Semantic-Aware Implicit Template Learning via Part Deformation Consistency (Supplement)

We provide additional experimental results/details and discussion in this supplement. The supplement consists of (1) experiment details (*e.g.*, dataset statistics, implementation details), (2) derivation of the closed form for the optimal global scaling factor r given deformation $\mathcal{D}$, (3) sensitivity test for different coefficients of proposed regularizations, (4) additional results with different semantic priors, (5-7) qualitative results on ShapeNetV2 [1], ScanObjectNN [2], and DFAUST [3], (8) analysis of learned implicit template fields, and (9) limitations. In each visualization, note that all shapes placed on the *ivory* background (and leftmost) are source shapes for label transfer tasks.

A. Experiment details

A.1. Data statistics

We mainly use four categories (*e.g.*, chair, table, airplane and car) of ShapeNetV2[1] [1] with labels from ShapeNetPart [4] and KeypointNet [5]. We follow DIF[2] [6] for data splitting and data preparation. Table 1 summarizes the data statistics. For additional experiments in Section 5.3, first, we use a subset of ShapeNetV2 [1] that consists of two subcategories in each category (*e.g.*, airplane: airliner/jet, car: sedan/jeep, chair: straight, chair/sofa, table: pedestal table/short table). As ShapeNetV2 contains incorrect subcategory labels, we manually cleaned subcategory labels and sample 50 shapes for each subcategory. To avoid any bias towards a specific shape structure, we use an equal number of subcategory shapes. Second, for DFAUST [3] dataset, we select two different subjects (50002 and 50026), where 50002 is of a larger size, being overweight and taller compared to 50026, and four distinct actions (*e.g.*, shake arms, chicken wings, running on spot, and jumping jacks).

A.2. Implementation details

For feature extraction, we mainly use BAE-Net[3] [7] and follow their training schemes. In detail, we prepare 8,192 surface points from voxel grid sampling and 32,768 query points from random sampling for each shape following IM-Net [8]. We train the encoder with a batch size of 1 for 60 epochs, using Adam [9] optimizer with a learning rate of 0.0001 for car and 0.00005 for the rest of categories. It takes about 2 hours on average to fully train the encoder with a single GPU (RTX 2080ti). For the hyperparameter k, we use 6/8/4/4 for airplane/car/chair/table category, and we use 256 for the global latent code dimension.

Table 1. **Data statistics.**

	Category			
	airplane	**car**	**chair**	**table**
Training data	3500	3000	4000	4000
Part labeled data	3195	2728	3551	3671
Keypoint labeled data	889	870	577	545
Evaluation data (recon)	100	100	100	100

For training deformation field $\mathcal{D}_{\theta_1}$ and template field $\mathcal{T}_{\theta_2}$, we train the model with a batch size of 128 for 60 epochs, using Adam [9] optimizer with the learning rate of 0.0001. It takes about 6 hours on average to fully train the model. Also, we use 256 for the dimension of each shape latent code and we use 64 for the dimension of each part deformation prior.

For the proposed regularizations, we grid search coefficients such as global scale consistency $\mathcal{L}_{\text{scale}}$ and global deformation consistency $\mathcal{L}_{\text{geo}}$ in the range of [10,500]/[50,100]. For the most *influential* regularization, which is part deformation consistency $\mathcal{L}_{\text{pdc\_geo}}$, we grid search the coefficient in the separate range according to the categories (*e.g.*, [1000,2000] for chair, [750,1000] for airplane and car, [250,500] for table). The final coefficients we used are described in Table 2. For the rest of the regularizers, we fix the coefficient as 50 for $\mathcal{L}_{\text{pdc\_sem}}$, 100 for $\mathcal{L}_{\text{normal}}$, 1e6 for $\mathcal{L}_{\text{emb}}$ in every category. For the regularizers for deformation smoothness $\mathcal{L}_{\text{smooth}}$, and minimal correction $\mathcal{L}_{\text{c}}$, we grid search in the range of [1,5,10]/[50,100,500] and apply 10/5/5/1 and 50/10/500/50 in sequential order to the four categories: airplane, car, chair, and table.

