Coarse-to-Fine 3D Part Assembly via Semantic Super-Parts and Symmetry-Aware Pose Estimation

Xinyi Zhang[1]*, **Bingyang Wei**[1]*, **Ruixuan Yu** (✉)[1], **Jian Sun**[2, 3]
[1]School of Airspace Science and Engineering, Shandong University, Weihai 264209, China
[2]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China
[3]Pazhou Laboratory (Huangpu), Guangzhou, Guangdong 510555, China
{zhangxinyi123, weibingyang}@mail.sdu.edu.cn,
yuruixuan@sdu.edu.cn, jiansun@xjtu.edu.cn

Abstract

We propose a novel two-stage framework, Coarse-to-Fine Part Assembly (CFPA), for 3D shape assembly from basic parts. Effective part assembly demands precise local geometric reasoning for accurate component assembly, as well as global structural understanding to ensure semantic coherence and plausible configurations. CFPA addresses this challenge by integrating semantic abstraction and symmetry-aware reasoning into a unified pose prediction process. In the first stage, semantic super-parts are constructed via an optimal transport formulation to capture high-level object structure, which is then propagated to individual parts through a dual-range feature propagation mechanism. The second stage refines part poses via cross-stage feature interaction and instance-level geometric encoding, improving spatial precision and coherence. To enable diverse yet valid assemblies, we introduce a symmetry-aware loss that jointly models both self-symmetry and inter-part geometric similarity, allowing for diverse but structurally consistent assemblies. Extensive experiments on the PartNet benchmark demonstrate that CFPA achieves state-of-the-art performance in assembly accuracy, structural consistency, and diversity across multiple categories. Code is available at https://github.com/zhangxinyi364/CFPA.

1 Introduction

3D part assembly, the task of estimating accurate 6-DoF poses for a set of basic parts to reconstruct a coherent 3D shape, is a fundamental yet challenging problem in robotics, vision, and digital design [1–5]. This task is particularly challenging due to the complex geometric dependencies between parts, the absence of explicit semantic structure, the pervasive presence of symmetries, and the vast number of potential combinations.

Recent advances in 3D part assembly have focused on modeling structural dependencies among parts. Graph-based approaches [6, 7] encode local geometric relationships through message passing on learned part graphs, but their receptive field is often limited and insufficient for capturing high-level object structure. Transformer-based methods [8, 9] improve long-range reasoning by leveraging global self-attention, yet they typically process parts as flat sequences and struggle to encode hierarchical or permutation-invariant semantics. Generative approaches, including VAEs [10], GANs [11], and score-based diffusion models [12], attempt to model joint distributions over object structure and part-level geometry [13–15]. While powerful, they often rely on implicit, handcrafted, or fixed part hierarchies, which can be brittle and category-specific. To address this, we propose to learn high-

*Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

level semantic super-parts via optimal transport, which provide meaningful guidance for subsequent coarse-to-fine pose prediction and support more coherent and semantically aware assembly.

Beyond semantic reasoning, another challenge in 3D part assembly is handling symmetry within individual parts and geometrically similar parts, common in real-world objects like self-symmetric chair seats or repeated structures like chair legs. These symmetries lead to multiple valid configurations that are indistinguishable under standard supervision. Some methods address this by introducing instance encoding, losses, or constraints to discourage geometrically similar parts from occupying equivalent position [8, 9, 16]. Others reduce part-wise pose ambiguity through part pose normalization and further leverage part-level geometric similarity to model part relations [6, 8, 17, 18]. Furthermore, some methods exploit symmetry to reduce complexity during fragment reassembly by aligning parts based on their symmetric relationships [19–21]. However, most existing methods focus on precise part assembly and overlook the interchangeable structural roles of self-symmetric parts or parts with minor geometric differences. In contrast, our approach jointly models both intra-part and inter-part symmetries, enabling the generation of accurate, diverse, and structurally plausible assemblies.

In this paper, we propose **CFPA** (Coarse-to-Fine Part Assembly), a two-stage framework that jointly models semantic structure and geometric symmetry. It first performs coarse pose estimation by constructing semantic super-parts via optimal transport, which guides prediction through a novel dual-range feature propagation. The part poses are then refined using cross-stage interaction and instance-level geometry encoding to enhance spatial precision and structural coherence. CFPA further incorporates a symmetry-aware loss that supervises multiple consistent pose configurations by explicitly modeling both intra-part and inter-part symmetries. Experiments on PartNet show that CFPA outperforms prior methods in pose accuracy, structural consistency, and assembly diversity.

