

---

# Attention-based Part Assembly for 3D Volumetric Shape Modeling

## - Supplementary Materials

---

**Chengzhi Wu<sup>1</sup>**   **Junwei Zheng<sup>1</sup>**   **Julius Pfrommer<sup>2,3</sup>**   **Jürgen Beyerer<sup>1,2,3</sup>**

<sup>1</sup>Karlsruhe Institute of Technology, Germany   <sup>2</sup>Fraunhofer IOSB, Germany

<sup>3</sup>Fraunhofer Center for Machine Learning, Germany

## 1 Network Architecture

### 1.1 Encoder and Decoder

The detailed network structures of the encoder and decoder are defined as follows. The accordingly feature layer indices are also given.

Layer	Kernel	Stride	Activation	Channel	Output shape
conv.	$4 \times 4 \times 4$	$2 \times 2 \times 2$	LeakyReLU	64	$16^3$
conv.	$4 \times 4 \times 4$	$2 \times 2 \times 2$	LeakyReLU	128	$8^3$
conv.	$4 \times 4 \times 4$	$2 \times 2 \times 2$	LeakyReLU	256	$4^3$
conv.	$4 \times 4 \times 4$	$1 \times 1 \times 1$	LeakyReLU	256	$1^3$
reshape	-	-	-	1	256

Table 1: Encoder Structure.

Layer	Kernel	Stride	Activation	Channel	Output shape	feature layer index
input	-	-	-	1	256	0
reshape	-	-	-	256	$1^3$	1
deconv.	$4 \times 4 \times 4$	$1 \times 1 \times 1$	LeakyReLU	256	$4^3$	2
deconv.	$4 \times 4 \times 4$	$2 \times 2 \times 2$	LeakyReLU	128	$8^3$	3
deconv.	$4 \times 4 \times 4$	$2 \times 2 \times 2$	LeakyReLU	64	$16^3$	4
deconv.	$4 \times 4 \times 4$	$2 \times 2 \times 2$	Sigmoid	1	$32^3$	5

Table 2: Decoder Structure.

### 1.2 Number of Model Parameters

The information is given in Table 3. Compared to the simple MLP method, our proposed attention-based methods have less parameters. This is a common outcome when high-dimensional fully connected layers are replaced with attention blocks. Additionally, when the feature channel dimension is preserved, the model has even less parameters since the weights in the attention blocks are shared over all feature channels.

	Simple MLP	Part Attention	Channel-wise Part Attention
Trainable parameters	47.61M	46.25M	29.61M

Table 3: Number of parameters in different models.

## 2 Loss Weights for Finetuning

Table 4 gives an ablation study of how we choose our loss weights for different loss terms in step 3. Here, we use a normal part attention model with input feature layers 0/3/5 and apply  $L_{AC}$  as an example. Since the autoencoder is already well trained for part generation in step 1, we find that the change in part mIoU is negligible in the finetuning step. Meanwhile, using a larger  $\omega_{trans}$  results in a smaller transformation MSE, and using a larger  $\omega_{shape}$  results in a better shape mIoU. Balancing the trade-off between all those numerical results and the visualization results, we use loss weights  $\omega_{PI} = 1, \omega_{part} = 1, \omega_{trans} = 10, \omega_{shape} = 10$ , and  $\omega_{AC} = 1$  if it is applied.

Loss weights					Part mIoU					Trans MSE	Shape mIoU
$\omega_{part}$	$\omega_{PI}$	$\omega_{trans}$	$\omega_{AC}$	$\omega_{shape}$	back	seat	leg	armrest	mean		
1	1	1	1	1	71.6%	73.1%	70.1%	61.2%	72.8%	31.9	76.2%
1	1	1	1	10	71.6%	73.2%	70.1%	61.2%	72.8%	31.8	<b>76.5%</b>
1	1	10	1	1	71.6%	73.1%	70.0%	61.2%	72.7%	31.5	76.3%
1	1	10	1	10	71.6%	73.2%	70.1%	61.3%	72.8%	<b>31.4</b>	<b>76.5%</b>
0.1	0.1	1	1	1	71.6%	73.2%	70.1%	61.2%	72.8%	31.6	76.3%

Table 4: The influence of using different loss weights choices. All models are based on an identical step 2 model.

## 3 Training Curves

During the training, models are also evaluated with the test dataset every 10 epochs along the training. In our paper, an ablation study of the input feature layers choice for attention models is presented. In this supplementary material, their full shape mIoU and the transformation matrices MSE curves are given as follows for more insights.

Figure 1 and figure 2 give the metric curves of using different single feature layer as input in a normal part attention model or a channel-wise part attention model, respectively. The curves are in accordance with the conclusions we give in our main paper. Additionally, under the condition of using feature layers 0/3/5 as input, the metric curves of using different network setting choices (attention mode, and the optional  $L_{AC}$ ) are also given in figure 3. Based on those results, we finally choose the normal part attention model with  $L_{AC}$ , and the channel-wise part attention model without  $L_{AC}$  as the main experimental settings.

