapes $(\mathbf{P}, \mathbf{Q})$, Encoder$(\cdot)$, Decoder$(\cdot)$
Output: Correspondence $\mathcal{C}$ for all $\mathbf{p} \in \mathbf{P}$
$\mathbf{Z}_p, \{f^p_{\theta_i}\} \leftarrow$ Encoder$(\mathbf{P})$
$\mathbf{Z}_q, \{f^q_{\theta_i}\} \leftarrow$ Encoder$(\mathbf{Q})$
$\mathbf{Q}' \leftarrow$ Decoder$(\{f^p_{\theta_i}(\mathbf{Z}_q)\})$        ▷ Cross-recon. from $\mathbf{P}$ to $\mathbf{Q}$
$\mathcal{C} \leftarrow \{\}$                                              ▷ Initialization
for $\{i \leftarrow 1$ to $|\mathbf{P}|\}$ do
    $\mathbf{p} \leftarrow \mathbf{P}_i$
    $\mathbf{q} \leftarrow$ NearestNeighborSearch$(\mathbf{Q}'_i, \mathbf{Q})$
    $\mathcal{C} \leftarrow \mathcal{C} \cup (\mathbf{p}, \mathbf{q})$
end for
```

B. Ablation Study on Losses

In this section, we conduct an ablation study of the components of our self-reconstruction loss (Eq. 1) on the motorcycle category of the ShapeNetPart dataset [39]. As shown in Table A1, our choice (b) shows the best performance among models trained with the seven loss variants from (a) to (g).

	MSE	EMD	CD	IoU (%)
(a)	✓	✓	✓	46.0
(b)	✓	✓		**48.5**
(c)	✓		✓	47.0
(d)		✓	✓	46.4
(e)			✓	45.8
(f)		✓		46.4
(g)	✓			44.9

Table A1. **Ablation study on the self-reconstruction loss.** The model trained with ((b): MSE and EMD) shows the best performance, justifying our choice for the self-reconstruction loss.

C. Multi-class Training

In this section, we provide the experiment results of previous approaches [3, 8, 17, 38] and ours trained with multiple classes (airplane and chair) in the ShapeNetPart dataset [39]. As shown in Table A2, RIST outperforms previous approaches [3, 8, 17, 38] by a large margin.

Method	Airplane	Chair	Average
FoldingNet [38]	20.9	23.9	22.4
AtlasNetV2 [8]	21.1	24.6	22.9
DPC [17]	22.7	25.6	24.2
CPAE [3]	16.6	14.8	15.7
RIST (ours)	34.4	34.7	**34.6**

Table A2. **Part label transfer with multi-classes training.**

D. Generalization to Unseen Classes

In this section, we evaluate the generalization ability of previous approaches [3, 8, 17, 38] and ours to unseen classes. Specifically, we train each method on the airplane category in the ShapeNetPart dataset [39] and test it on the chair category. As shown in Table A3, RIST shows a competitive result with an unseen category, outperforming previous approaches [3, 8, 38] except DPC [17].

FoldingNet [38]	AtlasNetV2 [8]	DPC [17]	CPAE [3]	RIST
24.8	23.0	**28.2**	15.6	27.3

Table A3. **Generalization results for the part label transfer.**

E. Inference on Aligned Shapes

In this section, we provide the results of previous approaches [3, 8, 17, 38] and ours evaluated on the ShapeNetPart [39], ScanObjectNN [35], and KeypointNet [40] datasets, but with aligned test shapes, as shown in Table A4, Table A5, and Figure A1, respectively. Note that under the aligned setting, the input shape pairs are perfectly aligned both at train and test time - which is an unrealistic setting in practice. For each method, we also include the results with rotated shapes to show the performance difference between aligned and rotated settings. It can be seen that while the drop in performance for previous approaches [3, 8, 17, 38] from the aligned to the rotated setting is drastic, the difference is negligible in RIST, demonstrating the robustness of our SO(3) correspondence establishment scheme against arbitrary rotations. While RIST is not always competitive on all settings, it is impractical to expect perfectly aligned shapes in real-world situations; on the realistic setting of SO(3) evaluation, RIST consistently shows the best results.