2 Related Work

3D Part Assembly 3D part assembly, which aims to estimate 6-DoF poses for object parts to form coherent shapes, has evolved from rule-based and probabilistic approaches to modern deep learning frameworks. Early methods [1, 22–25] relied on handcrafted rules or statistical models to retrieve and align parts from shape repositories. With the rise of deep learning, recent methods focus on learning inter-part relationships and global structure directly from data. Graph-based models [6, 7, 24] capture local dependencies via message passing, while Transformer-based frameworks [8, 9, 17] enable long-range reasoning and instance-level disambiguation through attention mechanisms and structural constraints. To improve diversity and probabilistic modeling, generative methods such as VAEs, GANs, and diffusion models have been explored [15, 26, 27]. Task-specific strategies have also been designed, and Li et al. [28] propose precise alignment through peg-hole constraints, while guided approaches such as GPAT [29], Img-PA [18] and Imagine [30] leverage external cues like planning sequences or images. Despite these advances, many existing methods lack explicit semantic structuring and struggle with symmetry-induced uncertainties, which are critical for producing consistent and diverse assemblies in complex scenarios. Our work addresses these gaps by introducing semantic super-parts and a symmetry-aware reasoning for pose prediction.

Multi-scale Feature for Shape Generation Multi-scale feature modeling [31–35] plays a crucial role in modeling complex object structures. Early works such as StructureNet [36] and PT2PC [37] introduced hierarchical representations via trees and graphs to encode semantic part relationships, mainly for shape generation. Later, assembly-focused models like DGL [6] and RGL [7] extended this idea by leveraging graph neural networks to reason over both global context and local interactions during part refinement. Recent transformer-based models further expand multi-scale reasoning. SPAFormer [38] integrates global attention with PCA-based symmetry grouping to enhance long-range dependencies, while 3DHPA [17] models part-whole hierarchies through super-part message passing based on geometric similarity. The works of [18, 36, 39] also conduct feature interaction within geometrically similar part sets. In parallel, Score-PA [15] employs diffusion processes to jointly learn global and local feature distributions for generative assembly. However, prior methods commonly rely on fixed part groupings and uniform feature aggregation, which limits their ability to model both overall structure and detailed geometry, especially for repetitive or symmetric parts. We learn semantic part abstractions in a data-driven way and adaptively propagate global and local context to improve structural alignment.

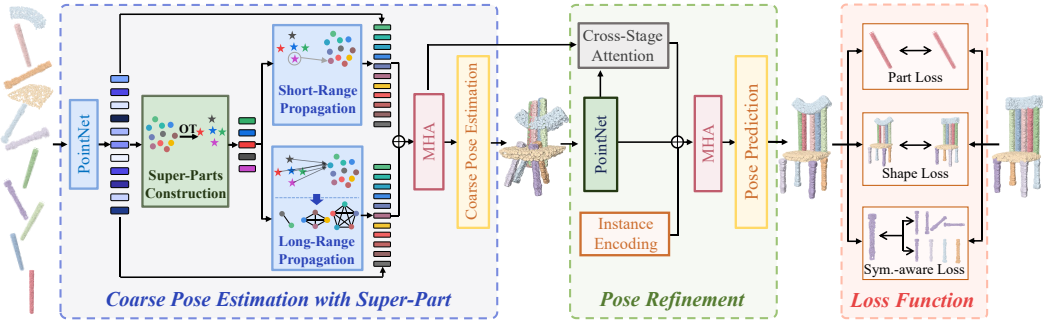

Figure 1: Pipeline of our CFPA. We first perform coarse pose estimation using super-parts derived via optimal transport, followed by pose refinement through cross-stage interactions. A symmetry-aware loss is proposed to improve pose accuracy while allowing structurally consistent variations.