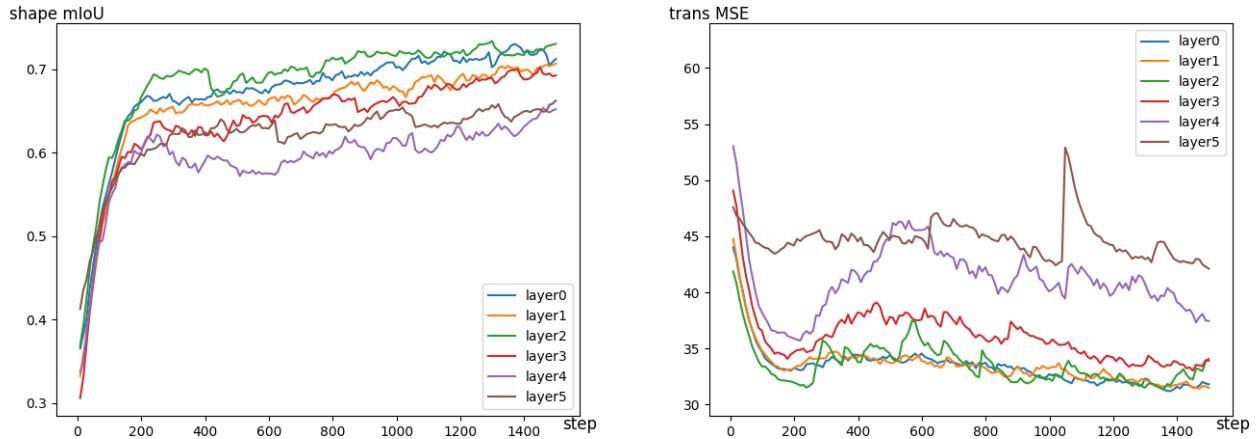


Figure 1: Metric curves of using different single feature layer as input in a normal part attention model.

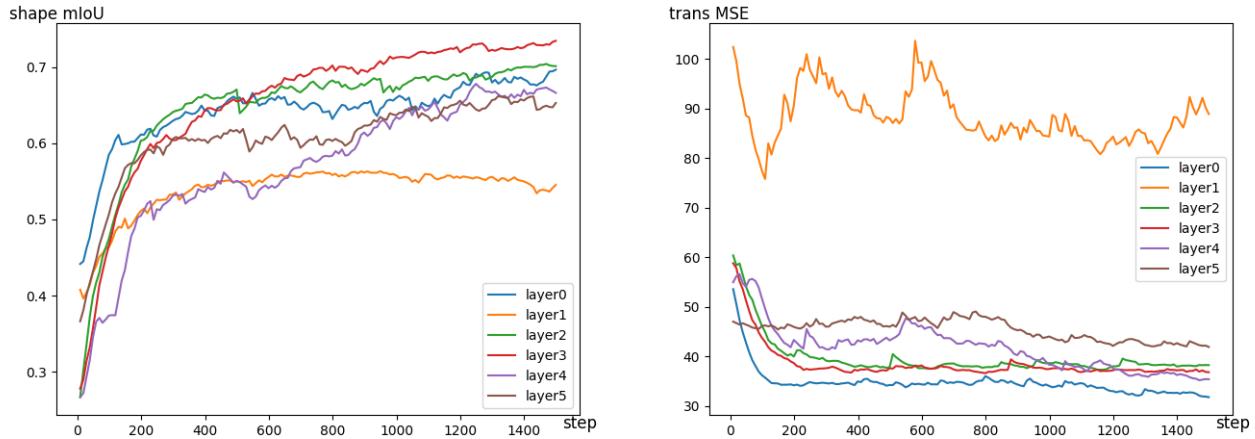


Figure 2: Metric curves of using different single feature layer as input in a channel-wise part attention model.

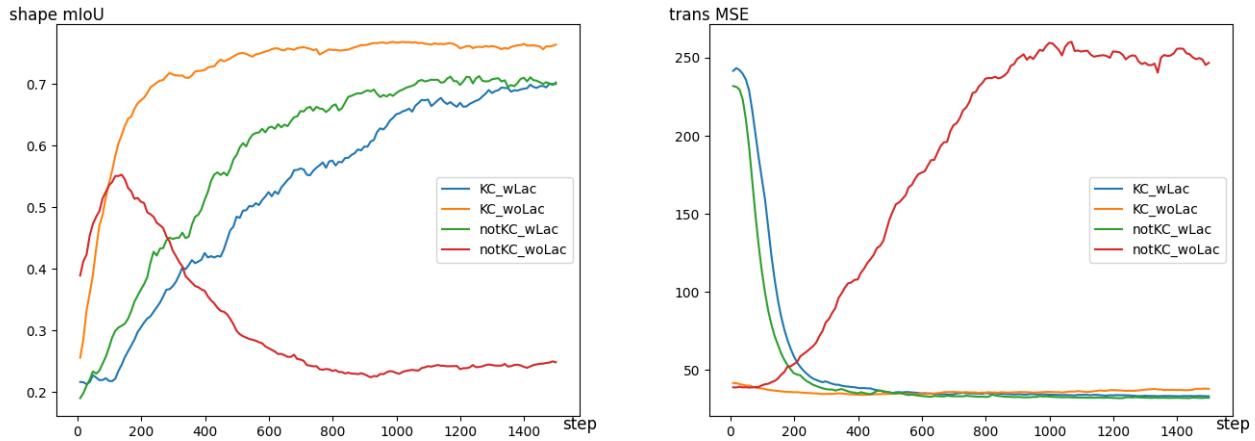


Figure 3: Metric curves when using feature layers 0/3/5 as input. KC stands for using the keeping channel dimension strategy (channel-wise part attention), while notKC stands for the normal part attention mode. wLac stands for  $L_{AC}$  is applied, while woLac stands for  $L_{AC}$  is not applied.