For baseline models, we validate surrogate tasks with provided pretrained models; if none exists, we train the model from scratch based on source code from the origi-

[1]Copyright (c) 2022 ShapeNET.
[2]https://github.com/microsoft/DIF-Net
[3]https://github.com/czq142857/BAE-NET

Table 2. **Regularization coefficients for each category.**

coef.	$\mathcal{L}_{\text{pdc\_geo}}$	$\mathcal{L}_{\text{scale}}$	$\mathcal{L}_{\text{geo}}$
airplane	750	10	50
car	1000	10	50
chair	2000	500	100
table	250	500	100

nal authors, *e.g.*, DIT[4] and AtlasNetV2[5]. Unlike implicit template learning models, AtlasNetV2 [10] does not learn a global template, rather it learns decomposed local patches. Here, for AtlasNetV2 [10], we evaluate the part label transfer task as identical as DIF [6], and also similarly evaluate the keypoint label transfer task, where we use the average points of each corresponding keypoint from source shapes as transferred keypoint labels. Lastly, all experiments are implemented in Pytorch[6] [11] and Pytorch3D [12] and conducted on 4 NVIDIA RTX A6000.

B. Proof of global scaling factor r

We propose global scale consistency regularization to preserve the scale of the implicit template field against strong deformations based on the following lemma and its proof.

Lemma 1. *Given a scalar field (shape) $X \in \mathbb{R}^{3\times M}$ and a non-rigid deformation $\mathcal{D} : \boldsymbol{x} \in \mathbb{R}^3 \rightarrow \Delta\boldsymbol{x} \in \mathbb{R}^3$, we define a global scaling factor r of $\mathcal{D}$ as an optimal solution to the following problem:*

$$r^* = \underset{r}{\operatorname{argmin}} \sum_{i=1}^{M} \|\boldsymbol{x}_i + \Delta\boldsymbol{x}_i - r\boldsymbol{x}_i\|_2^2. \tag{1}$$

Then, the optimal solution can be analytically obtained by

$$r^* = \frac{\sum_{i=1}^{M} \boldsymbol{x}_i^\top(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i)}{\sum_{j=1}^{M}(\boldsymbol{x}_j^\top \boldsymbol{x}_j)}, \tag{2}$$

where $\boldsymbol{x}_i \in X$ and $\Delta\boldsymbol{x}_i \in \mathcal{D}(x_i)$.

Proof. We defined the global scaling factor as the optimal solution to the following problem:

$$r^* = \underset{r}{\operatorname{argmin}} \sum_{i=1}^{M} \|\boldsymbol{x}_i + \Delta\boldsymbol{x}_i - r\boldsymbol{x}_i\|_2^2. \tag{3}$$

To find a closed-form solution, we differentiate (3) by r:

$$\begin{aligned}\frac{\partial}{\partial r}\sum_{i=1}^{M} \|\boldsymbol{x}_i + \Delta\boldsymbol{x}_i - r\boldsymbol{x}_i\|_2^2 &= \sum_{i=1}^{M}(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i - r\boldsymbol{x}_i)^\top \boldsymbol{x}_i \\ &= \sum_{i=1}^{M}(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i)^\top \boldsymbol{x}_i - \sum_{j=1}^{M} r\boldsymbol{x}_j^\top \boldsymbol{x}_j = 0 \end{aligned} \tag{4}$$

Finally, we can acquire the solution $r^* = \frac{\sum_{i=1}^{M} \boldsymbol{x}_i^\top(\boldsymbol{x}_i+\Delta\boldsymbol{x}_i)}{\sum_{j=1}^{M} \boldsymbol{x}_j^\top \boldsymbol{x}_j}$ from (4). □

In this paper, we encourage the learned implicit template to preserve its scale by the *average* scale of N deformed shapes in a single batch. That is, we simply estimate the global scaling of all N shapes in a mini-batch by $r_{\text{batch}} = \frac{\sum_{s=1}^{N}\sum_{i=1}^{M} \boldsymbol{x}_i^{s\top}(\boldsymbol{x}_i^s+\Delta\boldsymbol{x}_i^s)}{\sum_{s'=1}^{N}\sum_{j=1}^{M} \boldsymbol{x}_j^{s'\top}\boldsymbol{x}_j^{s'}} \approx 1$. This treats all shapes as a point cloud and finds one global scaling of it.