Inference	Method	Airplane	Cap	Chair	Guitar	Laptop	Motorcycle	Mug	Table	Average
Aligned	FoldingNet [38]	56.5	54.9	63.1	73.1	81.9	21.5	75.5	54.0	60.1
	AtlasNetV2 [8]	51.7	44.7	63.3	65.0	84.0	41.5	84.2	59.3	61.7
	DPC [17]	60.5	65.8	65.3	74.4	88.0	53.3	85.4	66.4	69.9
	CPAE [3]	61.3	61.6	72.6	78.9	89.9	55.4	86.5	72.5	**72.3**
	RIST (ours)	52.1	54.4	58.3	74.1	56.7	48.7	75.6	41.3	57.7
Rotated	FoldingNet [38]	17.8	34.7	22.5	22.1	36.2	12.6	50.0	34.6	28.8 (↓ 31.3)
	AtlasNetV2 [8]	19.7	31.4	23.6	22.7	36.0	13.1	49.7	35.2	28.9 (↓ 32.8)
	DPC [17]	22.7	37.1	25.6	31.9	35.0	17.5	51.3	36.8	32.2 (↓ 37.7)
	CPAE [3]	21.0	38.0	26.0	22.7	34.9	14.7	51.4	35.5	30.5 (↓ 41.8)
	RIST (ours)	52.1	54.5	58.3	74.1	56.5	48.6	75.0	41.3	**57.6** (↓ 0.1)

Table A4. **Average IoU (%) of part label transfer for eight categories in the ShapeNetPart dataset [39] on aligned and rotated shapes.** Note that each method is trained without rotation augmentation. RIST shows the most negligible performance drop (0.1% in IoU) with rotated shapes, while previous approaches [3, 8, 17, 38] show large performance drops (at least 30% in IoU).

Inference	FoldingNet [38]	AtlasNetV2 [8]	DPC [17]	CPAE [3]	RIST (ours)
Aligned	33.6	34.8	36.3	33.8	**39.6**
Rotated	23.2 (↓ 10.4)	23.6 (↓ 11.2)	23.9 (↓ 12.4)	24.4 (↓ 9.4)	**39.6** (−)

Table A5. **Average IoU (%) of part label transfer for the chair category in the ScanObjectNN dataset [35] on aligned and rotated shapes.** Note that each method is trained without rotation augmentation. RIST does not show any performance drop with rotated shapes, while previous approaches [3, 8, 17, 38] show large performance drops (at least 9% in IoU).

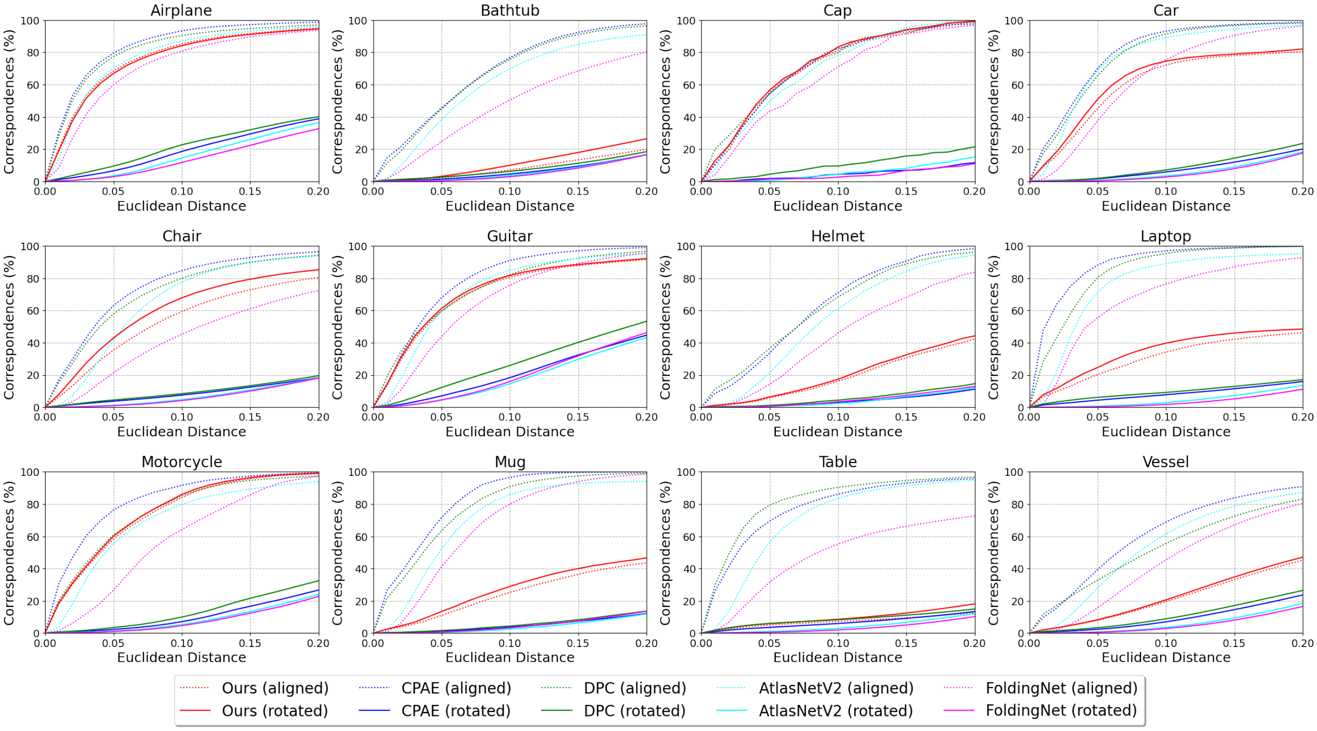

Figure A1. **Percentage of Correct Keypoints (PCK) for the 12 categories of the KeypointNet dataset [40] on aligned and rotated shapes.** Note that each method is trained without rotation augmentation. While previous approaches [3, 8, 17, 38] are vulnerable to rotations, RIST shows a negligible performance drop with rotated shapes.