## 4 More Qualitative Results

Results on the chair category are mainly presented in the paper. We hereby present more qualitative results on other categories including airplane, guitar, and lamp, on all kinds of tasks.

### 4.1 Shape Reconstruction

Following the same presentation way as in our paper, Figures 4-7 give some additional reconstruction results on those categories respectively.

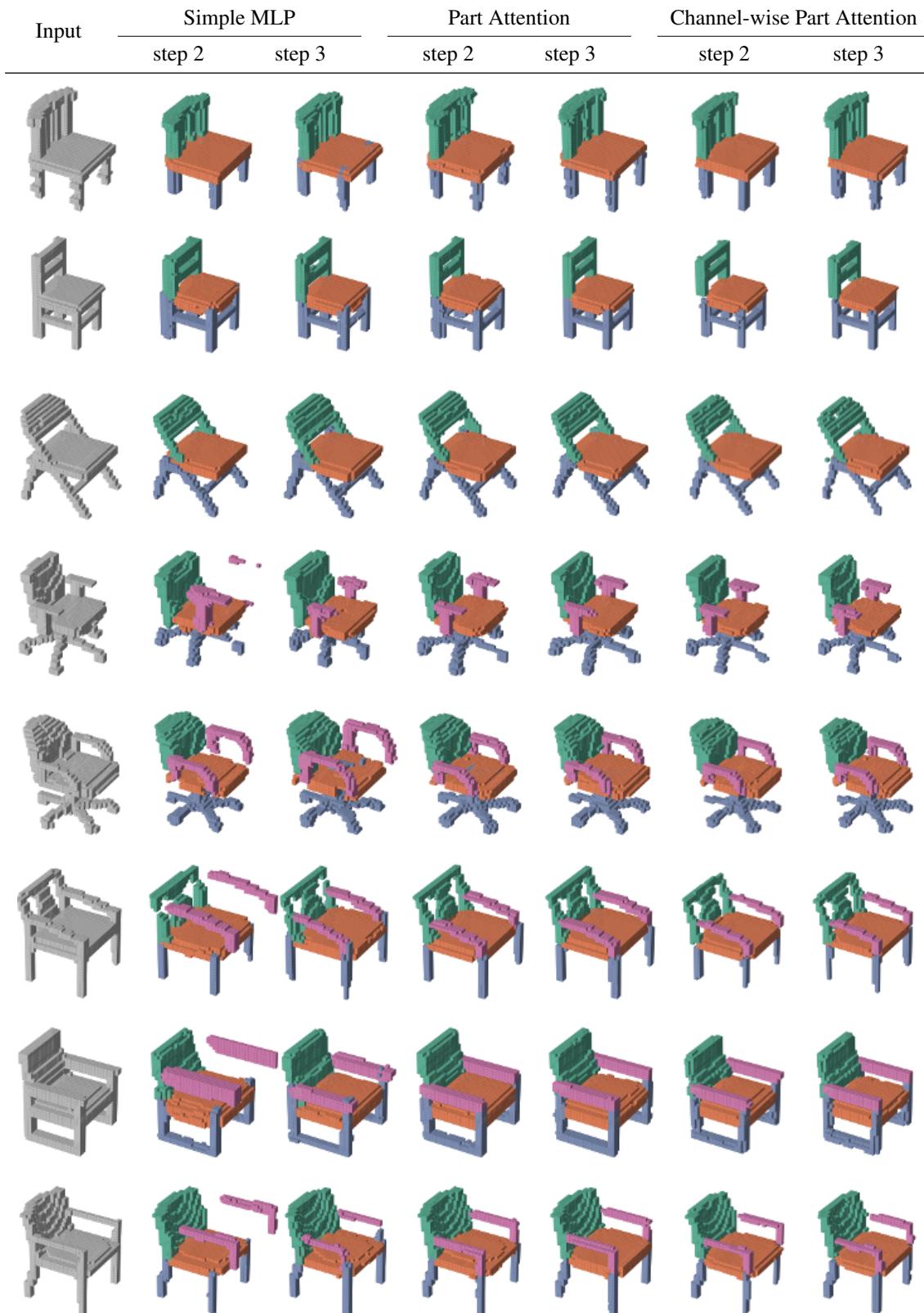
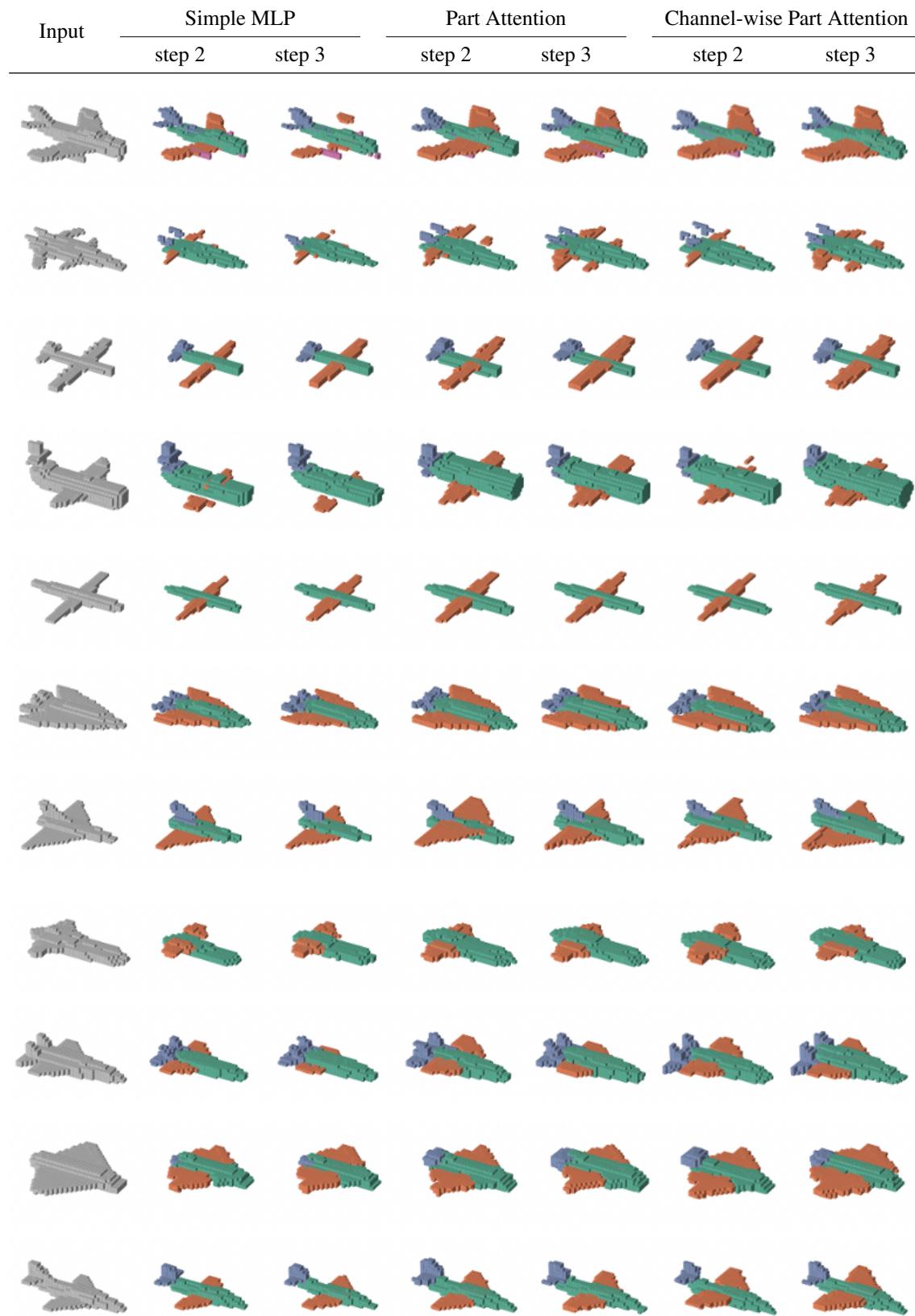