Further, we can impose a slightly different regularization with the expectation of r_s. Given a set of points $\{x_i^s\}_i$ in shape s, a shape-specific global scaling factor r_s and its expectation are defined as

$$\begin{aligned} r_s &= \frac{\sum_{i=1}^{M} \boldsymbol{x}_i^{s\top}(\boldsymbol{x}_i^s + \Delta\boldsymbol{x}_i^s)}{\sum_{j=1}^{M} \boldsymbol{x}_j^{s\top}\boldsymbol{x}_j^s} \\ \mathbb{E}_s[r_s] &= \sum_{s=1}^{N} \frac{1}{N} r_s \end{aligned} \tag{5}$$

Then, the regularizer is given as

$$\mathcal{L}_{\text{scale}} = |\mathbb{E}[r] - 1|. \tag{6}$$

We observed that in our preliminary experiments, these regularizations above allow more flexibility than enforcing the global scaling of each individual shape, *i.e.*, $\sum_{s=1}^{N} |r_s - 1|$. The final equation of global scaling consistency regularization is in (9) of the main paper.

C. Sensivity test for different coefficients

We analyze the effect of coefficients for suggested regularizations ($\mathcal{L}_{\text{pdc\_geo}}/\mathcal{L}_{\text{geo}}/\mathcal{L}_{\text{scale}}$). We perform experiments with a sub-dataset (same as Table 5 in the main paper) for ShapeNet airplane/chair and report mIoU as 2-shot part label transfer results. Figure 1 shows that $\mathcal{L}_{\text{pdc\_geo}}$ (highlighted with the red box for cases without $\mathcal{L}_{\text{pdc\_geo}}$) significantly improves performance. When the coefficient of $\mathcal{L}_{\text{pdc\_geo}}$ is properly set, our method is robust to the choice of weights for other losses: $\mathcal{L}_{\text{geo}}$ and $\mathcal{L}_{\text{scale}}$. As shown in the boxes highlighted in yellow and green, our method stably achieves good performance regardless of the coefficients for $\mathcal{L}_{\text{geo}}$ and $\mathcal{L}_{\text{scale}}$.

D. Different semantic priors

We provide an additional experimental result with RIM-Net [13], which is a self-supervised co-segmentation model for 3D object shapes. The pre-trained RIM-Net[7] is used and

[4] https://github.com/ZhengZerong/DeepImplicitTemplates

[5] https://github.com/TheoDEPRELLE/AtlasNetV2

[6] Copyright (c) 2016-Facebook, Inc (Adam Paszke), Licensed under BSD-style license

[7] https://github.com/chengjieniu/RIM-Net

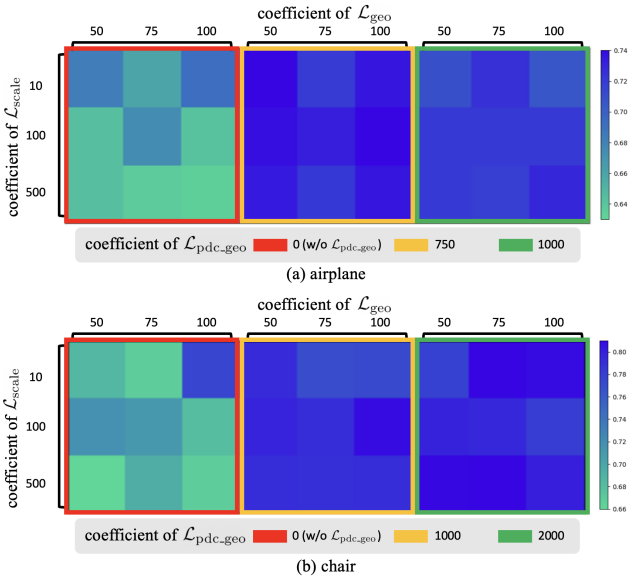

Figure 1. **Sensitivity test for $\mathcal{L}_{\text{pdc\_geo}}/\mathcal{L}_{\text{geo}}/\mathcal{L}_{\text{scale}}$.**