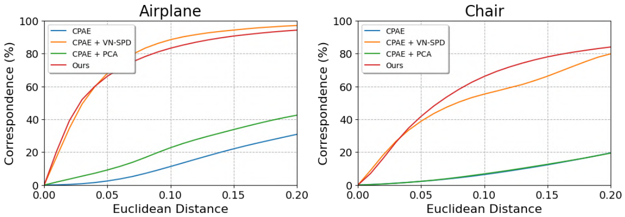

Figure A2. **Comparision of RIST with the combination of CPAE [3] and aligning methods; PCA and VN-SPD [12].** Note that RIST shows competitive results to the combined method of CPAE [3] and VN-SPD [12], which is a SE(3)-equivariant orientation predictor and requires additional parameters (17.4M).

We further evaluate CPAE [3] when integrated together with two 3D shape alignment methods. We present results when using PCA or VN-SPD [12] as the alignment method in Figure A2. PCA yields only marginal performance improvements, likely due to sign and order ambiguities. In contrast, using the SOTA learning-based alignment method, VN-SPD, produces competitive results with RIST, but its inconsistency in aligning shapes to a canonical orientation limits performance as also reported in Katzir *et al.* [12], sacrificing the efficiency.

F. Comparision with 3D Keypoint Estimators

We compare RIST with SC3K [45], a recent self-supervised method for coherent 3D keypoint estimation. However, since this 3D keypoint *estimation* method cannot be evaluated on 3D keypoint *matching*, we instead used the Dual Alignment Score (DAS)[4] of RIST for an empirical comparison with SC3K. Additionally, to facilitate comparison on the part label transfer task, we extended the number of keypoints estimated by SC3K to match the total number of points in a point cloud *e.g.*, 2048.

Method	Airplane	Car	Chair
SC3K [45]	81.3	73.8	**86.2**
RIST (ours)	**82.4**	**76.9**	81.8

Table A6. **Dual Alignment Score of SC3K [45] and RIST.** During the evaluation, we use 10 keypoints for both SC3K and RIST.

Method	Airplane	Car	Chair
SC3K [45]	22.3	23.0	24.7
RIST (ours)	**51.2**	**48.0**	**55.0**

Table A7. **Part label transfer results of SC3K [45] and RIST.** Note that we train SC3K [45] with 2048 keypoints.

As shown in Tables A6 and A7, RIST exhibits competitive DAS results compared to SC3K, although it is not trained for 3D keypoint estimation, and significantly outperforms SC3K on the part label transfer task.

G. Evaluation with Pseudo-Ground Truth

We utilize DIT [44] to establish pseudo-ground truth on ShapeNet [1] for a direct evaluation of RIST's dense semantic correspondence capabilities for airplane, car, and chair classes, using official checkpoints. As shown in Figure A3, the results show a similar trend of part label transfer results, showing that RIST outperforms previous approaches.

H. Implementation Details of SO(3)

In this section, we explain the implementation details of uniformly sampling random rotations and highlight the differences from the previous approach [3]. Cheng *et al.* [3] samples rotation angles from $\mathcal{N}(0, 0.2^2)$ and then clamps them to $[-\frac{1}{2}\pi, \frac{1}{2}\pi]$, which limits the range of rotation. In our work, we follow Shoemake *et al.* [30] to *uniformly* sample to cover full SO(3), which is more challenging.