Figure 4: Reconstruction results of our attention-based methods on the chair category in comparison of applying simple dense layers directly. Results before and after the finetuning are both presented.



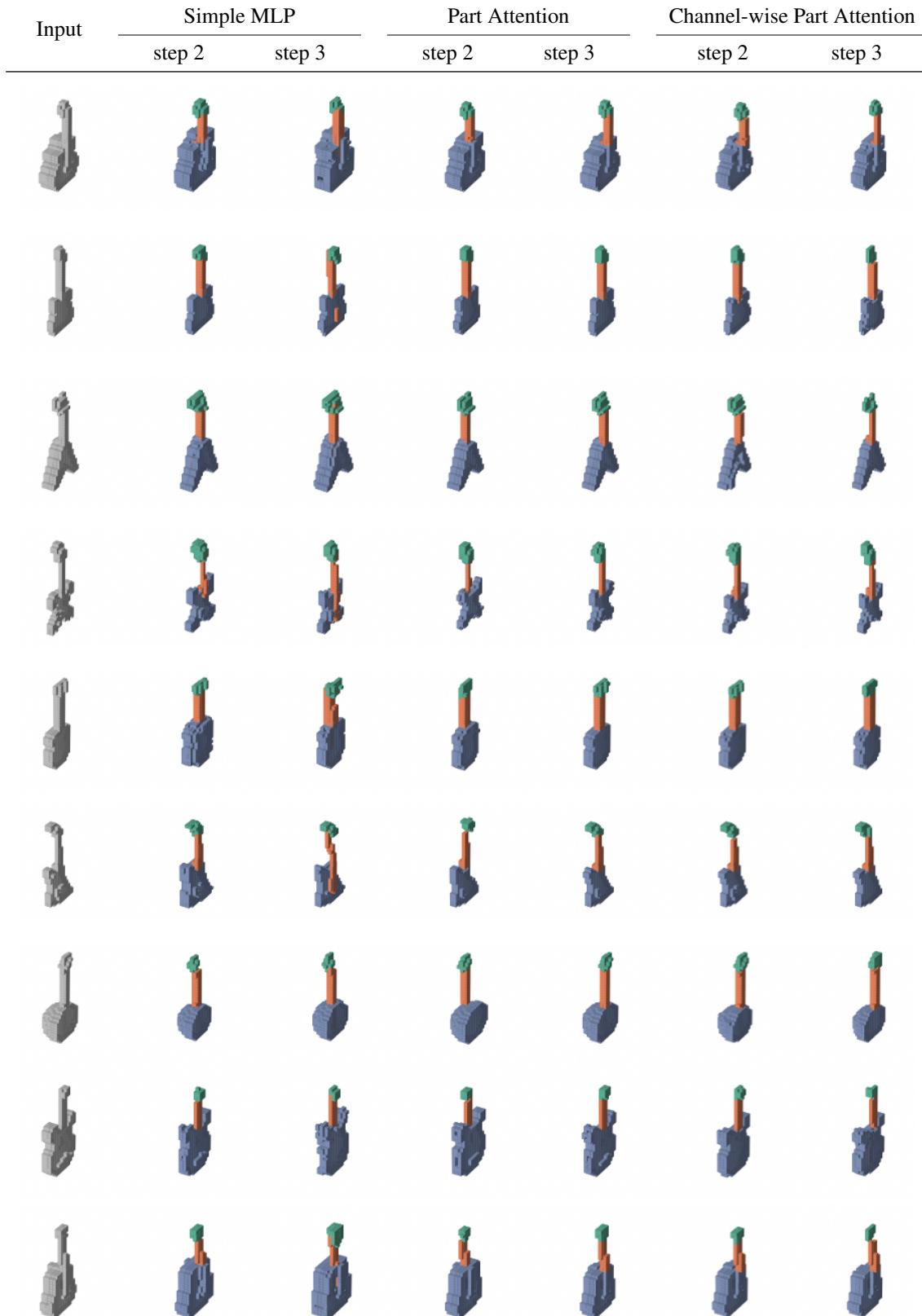


Figure 6: Reconstruction results of our attention-based methods on the guitar category in comparison of applying simple dense layers directly. Results before and after the finetuning are both presented.