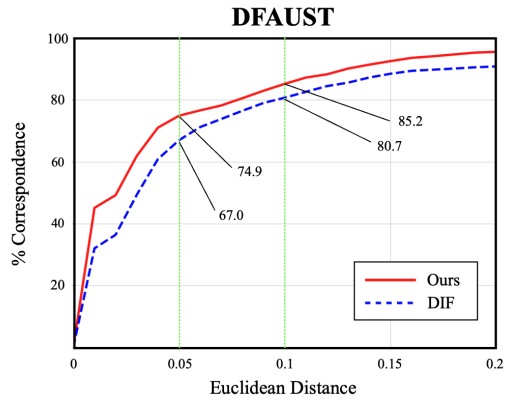

Figure 2. **Keypoint transfer performance for DFAUST [3].** We measure correspondence accuracy (PCK score).

we conduct 2-shot part label transfer with the subcategories for the chair of ShapeNetV2. Since RIM-Net provides different levels of part semantics, *e.g.*, two-part partitions for level 1, and eight-part partitions for level 3, we leverage this characteristic to evaluate our framework across different levels of part quality. Based on Table 3, we believe that even if the given prior does not have high-level semantics, our framework improves performance as long as the prior is *consistent*.

Table 3. **Utilizing different semantic priors**. 2-shot part label transfer in subcategories (straight chair and sofa) for ShapeNet chair.

	Ours-RIM(lv1)	Ours-RIM(lv3)	Ours-BAE	DIF
chair	70.1	78.3	80.7	67.2

Figure 3. **Visualization of learned implicit template fields for four categories by extracting different iso-surfaces.**

E. Qualitative results on ShapeNet

We provide additional visualizations as in Figure 4, Figure 5, and 6 to show keypoint/part label/texture transfer results on ShapeNetV2 [1]. These results illustrate the importance of imposing semantics in implicit template learning for generic object shapes.

F. Keypoint transfer on ScanObjectNN

In Figure 7, we conduct qualitative analysis via keypoint transfer task on ScanObjectNN [2] to validate the robustness of our method even with the domain gap between synthetic data and real-world data. We transfer the keypoint labels of chair/table categories in ShapeNetV2 [1] to corresponding shapes in ScanObjectNN [2]. Since shapes in ScanObjectNN are real-world scanned data, *i.e.*, they are not always watertight, we only use scanned surface points for inference. Our framework shows superior performance over baseline models, supporting the importance of understanding semantics for shape correspondence. Visualizations of the chair in Figure 7 are clear examples. Two keypoints of pea-green and orange are geometrically close but semantically different in the source chair. After transferring keypoints, our framework is the only model that enables disentangling two keypoints in the target chair, where arms and legs are separated, unlike the source chair.

G. Keypoint transfer on DFAUST

We provide detailed keypoint transfer results to evaluate unsupervised correspondence performance on non-rigid shapes [3] in Figure 2 and Figure 8. For evaluation, we assume that we do not know the ground truth correspondence between human shapes. Although learning a suitable global template for various actions with large local deformation scales is a challenging task, our method consistently shows better and more consistent correspondence accuracy compared to the baseline, as shown in the following figures. In particular, we present the full keypoint transfer result in Figure 2, where we use source shape/selected keypoints as in the shape in Figure 8 (leftmost shape in beige). The performance is PCK scores given a continuous range of thresholds. We observe our framework (red) consistently outperforms the baseline (blue), indicating that learned correspondences from our framework are more accurate. Visualizations in Figure 8 also support our framework, *e.g.*, our model transfers "hand" keypoints (highlighted in blue balls) while the baseline transfers them to the waist or the elbow.

H. Analysis of learned implicit template fields

Figure 3 illustrates the iso-surfaces of four categories (*e.g.*, airplane, car, chair, table) extracted from learned implicit template fields. Our framework learns a template by imposing semantically consistent mapping, rather than learning realistic templates. We observe that the iso-surface of the learned templates sometimes has stretched parts in more challenging categories, which is beneficial for semantically mapping shapes with high variability. Thus, it leads to improved performance on dense correspondence.

