Completing 3D Partial Assemblies with View-Consistent 2D-3D Correspondence

Supplementary Material

6. Autoregressive Learning Algorithm

Algorithm 1 Autoregressive Assembly Completion.

```
Input: Partial assembly $A$, image $I$, candidate parts $C$, number of missing parts $k$.
Initialize: Cross-modal encoder $\phi_{enc}$, part decoder $\phi_{dec}$, view-predictor $\phi_{view}$.
$T_v \leftarrow \phi_{view}\ (I)$    // view prediction
$F_A \leftarrow \phi_{enc-pnet}\ (A)$
$F_I \leftarrow \phi_{enc-mlp}\ (I)$
for $i = 1$ to $k$ do
    $D \leftarrow \|S_I - R(T_v(A))\|$    // differential map
    $F_{fuse} \leftarrow \phi_{enc-att}\ (F_A, F_I, D)$    // cross-modal fusion
    $p_i, T_i \leftarrow \phi_{dec}\ (F_{fuse}, C)$    // part retrieval, pose prediction
    $A \leftarrow A \cup T_i(p_i)$    // partial assembly update
    $F_A \leftarrow F_A \cup \phi_{dec-pnet}\ (T_i(p_i))$    // feature update
end for
$\phi_{view} \leftarrow \phi_{view} - \eta \nabla_{\phi_{view}} \mathcal{L}_{proj}$
$\phi_{enc,dec} \leftarrow \phi_{enc,dec} - \beta \nabla_{\phi_{enc,dec}} \mathcal{L}_{cls,pose}$
```

7. Dataset

We use PyTorch3D to render 8 texture-less images by rotating the assembly at 45-degree intervals in the yaw angle while keeping the pitch and roll angles fixed. This process is visualized in Fig. 10.

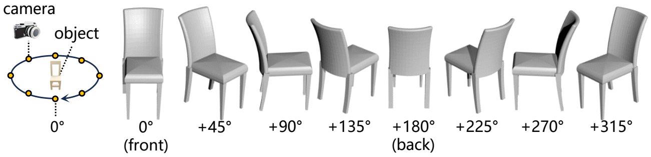

Figure 10. Visualization of the viewpoints and rendered images for the assembly.

8. Metrics

We use Match Accuracy (MA) to evaluate the quality of part retrieval, and Completion Chamfer Distance (CCD) and Part Accuracy (PA) for that of pose prediction. Note that MA and PA are infeasible to compute for generative completion methods (AdaPoinTr and XMFnet), since these methods do not involve concept of parts. Computation of these metrics is defined as follows.

- **Match Accuracy (MA)** is defined as the proportion of parts that are correctly matched with the ground-truth missing parts:

$$\mathcal{MA} = \frac{c}{k}. \tag{8}$$

In this context, a candidate is deemed correct when it exhibits an identical geometry as the ground truth. c denotes the number of correct candidates. Similar to bipartite part matching, we use Hungarian algorithm to compute an optimal permutation of the selected candidates that best matches the ground-truth missing parts and then compute this metric.

- **Completion Chamfer Distance (CCD)** is defined as the Chamfer distance between the completed parts $\mathcal{X}$ and ground-truth missing parts $\mathcal{Y}$:

$$d_c(\mathcal{X}, \mathcal{Y}) = \sum_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} \|x - y\|_2^2 + \sum_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \|x - y\|_2^2. \tag{9}$$

- **Part Accuracy (PA)** indicates the percentage of parts within a certain Chamfer distance threshold and is defined as:

$$\mathcal{PA} = \frac{1}{k} \sum_{i=1}^{k} \left(d_c \left(T_i \left(p_i \right), T_i^* \left(p_i^* \right) \right) < \tau_p \right), \tag{10}$$

where $\tau_p = 0.01$.

9. Baseline Methods

The implementation details of the baseline methods are described as follows.

- **3DPA** assembles a set of parts into a whole based on a reference image. The original method predicts the poses of parts through GNN-based relation modeling of part features, including 2D image features, 2D mask features, local 3D part features, global 3D assembly features, and one-hot instance encoding. This design requires segmented 3D parts. Therefore, we adopt [18] to obtain per-partial-assembly part instance segmentation and extract global 3D assembly features as context for completion, which are concatenated with the features of candidate parts for pose prediction.
- **FiT** assumes the partial assembly is pre-segmented and uses attention mechanism to model the relations between the parts of parital assembly and the candidates. Since no ground-truth part segmentation is provided, we adopt [18] to obtain part segmentation as 3DPA.
- **AdaPoinTr** completes a partial shape by generating the missing region. We adapt this method by setting the missing parts as missing region. Note that PoinTr adopts re-