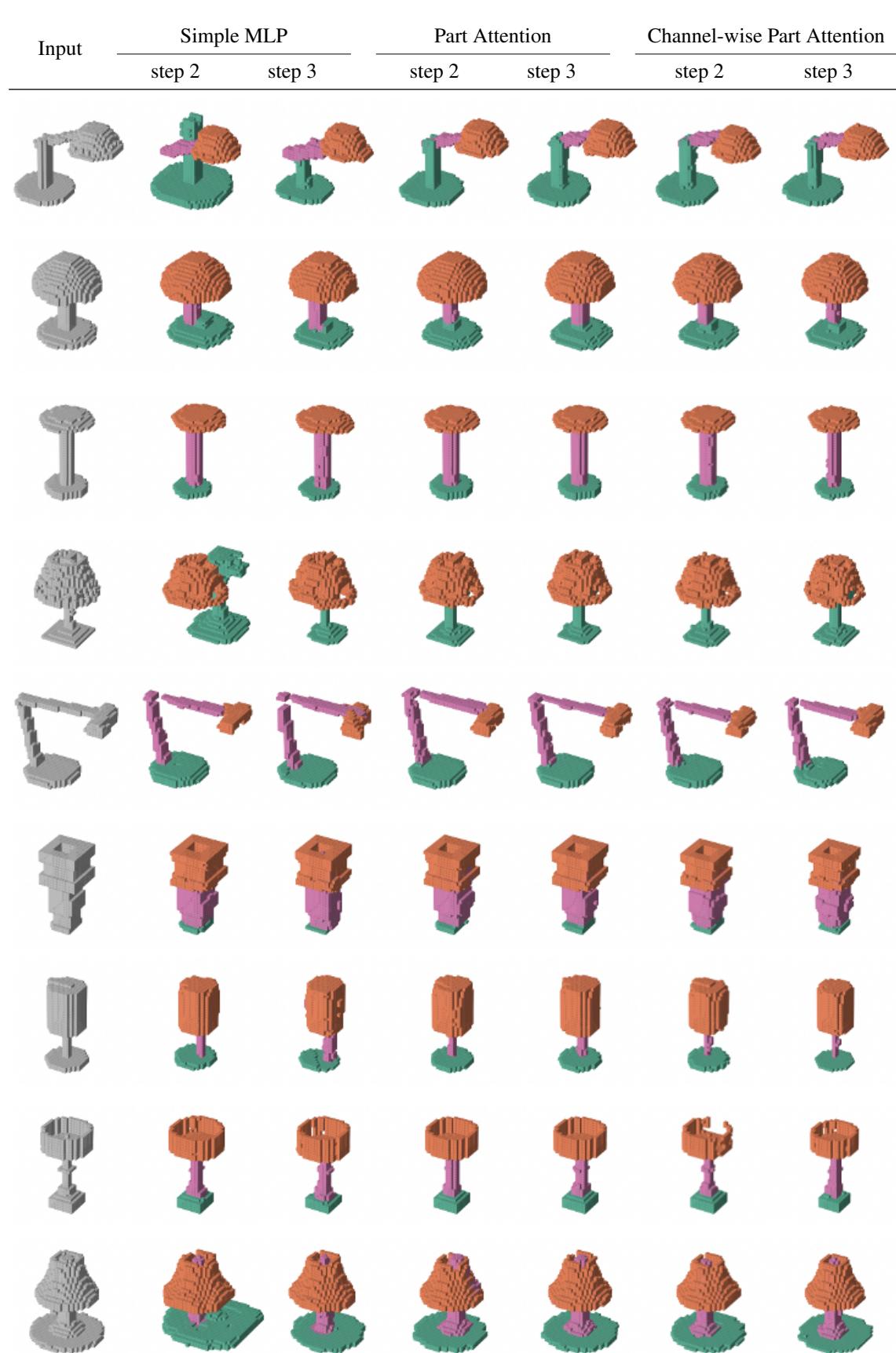


Figure 7: Reconstruction results of our attention-based methods on the lamp category in comparison of applying simple dense layers directly. Results before and after the finetuning are both presented.

## 4.2 Part Swapping

Figure 8 gives some experimental results of swapping shape parts in the latent space. From it, we can