Deformation-based In-Context Learning for Point Cloud Understanding

Supplementary Material

Overview of the Supplementary Material

To complement the main paper, this supplementary material offers expanded implementation details to facilitate reproducibility, together with additional experiments and analyses, structured as follows:

- Sec. A provides detailed descriptions of the model architectures, parameter settings, and fairness considerations in comparisons.
- Sec. B explains how the ModelNet40 and ScanObjectNN In-Context datasets are built.
- Sec. C reports additional evaluations using F-score and EMD to further demonstrate the advantages of DeformPIC.
- Sec. D presents additional analysis about task encoding mechanism.
- Sec. E presents qualitative comparisons between DeformPIC and PIC baselines across multiple tasks.
- Sec. F discusses the method's current limitations and outlines future directions.

A. More Implementation Details

Model Architecture. We adopt 4 transformer blocks in the DEN and 8 transformer blocks with AdaLN-Zero modulation [31] in the DTN by default. Both networks are configured with a hidden dimension of 384 and 6 attention heads. We apply a drop path rate of 0.1 for DTN to enhance regularization, while DEN does not employ drop path.

Model Comparison. We report detailed model comparisons in Table 4. To ensure fair comparisons with baseline models, we control the model capacity by matching the total parameter count rather than the layer depth. Specifically, our default configuration, DeformPIC-*8d*, has a comparable parameter scale to PIC-Cat and PIC-Sep (29.84M vs. 28.91M) and therefore reflects a fair capacity comparison. For completeness, we also implement a deeper variant, DeformPIC-*12d*, whose number of layers aligns with PIC-Cat and PIC-Sep but with higher parameters introduced by AdaLN-Zero modulation. Both variants are trained under identical settings, and their results exhibit consistent performance trends.

B. More Details on Dataset Construction

In this section, we present the construction details of ModelNet40 In-Context and ScanObjectNN In-Context, encompassing three types of tasks (reconstruction, denoising, and registration). For each sample in the original dataset, we first standardize the point cloud by translating it to have zero-mean XYZ coordinates and scaling it to a unit sphere. We then generate five levels of "disruption", corresponding to increasingly challenging transformations.

- **Reconstruction.** We downsample the original 1,024-point cloud to $\{512, 256, 128, 64, 32\}$ points, producing inputs of varying sparsity while keeping the original point cloud as the target..
- **Denoising.** We corrupt the point cloud by replacing $\{100, 200, 300, 400, 500\}$ points with Gaussian noise sampled from $\mathcal{N}(0, I)$.
- **Registration** We randomly rotate the original point cloud within angle ranges of $\{\pm 20°, \pm 40°, \pm 60°, \pm 80°, \pm 100°\}$.

The full set of perturbed samples across all tasks and disruption levels constitutes our ModelNet40 In-Context and ScanObjectNN In-Context benchmarks.

C. More Results

To verify that our conclusions are not biased by the use of a single distance metric, we additionally report results under F-score@τ [36] and Earth Mover's Distance (EMD) on representative settings of the three in-context tasks (Reconstruction, Denoising, and Registration). The F-score@τ metric evaluates both the precision and recall of the predicted point cloud within a distance threshold τ from the ground truth.

$$\text{F-score@}\tau = 2 \times \frac{P(\tau) \times R(\tau)}{P(\tau) + R(\tau)}, \tag{12}$$

where the precision $P(\tau)$ and recall $R(\tau)$ are given by:

$$P(\tau) = \frac{|\{p \in \hat{\mathcal{P}} \mid \min_{g \in \mathcal{P}} \|p - g\| < \tau\}|}{|\hat{\mathcal{P}}|}, \tag{13}$$

$$R(\tau) = \frac{|\{g \in \mathcal{P} \mid \min_{p \in \hat{\mathcal{P}}} \|g - p\| < \tau\}|}{|\mathcal{P}|}. \tag{14}$$

Here, $\hat{\mathcal{P}}$ and $\mathcal{P}$ denote the predicted and ground-truth point sets, respectively. Intuitively, precision measures how many predicted points lie close to the real surface (i.e., accuracy of generated geometry), whereas recall measures how well the predicted point cloud covers the true surface (i.e., completeness). Thus, F-score@τ thus provides a balanced assessment of both qualities: higher values indicate more accurate and complete reconstructions.