Symmetry for Shape Understanding Symmetry is a fundamental property in 3D shapes, playing a crucial role in structural reasoning and part assembly. It includes self-symmetry, where a part remains invariant under transformations like rotation, translation, and reflection, and inter-part geometric similarity, where multiple parts share nearly identical shapes. Early works focused on detecting reflective and rotational symmetries for shape matching and alignment [40–44]. Later methods integrated symmetry into deep models for structural generation, such as StructureNet [36], which encodes symmetric part relationships via hierarchical graphs, as well as SAGNet [45] and ShapeFlow [46], which treat symmetry as an inductive bias to guide part deformation. SDM-NET [47] implicitly encodes symmetry in latent codes to guide part generation. Recent methods like DGL [6], Score-PA [15], and Img-PA [18] group geometrically similar parts for feature interaction, while SPAFormer [38] and 3DHPA [17] use PCA or bounding box similarity to enhance attention during part assembly. Symmetry is also particularly critical in deterministic fragment assembly, where it guides the precise reconstruction of broken objects using symmetry-based priors to align fractured parts [16, 19, 21]. These approaches typically encode symmetry as an architectural prior. In this work, we model inter-part similarity to enhance representations and further propose a symmetry-aware loss that accounts for both self-symmetry and geometric similarity, enabling diverse yet valid assemblies.

3 Method

Let $\{P_i\}_{i=1}^N$ denote a set of 3D part point clouds, where each part $P_i \in \mathbb{R}^{D\times 3}$ contains D points. Our goal is to predict the corresponding 6-DoF part poses $\{t_i^*, r_i^*\}_{i=1}^N$, where $t_i^* \in \mathbb{R}^3$ is the translation vector and $r_i^* \in \mathbb{R}^4$ is the rotational quaternion. The assembled shape is then given by $\mathcal{S}^* = \bigcup_{i=1}^N (r_i^* \circ P_i + t_i^*)$, with $\circ$ denoting the rotation operation, which can be implemented via Rodrigues' formula [48], and $\bigcup$ denoting the union operation.

Effective 3D part assembly requires not only the precise prediction of part poses for correct component assembly, but also a deep understanding of the global structure to ensure semantic consistency and generate plausible configurations. We propose CFPA (Coarse-to-Fine Part Assembly), a two-stage framework that integrates structural priors, geometric reasoning, and symmetry-aware supervision into the pose prediction pipeline. The overall architecture is illustrated in Figure 1. It begins with coarse pose estimation using semantic super-parts and a dual-range feature propagation strategy (Section 3.1), followed by a pose refinement stage that leverages cross-stage attention and instance encoding (Section 3.2). Additionally, we introduce a symmetry-aware loss that encourages both pose accuracy and assembly diversity (Section 3.3) by exploiting intra-part and inter-part symmetries.

3.1 Coarse Pose Estimation with Semantic Super-Part

We conduct coarse part-wise poses estimation guided by semantic super-parts, which serve as high-level structural priors capturing object-level semantics. These super-parts are constructed via an optimal transport formulation that yields compact and coherent part groupings in feature space. Their representations are then propagated to individual parts through a dual-range mechanism that integrates both local and global structural cues, enabling accurate pose prediction.

3.1.1 Semantic Super-Parts Construction via Optimal Transport

Given a set of parts $\{P_i\}_{i=1}^N$, we first extract part-wise features $\{f_i\}_{i=1}^N$ using a shared lightweight PointNet [49]. To capture the high-level structural information of the object, we then construct a set of semantic super-parts $\{h_j\}_{j=1}^M$ from $\{f_i\}_{i=1}^N$, with $M \leq N$. These super-parts serve as compact, semantically meaningful representations that encode the global structure of the target shape.

To encourage compact and semantically consistent part grouping, we formulate the assignment of part features to super-parts as an entropy-regularized optimal transport problem, which provides a principled and differentiable way to compute soft assignments. Specifically, we learn a transport matrix $T = \{T_{ij}\} \in \mathbb{R}^{N \times M}$ that defines soft correspondences between individual part features and super-part representations. Each super-part is then computed as a weighted aggregation of part features according to the transport matrix T:

$$h_j = \sum_{i=1}^{N} T_{ij} f_i, \quad j = 1, \ldots, M. \tag{1}$$

The transport matrix T can be obtained by minimizing the following entropy-regularized objective using Sinkhorn's algorithm [50]:

$$T^* = \arg\min_{T} \sum_{i=1}^{N} \sum_{j=1}^{M} T_{ij} C_{ij} - \epsilon \sum_{i=1}^{N} \sum_{j=1}^{M} T_{ij} \log T_{ij}, \tag{2}$$

where $C_{ij} = \langle f_i, h_j \rangle$ denotes the inner product cost between part feature f_i and super-part feature h_j, and $\epsilon > 0$ is a regularization coefficient that balances transport cost and entropy.