observe that compared to the simple MLP method, attention-based method performs much better in scaling and translating all shape parts to a relatively correct place when a part is swapped.

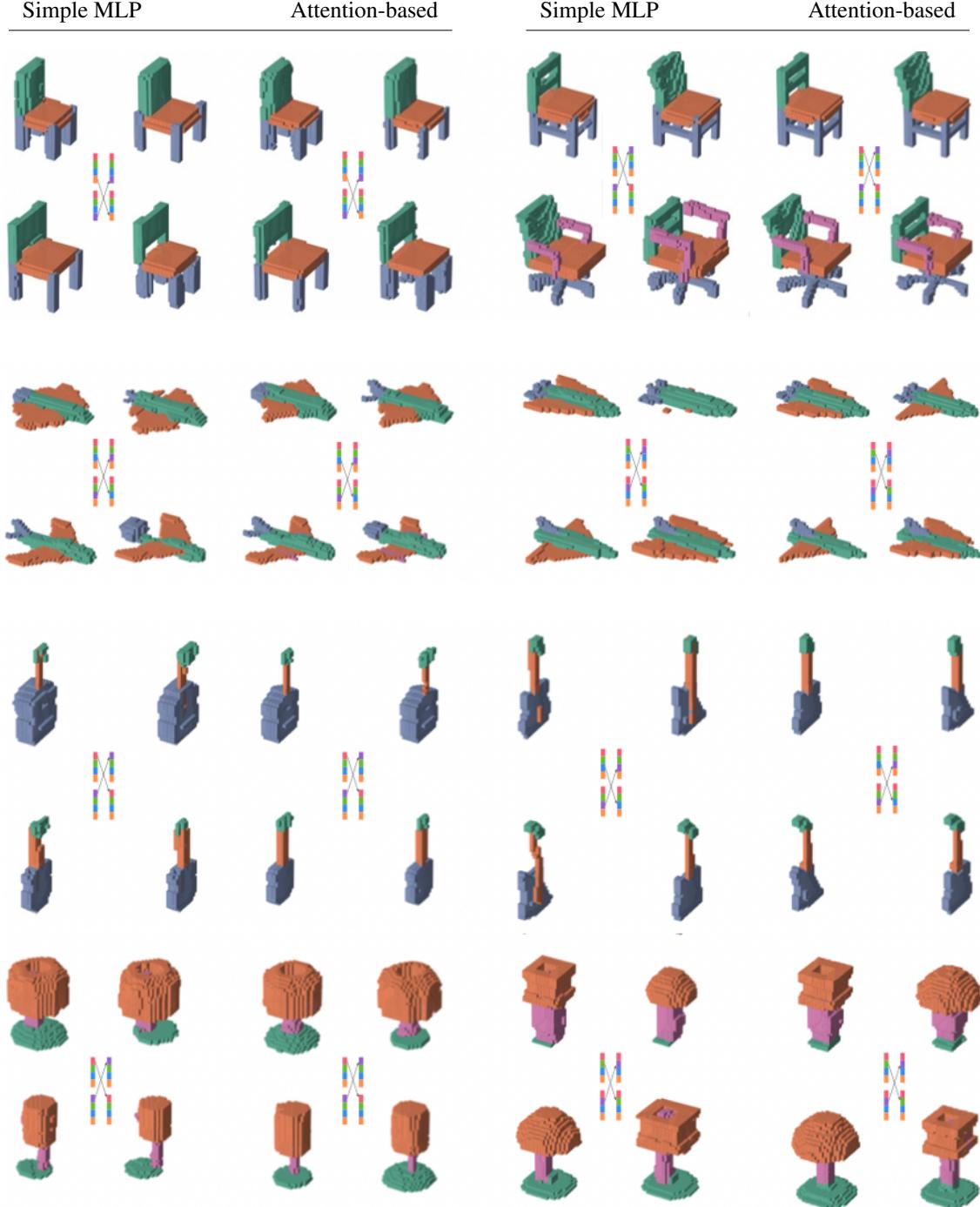


Figure 8: Part swapping results of our attention-based methods in comparison to applying simple dense layers directly on all four categories.