The EMD, also known as the Wasserstein distance, measures the minimum cost of transporting the predicted point distribution to exactly match the ground-truth distribution.

$$\text{EMD}(\hat{\mathcal{P}}, \mathcal{P}) = \min_{\phi: \hat{\mathcal{P}} \to \mathcal{P}} \frac{1}{|\hat{\mathcal{P}}|} \sum_{p \in \hat{\mathcal{P}}} \|p - \phi(p)\|_2, \tag{15}$$

Table 4. Comparison of model variants in terms of computational complexity, and in-context performance in ShapeNet In-Context [10].

Models	Flops(G)	Param.(M)	Reconstruction CD ↓						Denoising CD ↓						Registration CD ↓					
			L1	L2	L3	L4	L5	Avg.	L1	L2	L3	L4	L5	Avg.	L1	L2	L3	L4	L5	Avg.
PIC-Cat [10]	11.31	28.91	3.2	3.6	4.6	4.9	5.5	4.3	3.9	4.6	5.3	6.0	6.8	5.3	10.0	11.4	13.8	16.9	18.6	14.1
PIC-Sep [10]	8.14	28.91	4.7	4.3	4.3	4.4	5.7	4.7	6.3	7.2	7.9	8.2	8.6	7.6	8.6	9.2	10.2	11.3	12.4	10.3
DeformPIC-*12d*	5.33	40.48	2.1	2.2	2.4	2.7	3.9	2.7	2.8	3.3	3.5	3.7	3.8	3.4	1.8	1.8	1.9	1.9	2.0	1.9
DeformPIC-*8d*	4.87	29.84	2.2	2.3	2.5	2.8	3.9	2.7	2.8	3.3	3.6	3.8	3.9	3.5	1.9	1.9	1.9	2.0	2.1	2.0

Unlike Chamfer Distance, which matches each point independently to its nearest neighbor, EMD enforces a one-to-one global correspondence, thereby capturing both local geometry and global structure consistency.

Table 5. Comparison with state-of-the-art methods in F-score@τ ($\tau = 0.001$) on ShapeNet In-Context [10].

Models	Rec. F-score@τ	Den. F-score@τ	Reg. F-score@τ
PIC-Cat [10]	38.36	33.06	26.98
PIC-Sep [10]	42.78	32.86	41.57
DeformPIC-*8d*	**57.98**	**49.95**	**66.04**

In the F-score metrics (Table 5), DeformPIC-*8d* achieves the best performance on all three tasks (reconstruction, denoising, and registration), surpassing PIC-Cat and PIC-Sep by large margins. Notably, its registration F-score reaches 66.04, which is 39.06 points higher than PIC-Cat and 24.47 higher than PIC-Sep, indicating substantially improved robustness under large geometric perturbations. A similar trend appears in reconstruction and denoising, where DeformPIC-*8d* outperforms baselines by 15–20 points, highlighting its ability to model both global structures and fine-grained local variations.

Table 6. Comparison with state-of-the-art methods in EMD ($\times$1000) on ShapeNet In-Context [10].

Models	Rec. EMD	Den. EMD	Reg. EMD
PIC-Cat [10]	56.30	52.89	75.14
PIC-Sep [10]	60.83	65.27	64.28
DeformPIC-*8d*	**37.01**	**47.32**	**24.05**

In the EMD metrics (Table 6, DeformPIC-*8d* again exhibits strong performance arcoss three tasks, achieving the lowest EMD values (37.01, 47.32 and 24.05, respectively), reflecting more accurate geometric preserving ability. Overall, the results indicate that DeformPIC-*8d* generalizes significantly better in challenging in-context settings, particularly when structural consistency and alignment accuracy matter most.

D. More Analysis

Our DEN aims to extract task features that encode both task semantics (i.e., *what* the task is) and transformation cues (i.e., *how* to transform). To verify this, we construct a *static* DEN variant that is supervised to predict a fixed task ID.

Table 7. Static vs. Dynamic Encoding on DEN.

Encoding Strategies	Rec. CD ↓	Den. CD ↓	Reg.CD ↓
Static (Predict Task ID)	3.2	4.0	9.9
Dynamic (Predict Task Token)	**2.7**	**3.5**	**2.0**

As shown in Table 7, static encoding results in a clear performance drop, suggesting that task-level information alone is insufficient. In contrast, the dynamic DEN captures prompt-conditioned, instance-specific deformation cues, which are crucial for effective geometric transfer.