The resulting semantic super-parts $\{h_j\}_{j=1}^M$ encapsulate the global structural context implied by the input part set. These high-level representations serve as semantic anchors for subsequent dual-range feature propagation, providing informative priors for coarse part-wise pose estimation.

3.1.2 Dual-Range Feature Propagation

Leveraging the semantic super-parts, we propose a dual-range feature propagation to enhance part features by incorporating the local and global structural context, supporting accurate pose prediction.

Short-Range Feature Propagation To embed localized structural priors into part features, we perform short-range propagation from the nearest semantic super-part. For each part feature f_i, we identify its closest super-part from $\{h_j\}_{j=1}^M$ via Euclidean distance in the feature space, denoted by $h_i^\star$. The paired representations of $f_i, h_i^\star$ are then concatenated and passed through a multi-layer perceptron (MLP) to produce an enhanced representation:

$$\hat{f}_i = \text{MLP}\left([f_i, h_i^\star]\right), \quad i = 1, \ldots, N. \tag{3}$$

This operation enriches part representations with structural priors from their nearest semantic super-parts. The features $\{\hat{f}_i\}_{i=1}^N$ preserve local geometric detail while integrating contextual cues from nearest semantic super-part, enabling more coherent representations for downstream pose estimation.

Long-Range Feature Propagation To capture holistic structural dependencies at the global scale, we introduce a long-range feature propagation mechanism that integrates semantic information from all super-parts and reinforces spatial coherence through geometry-aware message passing. Specifically, we compute the Super-to-Base attention-weighted aggregation over all super-parts as:

$$f_i^{\text{S2B}} = \sum_{j=1}^{M} \alpha_{ij} \cdot (h_j W_V), \quad \textit{with } \{\alpha_{ij}\}_{j=1}^M = \text{Softmax}\left(\{(f_i W_Q)(h_j W_K)^\top\}_{j=1}^M\right), \tag{4}$$

where W_K, W_V, W_Q are learnable projection matrices. This mechanism enables each part to selectively integrate high-level semantic cues from the entire structural representation. To promote spatial coherence and suppress isolated responses, we further refine the attended features through message passing among geometrically similar parts:

$$f_i^{\text{S2B-MP}} = \sum_{g \in \Omega(i)} \beta_{ig} f_g^{\text{S2B}}, \quad i = 1, ..., N, \tag{5}$$

where $\Omega(i)$ denotes the set of geometrically similar parts to i, and β_{ig} is normalized affinity weights for the subgraph of geometrically similar parts (details in Appendix). The final long-range propagated feature is obtained by fusing original part feature f_i and refined representation $f_i^{\text{S2B-MP}}$ through MLP:

$$\tilde{f}_i = \text{MLP}\left(\left[f_i, f_i^{\text{S2B-MP}}\right]\right), \quad i = 1, ..., N. \tag{6}$$

This long-range propagation complements the short-range path by incorporating global semantic structure and promoting geometric consistency across parts, leading to more contextually informed part representations for downstream pose prediction.

3.1.3 Coarse Part Pose Estimation

Based on the short-range and long-range propagated features $\{\hat{f}_i\}_{i=1}^N$ and $\{\tilde{f}_i\}_{i=1}^N$, we estimate the coarse poses for individual parts. For each part, the two features are concatenated and passed through a multi-head attention (MHA) module to capture inter-part dependencies:

$$\{f_i^{\text{Coarse}}\}_{i=1}^N = \text{MHA}\left(\{[\hat{f}_i, \tilde{f}_i]\}_{i=1}^N\right). \tag{7}$$

The resulted coarse features $\{f_i^{\text{Coarse}}\}_{i=1}^N$ are then processed by MLP to predict rigid transformation, consisting of rotational quaternion r_i and translation vector t_i, and the transformed part $\bar{P}_i$ is achieved:

$$r_i, t_i = \text{MLP}(f_i^{\text{Coarse}}), \quad \bar{P}_i = r_i \circ P_i + t_i, \quad i = 1, ..., N. \tag{8}$$

By integrating semantic priors from super-parts and capturing structural dependencies via dual-range propagation, the coarse stage yields an initial pose estimation that is globally coherent and semantically informed. This provides a reliable initialization for subsequent pose refinement.