### 4.3 Part Mixing for Random Assembly

Figure 9 gives some part assembly results of using random parts from the shape category to compose new shapes. From the figure we can observe that for the newly composed shapes, the transformation matrices of different parts are learned coherently for the assembly.

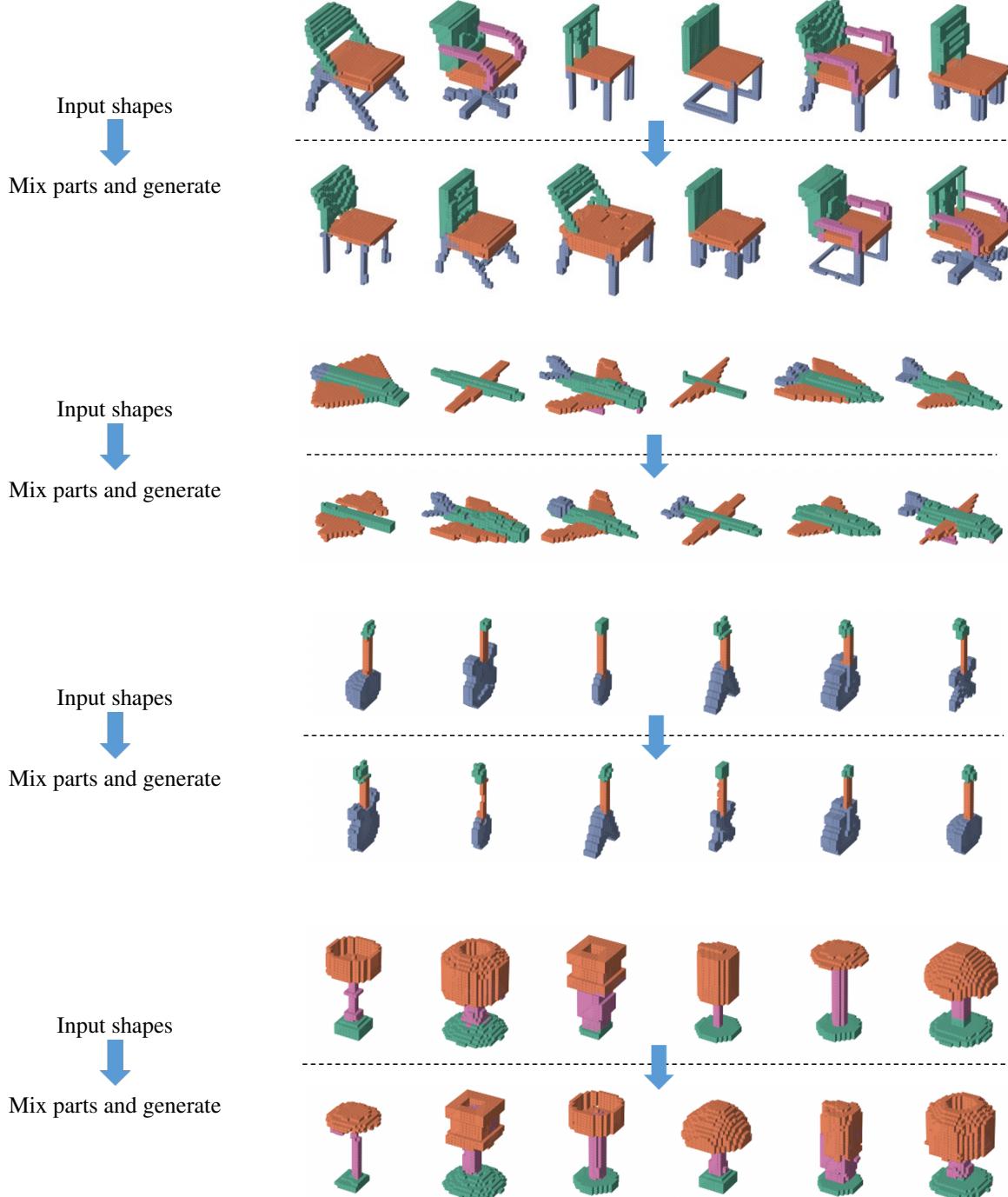


Figure 9: Part assembly results of mixing random parts from different input shapes of one category to generate new shapes.

#### 4.4 Shape Interpolation

Additionally, same as most other 3D shape modeling papers, we also give some shape interpolation results as presented in Figure 10.