I. Limitations

Since we utilize self-supervised segmentation models for knowledge distillation, the performance can be highly dependent on the part segmentation quality of the feature extractor. However, we have demonstrated that our framework consistently improves correspondence performance with various semantic priors, even with coarse part semantics such as BAE-Net [7] with $k = 2$ or level 1 RIM-Net [13]. Note that, unlike 2D domain, there are no powerful self-supervised segmentation models available yet, such as DINO [14]. If such models emerge for 3D domain, our framework can potentially achieve even better performance. These are left for future works.

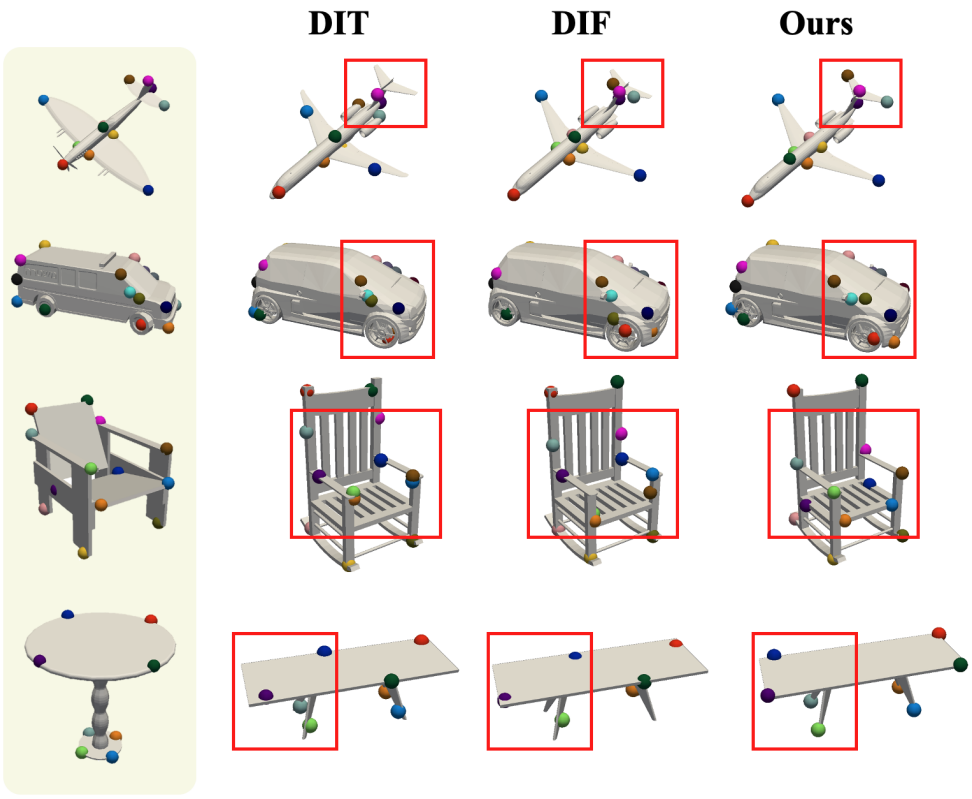