E. More Visualization

In this section, we provide more qualitative comparisons of our DeformPIC with PIC-Cat [10] on ShapeNet In-Context in Fig. 5, 6, 7, 8. Across all tasks, the results indicate that PIC-Cat [10] frequently exhibits geometric inconsistencies with respect to the query point cloud. In some cases, it even produces noticeable shape distortions. These observations further highlight the strong geometric preservation ability of our DeformPIC model.

F. Discussion

Limitations. Although our method achieves excellent performance on out-of-domain data, it is still premature to consider it a general-purpose model for point cloud understanding, given that its capabilities are limited to four tasks: reconstruction, denoising, registration, and part segmentation. In addition, current point cloud ICL techniques operate only on object level point clouds, and it remains unclear whether they can be effectively extended to indoor or outdoor scene level data.

Future Works. We hypothesize that a key factor underlying the success of ICL in LLMs is the availability of large-scale training data and diverse task types. Motivated by this, our future work will investigate training point cloud ICL models on substantially larger and more varied datasets to move toward truly point cloud understanding generalist. In addition, we will explore applying these models to specific scene level settings, such as embodied perception and autonomous driving.

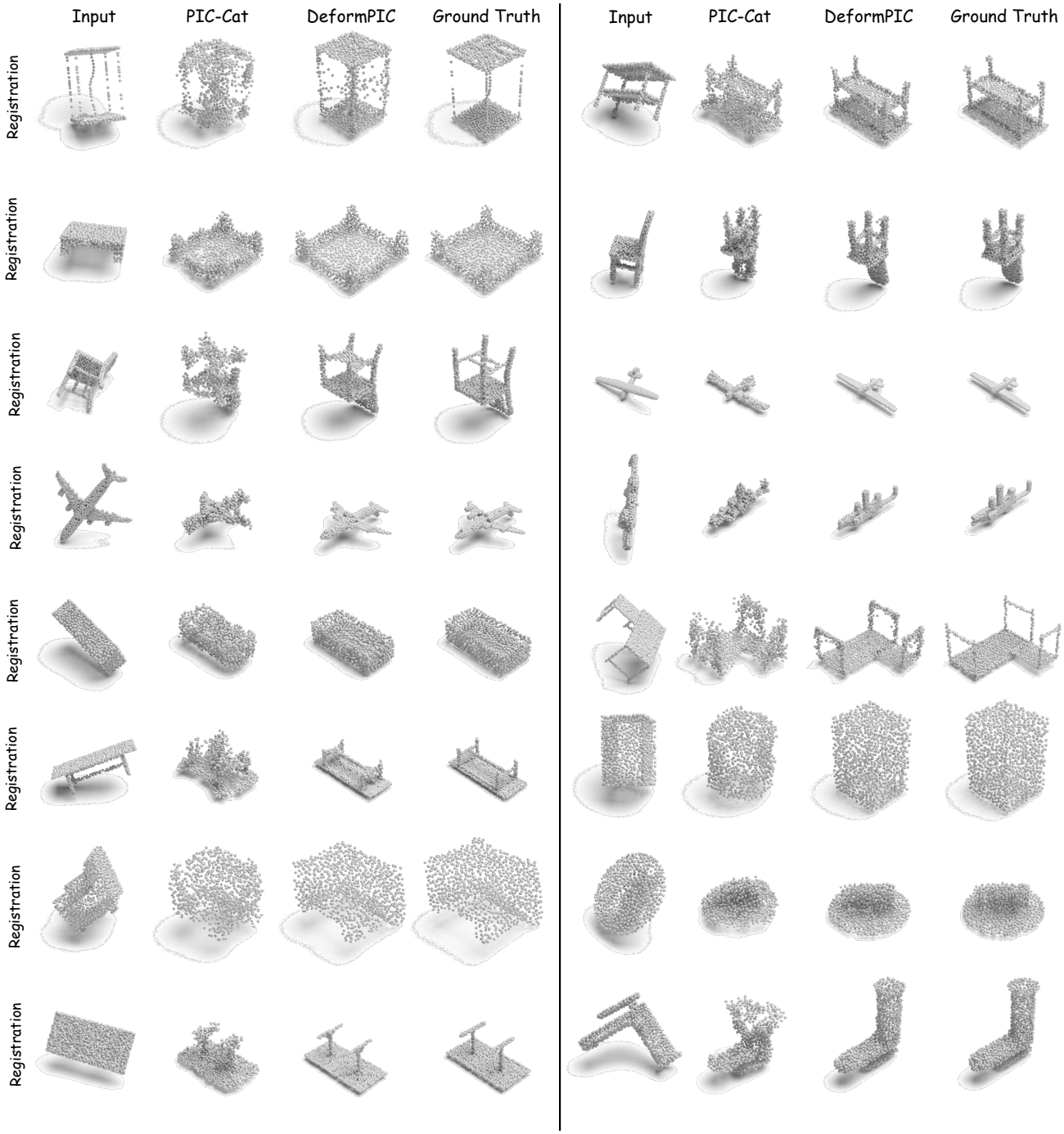

Figure 5. **Visualization results on registration task compared with the PIC [10] and our approach.**

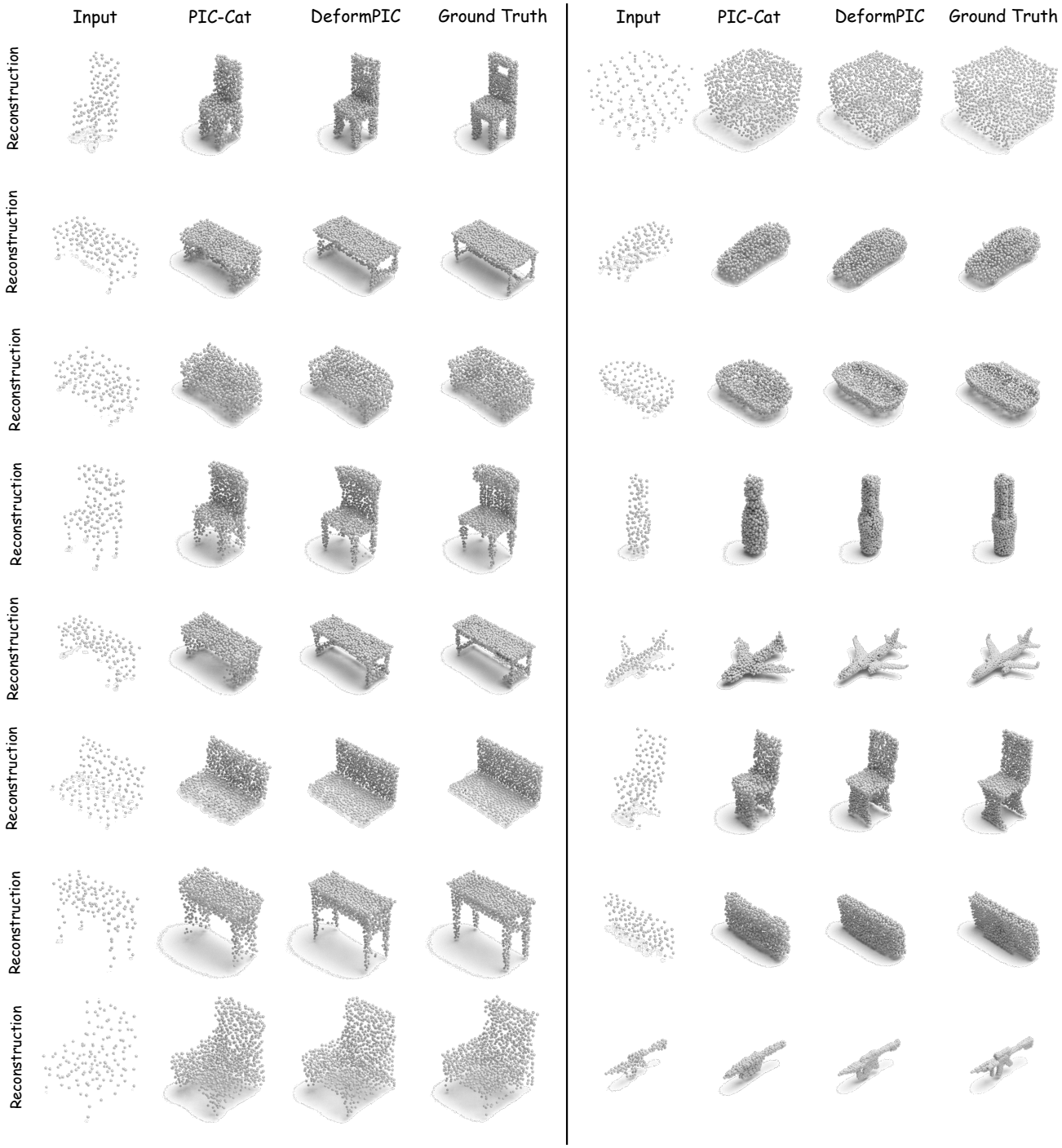

Figure 6. **Visualization results on reconstruction task compared with the PIC [10] and our approach.**