I. Part Label Transfer Results on More Classes

We initially presented evaluations only for the classes of ShapeNetPart [39] that are shared with those of KeypointNet [40]. In Table A8, we further present part label transfer results on the remaining classes of ShapeNetPart [39].

Method	Bag	Car	Ear.	Knife	Lamp	Pistol	Rocket	Skate.
CPAE [21]	43.2	20.3	33.4	36.3	31.1	26.8	27.7	52.0
RIST (ours)	**50.8**	**48.0**	**36.3**	**57.9**	**35.9**	**54.7**	**34.4**	**54.4**

Table A8. **Part label transfer results on ShapeNetPart [39].** RIST consistently outperforms the previous state-of-the-art method on the remaining classes of ShapeNetPart [39].

J. Matching with Local Shape Transform

In this section, We experiment with a variant of RIST (RIST_{LST}), which matches 3D shapes using similarity between SO(3)-invariant Local Shape Transform (LST). As shown in Table A9, our current scheme of comparing point positions yields noticeably better results on ShapeNetPart [39], meaning that our trained decoder is better adept at handling topology-varying structures

K. Alignment of Qualitative Results

In this section, we provide both unaligned and aligned qualitative results for a better understanding of how our qualitative results were drawn. As shown in Figure A4, both

[4] A metric for 3D keypoint estimation task SC3K [45] used.

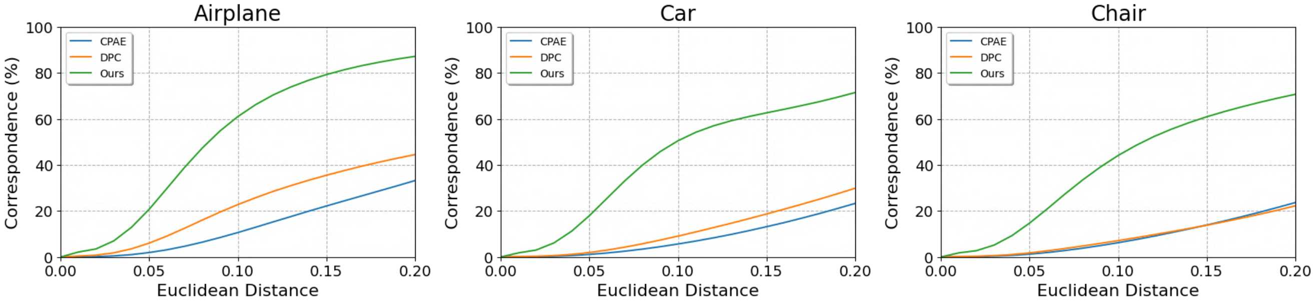

Figure A3. **3D semantic correspondence results of RIST using DIT [44] as pseudo-ground truth on ShapeNet [1].** We use the official checkpoints of DIT to generate pseudo-ground truth of 3D semantic correspondence for rotated 3D shapes.

Method	Airplane	Chair
CPAE	17.0	24.5
RIST$_{\text{LST}}$	48.6	50.3
RIST	**51.2**	**55.0**

Table A9. **Part label transfer results of RIST$_{\text{LST}}$.** Note that we use randomly rotated 3D shapes for the evaluation.

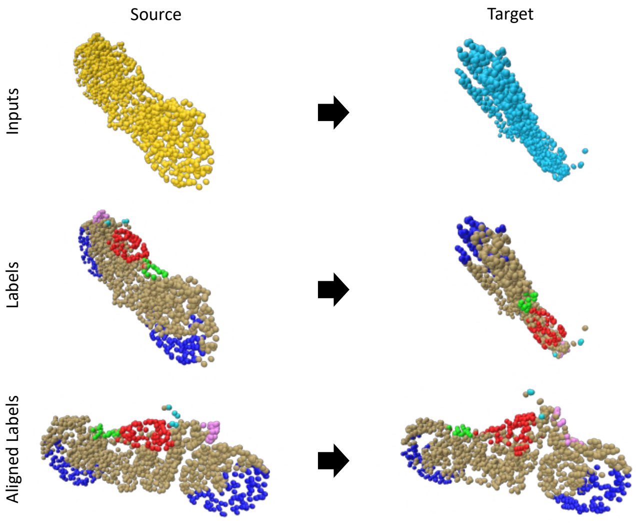

Figure A4. **Visualization for aligning qualitative results of part label transfer on the ShapeNetPart dataset [39].**

source and target shapes are randomly rotated at the inference time. Note that we use the part segmentation labels transferred by RIST for the target shape in Figure A4.