Figure 4. **Additional comparison on keypoint transfer in ShapeNetV2.**

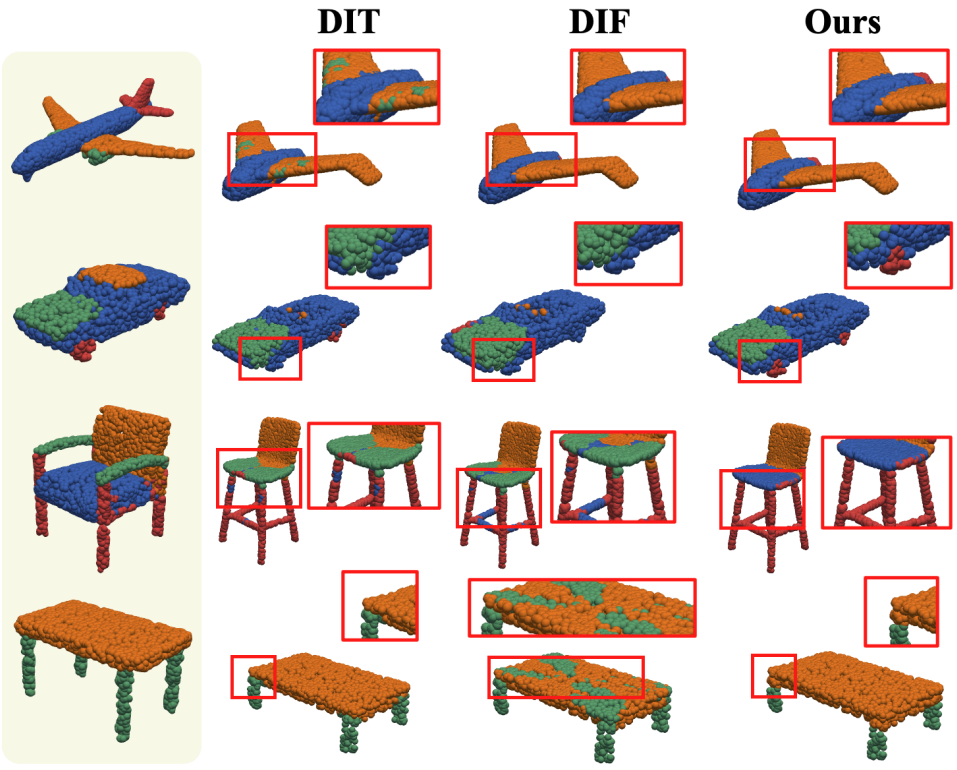

Figure 5. **Additional comparison on part label transfer ShapeNetV2.**

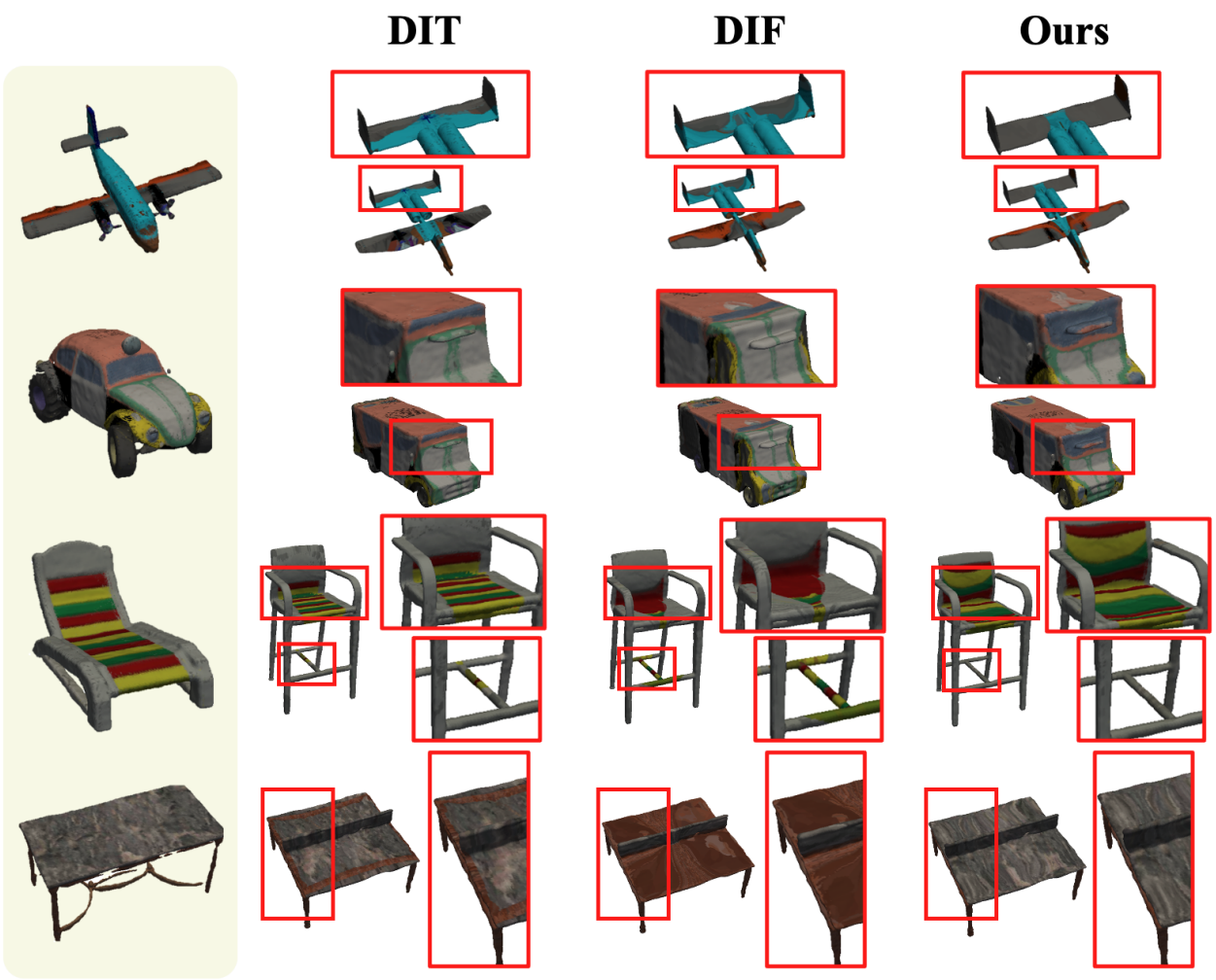

Figure 6. **Additional comparison on texture transfer ShapeNetV2.**

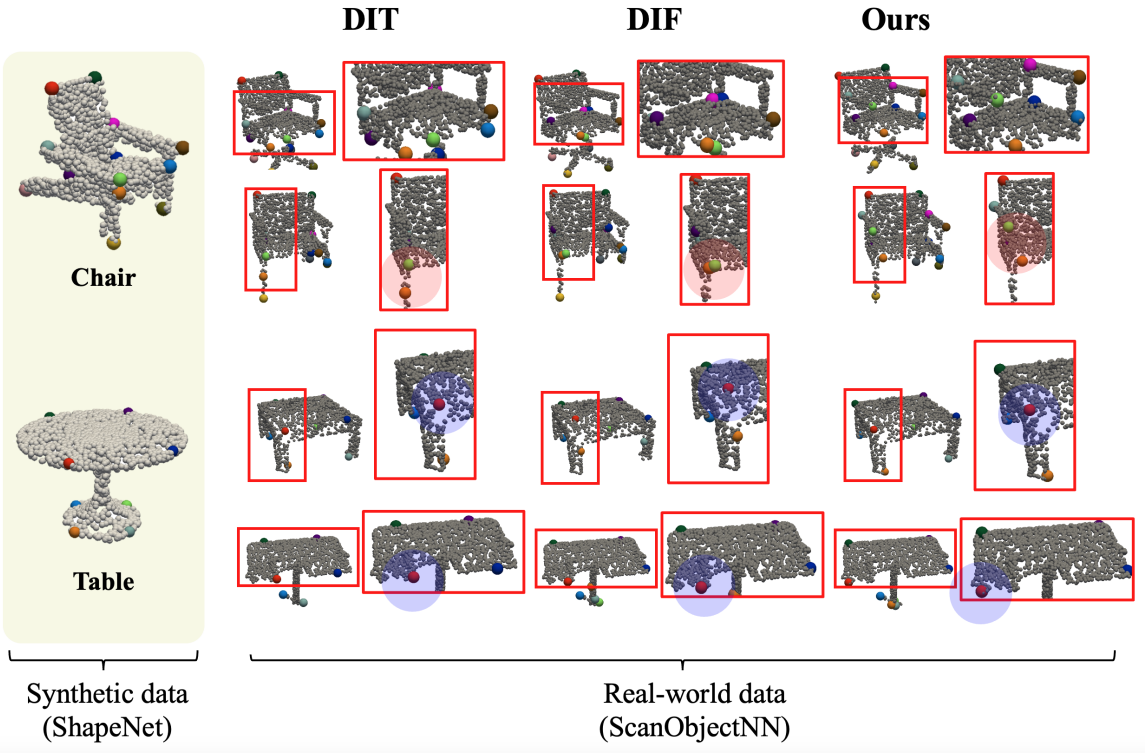

Figure 7. **Comparison on keypoint transfer in ScanObjectNN [2].** We transfer keypoint from synthetic data (ShapeNetV2) to real-world scanned data (ScanObjectNN) to validate the robustness towards the domain gap (zoom-in for better visualization).

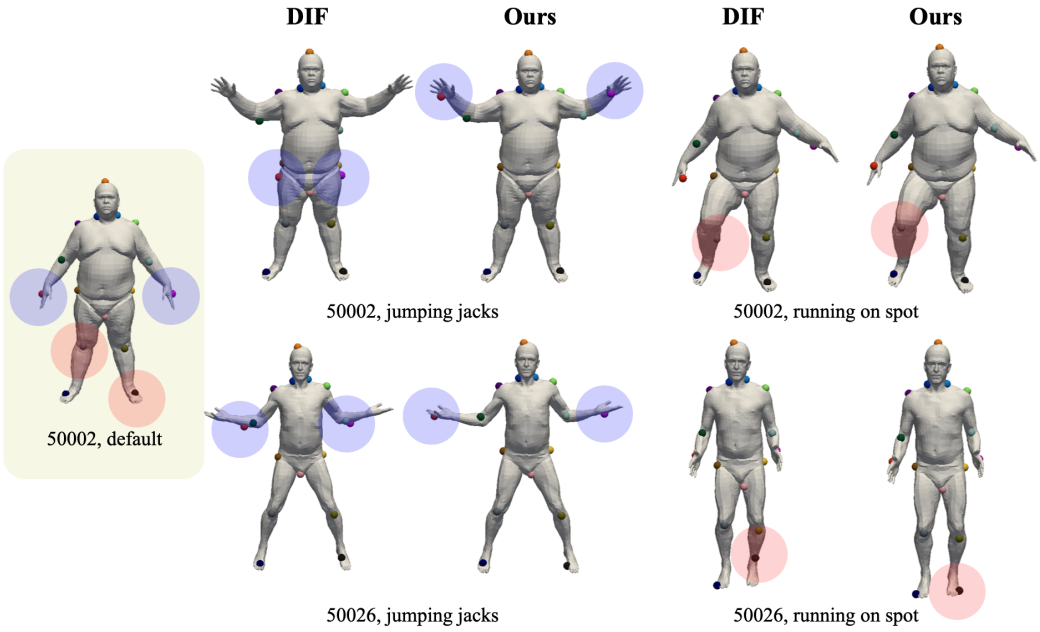

Figure 8. **Comparison on keypoint transfer in DFAUST [3].** Blue balls indicate keypoints for both hands, while red balls indicate keypoints for the right knee and the left foot (zoom-in for better visualization).

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv:1512.03012*, 2015. 1, 3

[2] 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. 1, 3, 7

[3] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *CVPR*, 2017. 1, 3, 4, 7

[4] Li Yi, Vladimir G. Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas J. Guibas. A scalable active framework for region annotation in 3d shape collections. *TOG*, 2016. 1

[5] 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. 1

[6] Yu Deng, Jiaolong Yang, and Xin Tong. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *CVPR*, 2021. 1, 2

[7] Zhiqin Chen, Kangxue Yin, Matthew Fisher, Siddhartha Chaudhuri, and Hao Zhang. Bae-net: Branched autoencoder for shape co-segmentation. In *ICCV*, 2019. 1, 4

[8] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. 2019. 1

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[10] 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. 2

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 2

[12] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 2

[13] Chengjie Niu, Manyi Li, Kai Xu, and Hao Zhang. Rim-net: Recursive implicit fields for unsupervised learning of hierarchical shape structures. In *CVPR*, 2022. 2, 4

[14] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 4