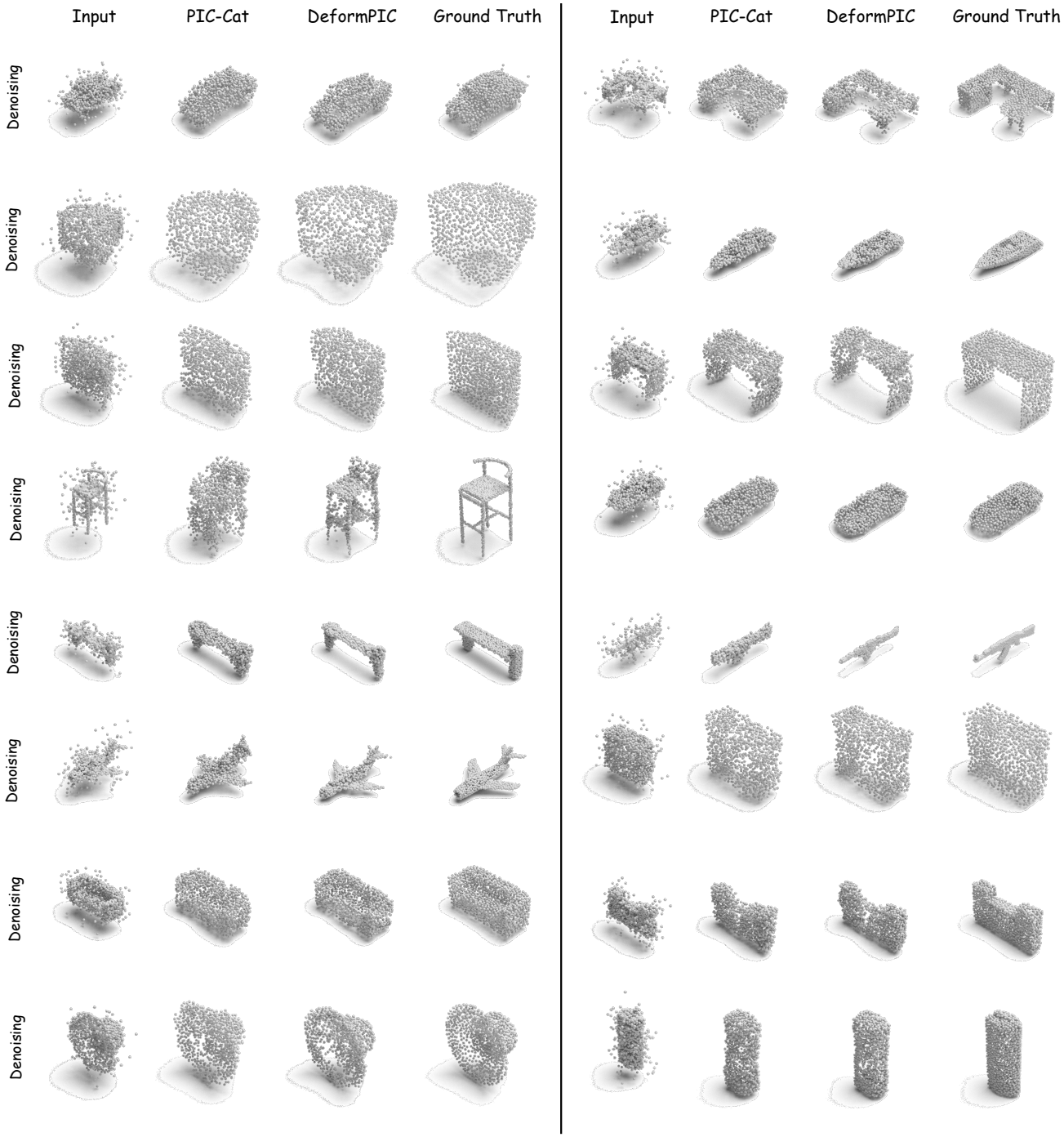

Figure 7. **Visualization results on denoising task compared with the PIC [10] and our approach.**

