Manual-PA: Learning 3D Part Assembly from Instruction Diagrams

Jiahao Zhang[1] Anoop Cherian[2] Cristian Rodriguez[3] Weijian Deng[1] Stephen Gould[1]

[1]The Australian National University [2]Mitsubishi Electric Research Labs [3]The Australian Institute for Machine Learning

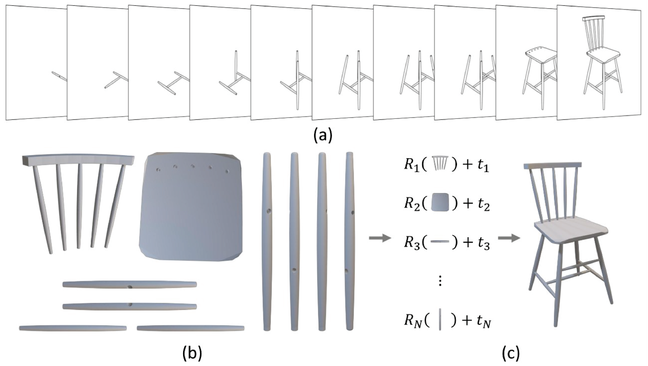
[1]{first.last}@anu.edu.au [2]cherian@merl.com [3]crodriguezop@gmail.com

Code & Datasets

ICCV OCT 19-23, 2025 HONOLULU HAWAII

1. Introduction

▶ Problem Statement



Given **(a)** a **diagrammatic** manual book demonstrating the step-by-step assembly process and **(b)** a set of **texture-less** furniture parts, the goal is to **(c)** infer the order of parts for the assembly from the manual sequence and predict the **6DoF pose** for each part such that the spatially transformed parts assembles the furniture described in the manual.

2. Challenges & Contributions

▶ Key Challenges

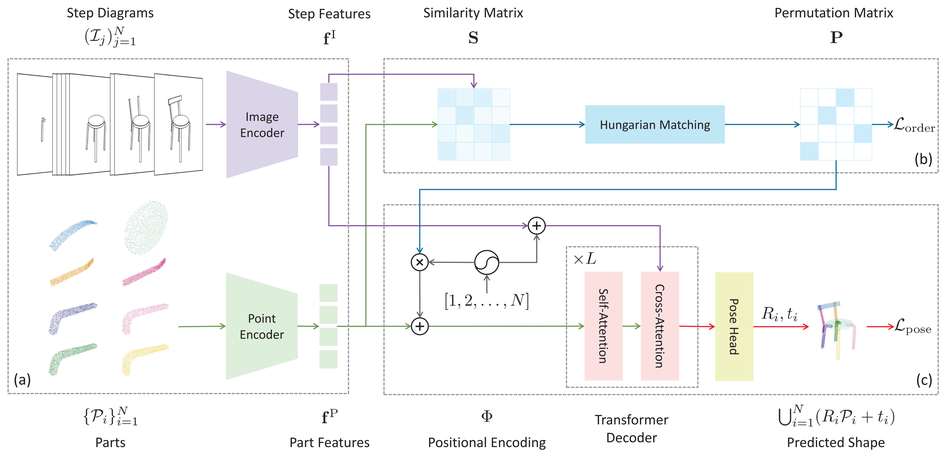
1. **Cross-Modal Alignment:** How to match 2D diagrams with 3D parts to infer the assembly sequence?
2. **Guided Pose Estimation:** How to leverage the inferred sequence as guidance without accumulating errors?

▶ Contributions

1. **A Novel Task:** We introduce and formulate the new task of manual-guided 3D part assembly.
2. **A Novel Framework (Manual-PA):** We propose an end-to-end Transformer-based model that learns the assembly order via contrastive learning and uses it as soft guidance for robust pose estimation.
3. **SOTA Results:** Our method significantly outperforms prior works on the PartNet benchmark and shows strong zero-shot generalization to the real-world IKEA-Manual dataset.

3. Method - Manual-Guided Part Assembly

▶ Overall Pipeline



▶ Manual-Guided Part Permutation Learning

1. Similarity

$$\sigma = \operatorname{argmax}(\mathbf{P})$$

2. Hungarian Matching

$$\begin{aligned} \text{minimize} \quad & \textstyle\sum_{i=1}^N \sum_{j=1}^N \mathbf{C}_{ij}\mathbf{P}_{ij} \\ \text{subject to} \quad & \textstyle\sum_{j=1}^N P_{ij} = 1, \quad \forall i \in \{1, \ldots, N\} \\ & \textstyle\sum_{i=1}^N P_{ij} = 1, \quad \forall j \in \{1, \ldots, N\} \\ & P_{ij} \in \{0, 1\}, \quad \forall i, j \in \{1, \ldots, N\} \end{aligned}$$