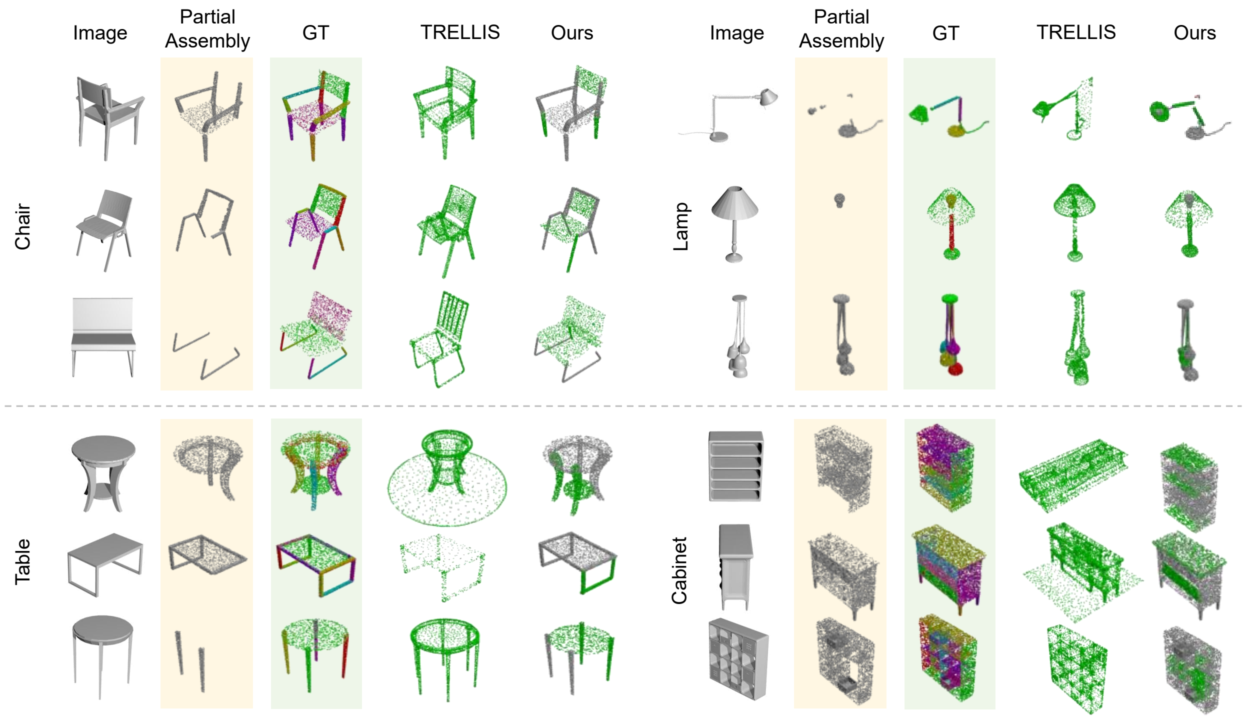

Figure 11. Completion results compared with TRELLIS.

construction as the training task, and the completion result contains both the input partial assembly and missing region. Therefore, we remove the input partial assembly from the completion results by computing the Chamfer distance between the completion result and the partial assembly, and set a distance threshold of 0.01. We employ a dense prediction during inference.

- **XMFnet** and **EGIInet** complete a partial shape with an additional image. We prepare the partial assembly and image in our setting as the input of this method.
- **TRELLIS** is a 3D reconstruction model built on a unified Structured Latent representation. We compare our method with reconstruction results of TRELLIS in Fig. 11. We compare our model ($k = 3$) with the *Large image-to-3D model* of TRELLIS, which is trained with 500K high-quality objects. As demonstrated in the figure, TRELLIS learns strong 3D prior for complicated shape reconstruction, but tends to manifest hallucinations in unobservable regions and geometric detail distortions without guidance of partial shape.

10. More Completion Results

Completion results of $k = 1$**.** Fig. 12 provides more visualization results of our method in the easy mode. Colorful assemblies denote the ground truths. Different colors represent different parts. The completion results are visualized in grad and green, where gray points represent partial assemblies and green points represent missing parts (correctly completed).

Method	Chair→ Table			Chair→ Lamp			Chair→ Cabinet		
(k=1)	PA	MA	CCD	PA	MA	CCD	PA	MA	CCD
3DPA	2.31	43.07	3.39	0.93	30.02	6.93	0.70	20.28	1.47
FiT	6.67	39.74	3.66	2.34	32.55	**5.33**	0.69	30.77	1.34
Ours	**29.46**	**49.59**	**2.29**	**3.50**	**35.60**	5.42	**7.19**	**34.27**	**1.31**

Table 4. Unseen-category generalization. We evaluate the chair model trained with $k = 1$ on the other categories, compared with assembly-based methods.