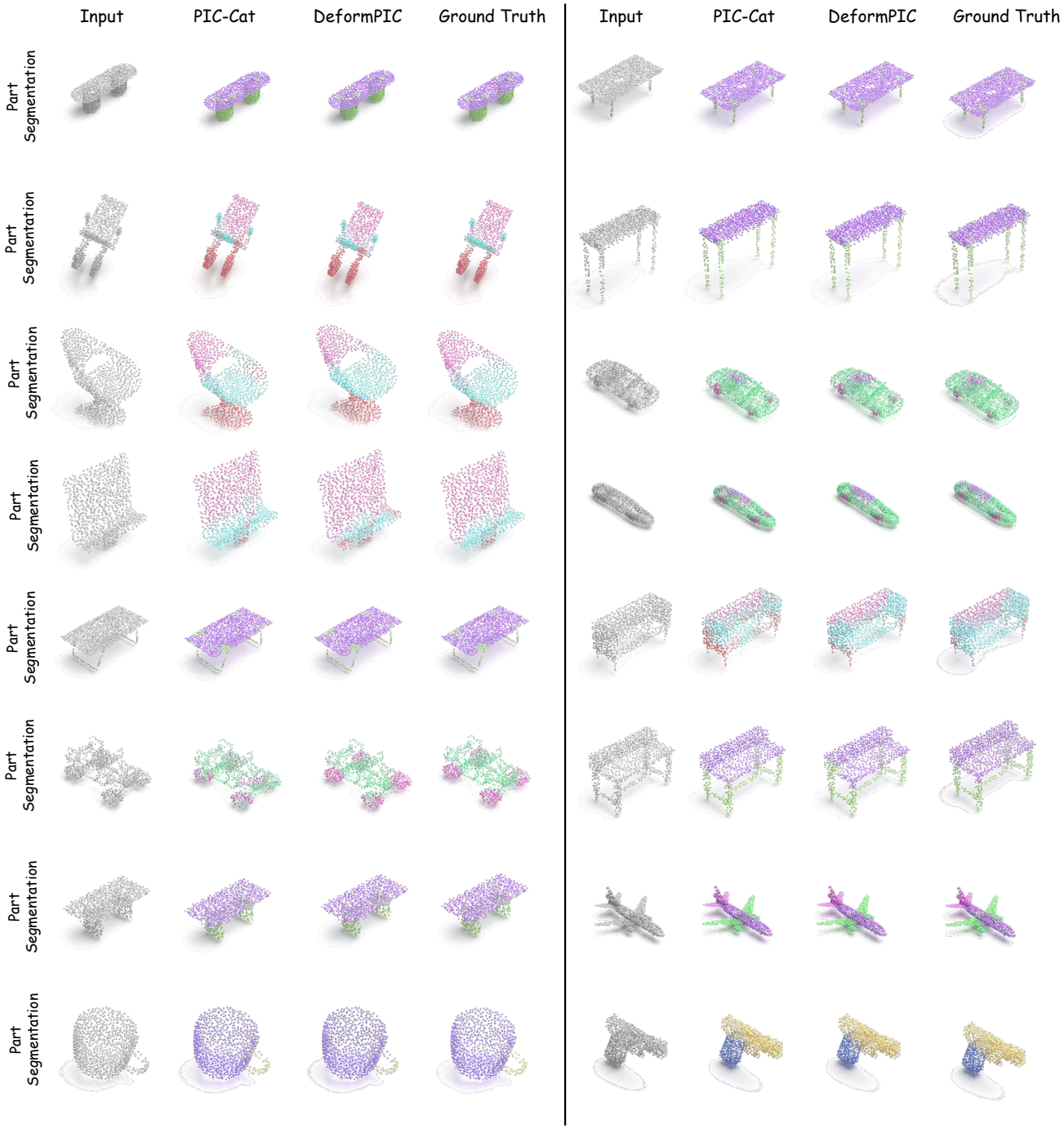

Figure 8. **Visualization results on part segmentation task compared with the PIC [10] and our approach.**