3. Order

$$\mathbf{S}_{ij} = \operatorname{sim}(\mathbf{f}_i^{\mathrm{P}}, \mathbf{g}_j^{\mathrm{I}})$$

▶ Manual-Guided Part Pose Estimation

1. Sinusoidal Encoding

$$\phi(x)_{2i} = \sin\left(\frac{x}{\tau_p^{2i/d}}\right), \ \phi(x)_{2i-1} = \cos\left(\frac{x}{\tau_p^{2i/d}}\right)$$

2. Permutation

$$\hat{\mathbf{p}}^{\mathrm{P}} = \hat{\mathbf{P}}^T \Phi$$

▶ Losses

1. Loss for Permutation Learning

(a) Contrastive Loss

$$\mathcal{L}_{\text{order}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\operatorname{sim}(\mathbf{f}^{\mathrm{P}}_{\sigma(i)}, \mathbf{g}^{\mathrm{I}}_i)/\tau)}{\sum_{j=1}^{B} \exp(\operatorname{sim}(\mathbf{f}^{\mathrm{P}}_{\sigma(i)}, \mathbf{g}^{\mathrm{I}}_j)/\tau)}$$

2. Losses for Pose Estimation

(a) Translation Loss (L2)

$$\mathcal{L}_C = \frac{1}{N}\sum_{i=1}^{N} \mathrm{CD}(\hat{R}_{[i]}\mathcal{P}_i, R_i\mathcal{P}_i)$$

(b) Rotation Loss (Chamfer)

$$\mathcal{L}_T = \frac{1}{N}\sum_{i=1}^{N} \|\hat{t}_{[i]} - t_i\|_2$$

(c) Rotation Loss (L2)

$$\mathcal{L}_E = \frac{1}{N}\sum_{i=1}^{N} \|\hat{R}_{[i]}\mathcal{P}_i - R_i\mathcal{P}_i\|_2$$

(d) Shape Loss (Chamfer)

$$\mathcal{L}_S = \mathrm{CD}\left(\bigcup_{i=1}^{N}(\hat{R}_{[i]}\mathcal{P}_i + \hat{t}_{[i]}), \bigcup_{i=1}^{N}(R_i\mathcal{P}_i + t_i)\right)$$

(e) Final Loss

$$\mathcal{L}_{\text{pose}} = \lambda_T\mathcal{L}_T + \lambda_C\mathcal{L}_C + \lambda_E\mathcal{L}_E + \lambda_S\mathcal{L}_S$$

4. Results

Method	Condition	SCD↓		PA↑		SR↑	
		Chair	Table	Chair	Table	Chair	Table
Fully-Supervised on PartNet [26]							
$\text{DGL}_{\text{NIPS'20}}$ [52]	-	9.1	5.0	39.00	49.51	-	-
$\text{IET}_{\text{RA-L'22}}$ [55]	-	5.4	3.5	62.80	61.67	-	-
$\text{Score-PA}_{\text{BMVC'23}}$ [8]	-	7.4	4.5	42.11	51.55	8.320	11.23
$\text{CCS}_{\text{AAAI'24}}$ [56]	-	7.0	-	53.59	-	-	-
$\text{3DHPA}_{\text{CVPR'24}}$ [9]	-	5.1	2.8	64.13	64.83	-	-
$\text{RGL}_{\text{WACV'22}}$ [27]	Sequence	8.7	4.8	49.06	54.16	-	-
$\text{SPAFormer}_{\text{ArXiv'24}}$ [51]	Sequence	6.7	3.8	55.88	64.38	16.40	33.50
$\text{Joint-PA}_{\text{CVPR'24}}$ [24]	Joint	6.0	7.0	72.80	67.40	-	-
$\text{Image-PA}_{\text{ECCV'20}}$ [22]	Image	6.7	3.7	45.40	71.60	-	-
$\text{Image-PA}^{\dagger}_{\text{ECCV'20}}$	Diagram	5.9	3.9	62.67	70.10	19.97	32.83
Manual-PA w/o Order	Manual	3.0	3.6	79.07	74.03	34.13	37.71
Manual-PA (Ours)	Manual	**1.7**	**1.8**	**89.06**	**87.41**	**58.03**	**56.66**
Zero-Shot on IKEA-Manual [47]							
$\text{3DHPA}_{\text{CVPR'24}}$	-	34.3	37.8	1.914	4.027	0.000	0.000
$\text{Image-PA}^{\dagger}_{\text{ECCV'20}}$	Diagram	17.3	14.7	19.07	36.74	0.000	10.53
Manual-PA w/o Order	Manual	12.8	8.9	38.36	42.01	1.754	10.53
Manual-PA (Ours)	Manual	**11.4**	**4.8**	**42.51**	**49.72**	**3.509**	**15.79**