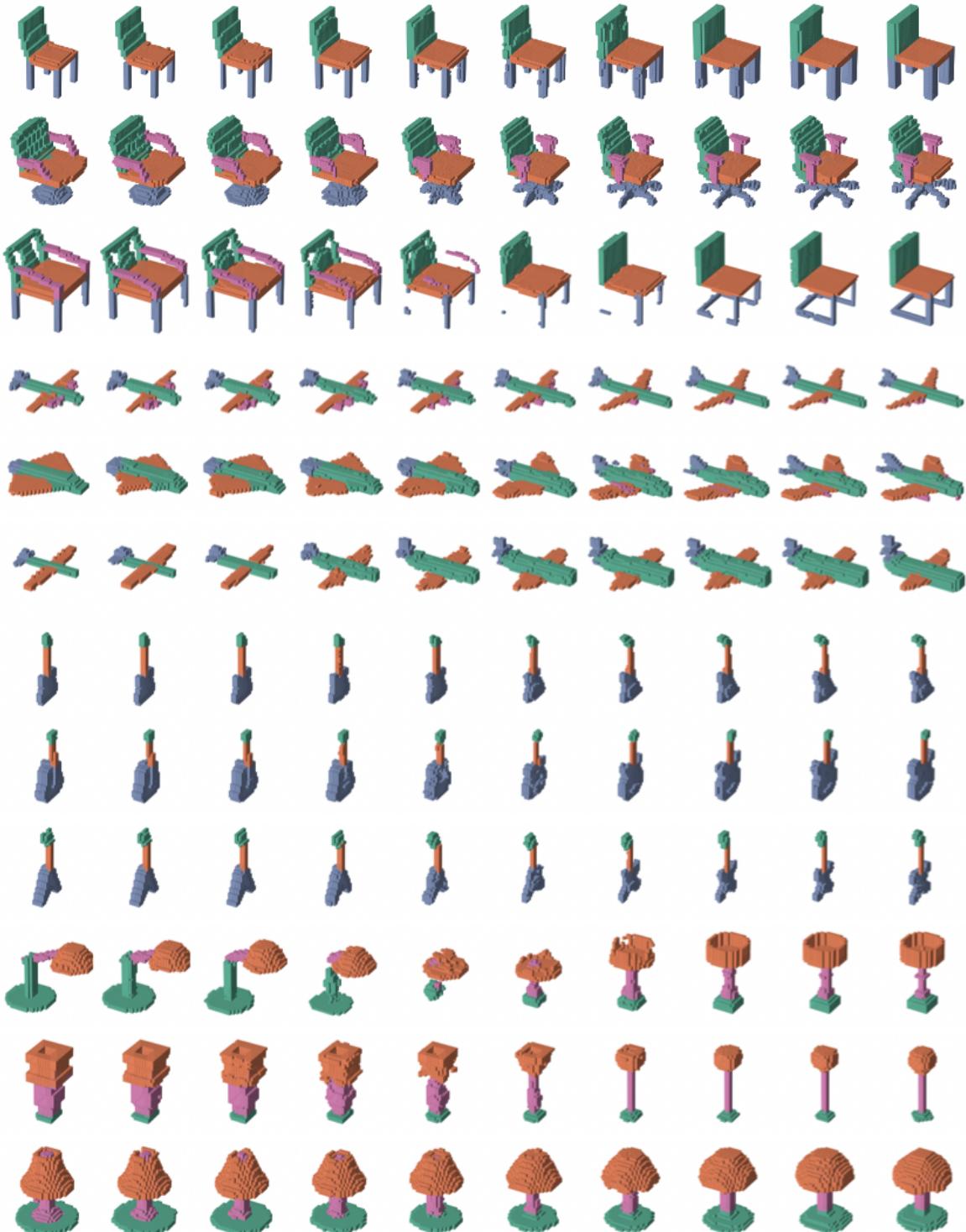


Figure 10: Interpolation results between shape pairs from the same category.

---

## 5 Failure Cases

We show some failure cases in our experiments in Figure 11. For each pair, the left shape is the ground truth, while the right shape is the reconstructed one. It can be seen that for some chairs that have uncommon parts, our model fails to reconstruct the parts correctly. However, our model can still assemble the parts into chair-like objects.

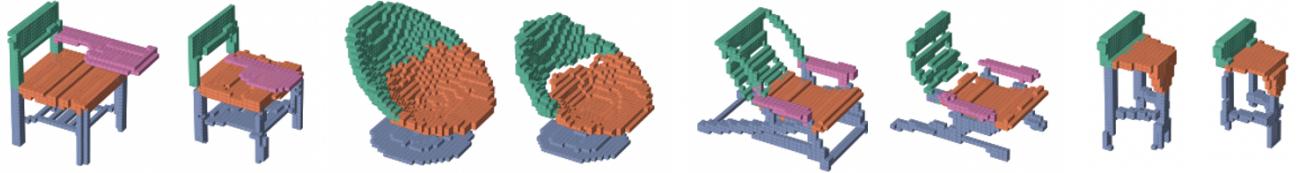


Figure 11: Failure cases.

## 6 Higher Resolution

Training a higher resolution of 3D volumetric data could be really time consuming and needs a lot more computation resources. Hence in our paper, all the presented results are from resolution of  $32^3$ . However, it is surely possible to apply our method to a higher resolution. We hereby show some demo reconstruction results from resolution of  $64^3$  in Figure 12.



Figure 12: Reconstruction results of chair shapes in resolution of  $64^3$ .