Comparison with single-view generation method. Fig. 13 provides more completion results of chairs from IKEA-Manual. Annotations of chairs from the IKEA-manual dataset are more coarse-grained than that of the PartNet dataset. For example, the backrest of chair is usually integrated in the IKEA-manual dataset, resulting in inaccuracies in completing such parts.

Generation to novel categories. To further evaluate the image-guided generalization ability of our method, we test our model trained on unseen categories. The results are summarized in Tab. 4. Each model is trained on category A and tested on category B with $k = 1$. Our method outperforms baseline methods, especially on part accuracy.

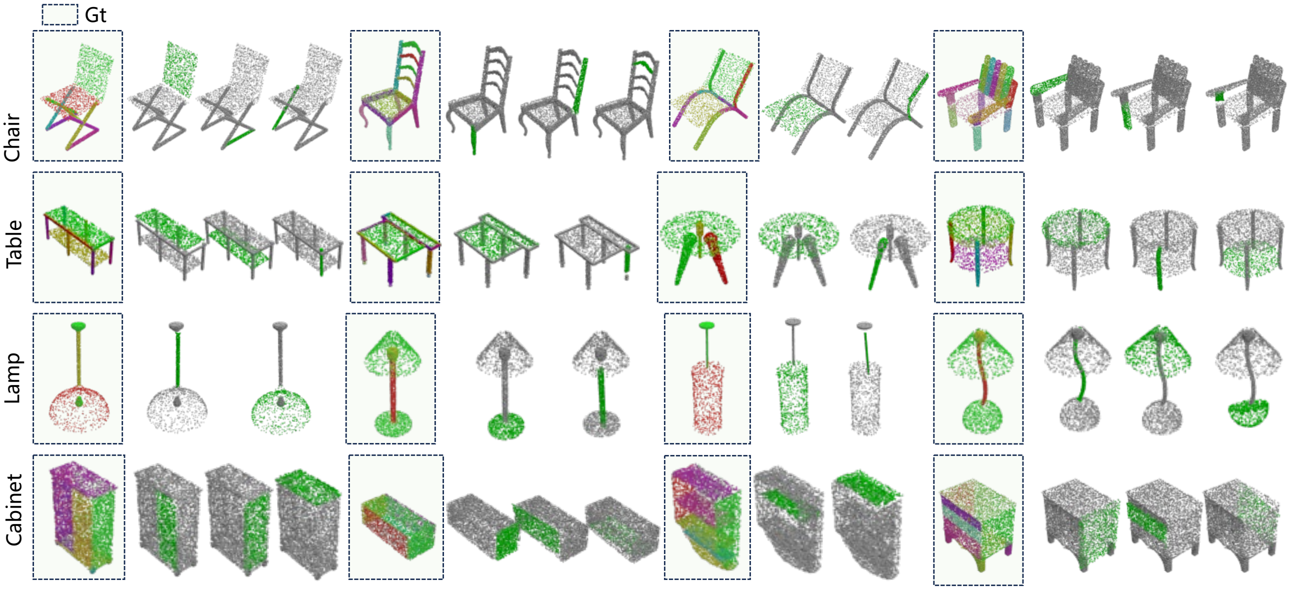

Figure 12. Visualization of completion results in easy mode ($k = 1$).

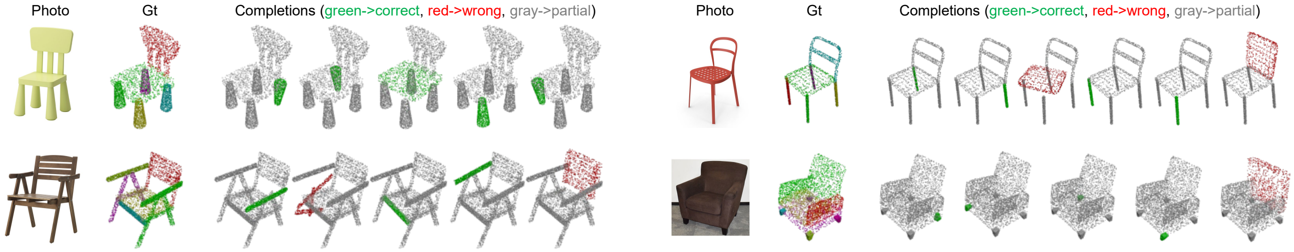
Photo
Gt
Completions (green->correct, red->wrong, gray->partial)
Photo
Gt
Completions (green->correct, red->wrong, gray->partial)

Figure 13. Additional completion results of chairs from IKEA-Manual.