L. Qualitative Results of Part Label Transfer

We provide additional qualitative results on the ShapeNetPart dataset [39] that were not included in our manuscript due to space constraints, as shown in Figures A5 and A6.

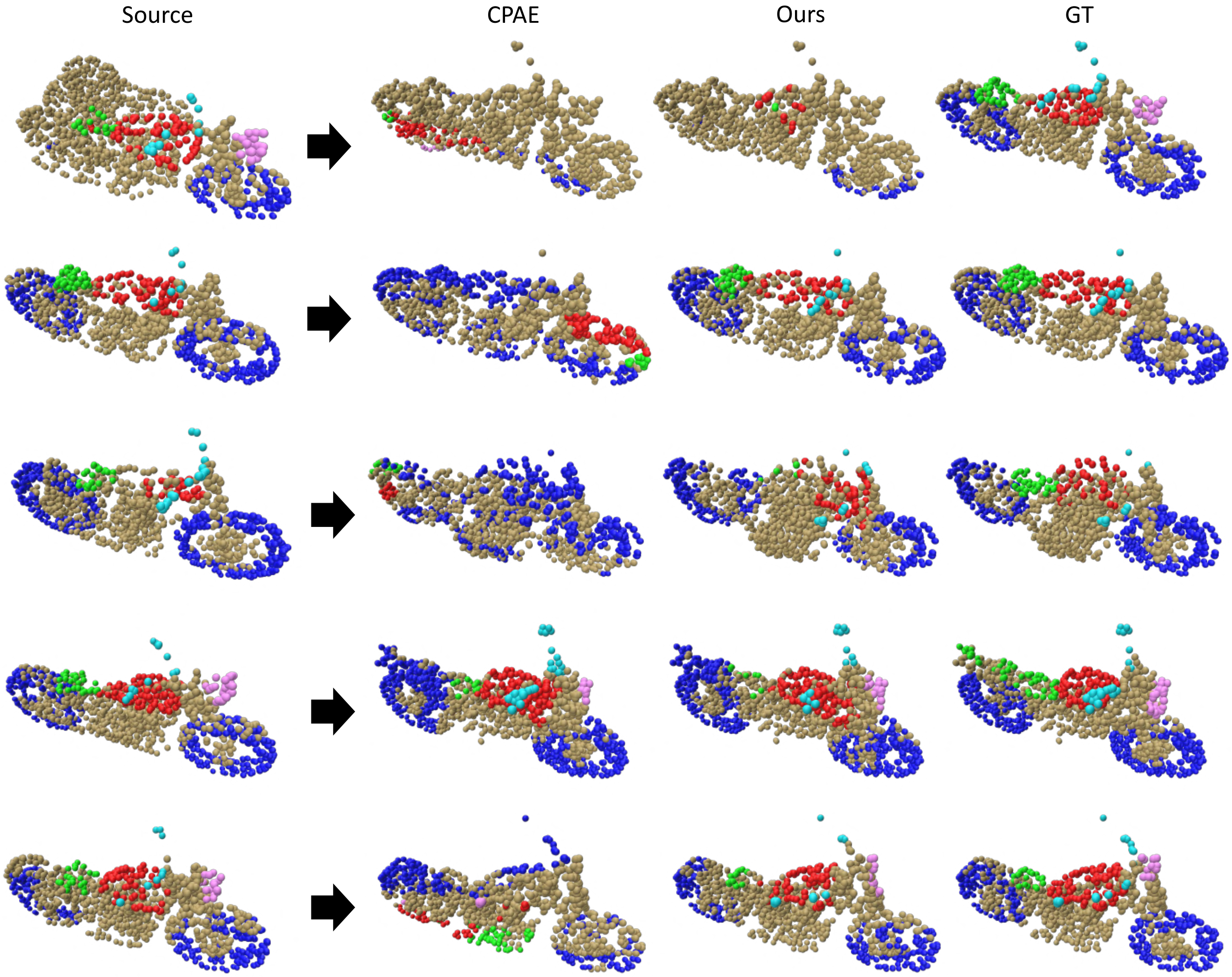

Figure A5. **Qualitative results of part label transfer on the motorcycle class in the ShapeNet part dataset [39].** Note that the input shapes were arbitrarily rotated, differently for each target column, but have been aligned for better visibility of part label transfer results. RIST shows to outperform CPAE [3] consistently, showing a high resemblance to ground truth results.

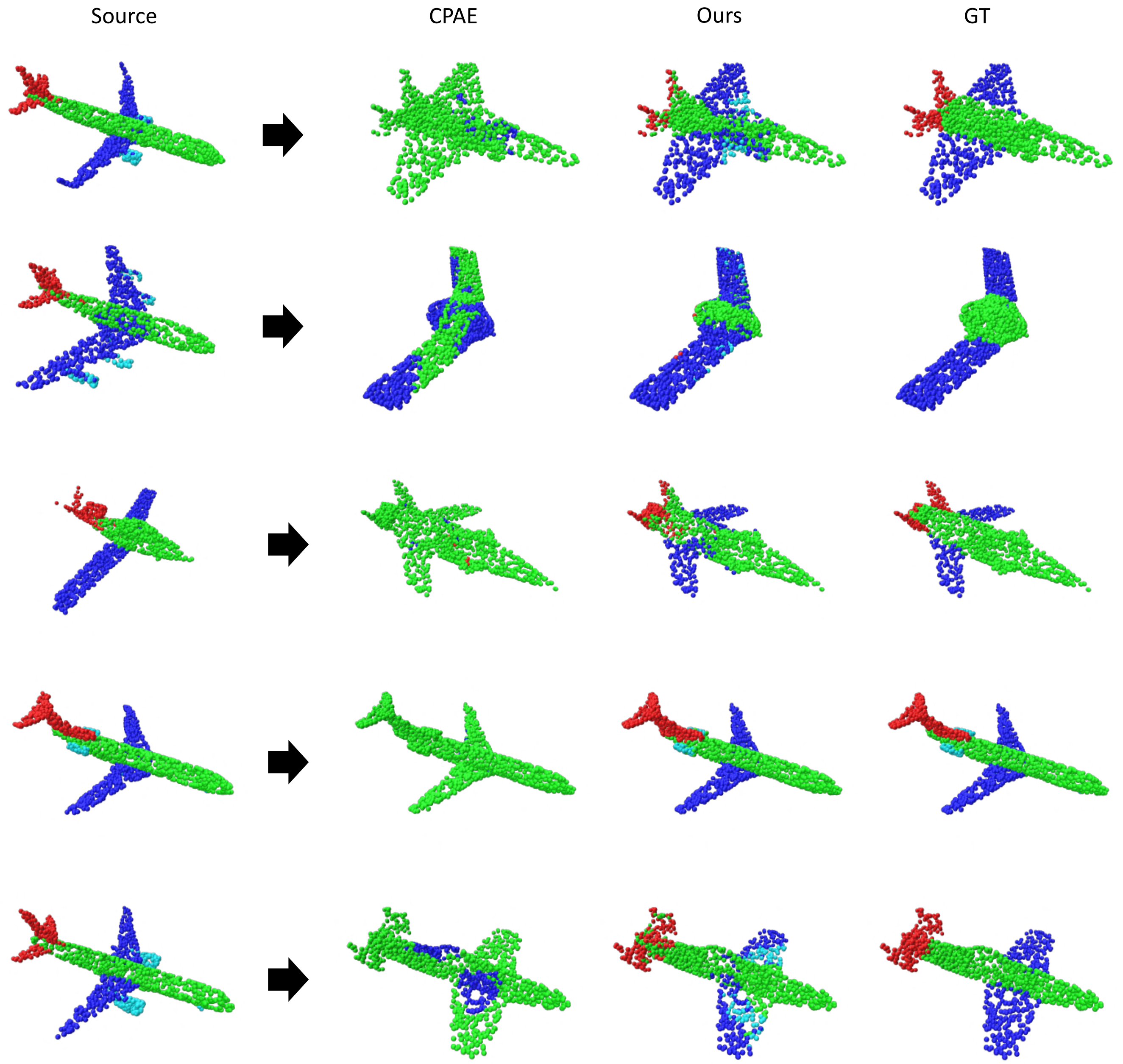

Figure A6. **Qualitative results of part label transfer on the airplane class in the ShapeNet part dataset [39].** Note that the input shapes were arbitrarily rotated, differently for each target column, but have been aligned for better visibility of part label transfer results. RIST shows to outperform CPAE [3] consistently, showing a high resemblance to ground truth results.

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