3.2 Pose Refinement

Building upon the coarsely transformed parts $\{\bar{P}_i\}_{i=1}^N$, we perform pose refinement by incorporating coarse-stage semantic guidance and modeling part-level geometric relations.

Specifically, we first extract part-wise features from $\{\bar{P}_i\}_{i=1}^N$ using a shared, lightweight PointNet [49], which produces the feature set $\{g_i\}_{i=1}^N$. These features serve as queries in a cross-stage attention module, with keys and values taken from the coarse-stage features $\{f_i^{\text{Coarse}}\}_{i=1}^N$. The resulting attended features $\{\tilde{g}_i\}_{i=1}^N$ encode coarse-to-fine guidance for each part. To further enhance spatial coherence and structural consistency, we incorporate instance encoding $\{e_i\}_{i=1}^N$, which encodes geometric similarity and inter-part relations following [8] (details in Appendix). Finally, we concatenate $g_i, \tilde{g}_i, e_i$ for each part, and apply MHA to obtain refined part feature $\{f_i^{\text{Refine}}\}_{i=1}^N$, and regress the final pose by MLP:

$$\{f_i^{\text{Refine}}\}_{i=1}^N = \text{MHA}\left(\{[g_i, \tilde{g}_i, e_i]\}_{i=1}^N\right), \quad r_i^*, t_i^* = \text{MLP}\left(f_i^{\text{Refine}}\right), \quad i = 1, ..., N, \tag{9}$$

where r_i^* and t_i^* denote the predicted rotational quaternion and translation vector for part P_i, respectively. The final transformed part and the assembled shape are computed as:

$$P_i^* = r_i^* \circ P_i + t_i^*, \qquad \mathcal{S}^* = \bigcup_{i=1}^N P_i^*. \tag{10}$$

This refinement stage enhances transformation precision by combining semantic guidance from the coarse stage with fine-grained geometric reasoning, ultimately yielding more accurate part-wise pose predictions and structurally coherent assemblies.

3.3 Training Objective

We design the training objective to optimize part-level pose accuracy and global structural consistency, while allowing symmetry-consistent diversity in part configurations, through a combination of part, shape, and symmetry-aware losses defined as follows.

Part and Shape Losses Denoting the predicted pose for part P_i as $\mathcal{F}(P_i) = \{r_i^*, t_i^*\}$, to ensure accurate part placement, we supervise the predicted transformation against its ground-truth pose $\{r_i^{\text{GT}}, t_i^{\text{GT}}\}$ using a standard part loss [6, 7, 15]:

$$\mathcal{L}_{Part} = \sum_{i=1}^{N} L_{pose}\left(\mathcal{F}(P_i), \{r_i^{\text{GT}}, t_i^{\text{GT}}\}\right), \quad \textit{with} \quad L_{pose}(\mathcal{F}(P_i), \{r_i^{\text{GT}}, t_i^{\text{GT}}\}) = \|t_i^* - t_i^{\text{GT}}\|_2 + \gamma d_c\left(r_i^* \circ P_i, r_i^{\text{GT}} \circ P_i\right), \tag{11}$$

where d_c denotes the Chamfer distance [18] between transformed point clouds, and γ balances translation and rotation terms. To further encourage global consistency, we supervise the assembled shape $\mathcal{S}^*$ against the ground-truth shape $\mathcal{S}^{\text{GT}}$ by shape loss as [6–8, 17]:

$$\mathcal{L}_{Shape} = d_c\left(\mathcal{S}^*, \mathcal{S}^{\text{GT}}\right). \tag{12}$$

The part and shape losses provide complementary supervision at the local and global levels, jointly promoting accurate pose estimation and structurally coherent assemblies.

Symmetry-Aware Loss To support diverse yet semantically valid part configurations, we introduce a symmetry-aware loss that explicitly accommodates multiple pose solutions induced by inherent part symmetries. It accounts for two cases, (1) self-symmetry where a part is indistinguishable from its flipped counterpart (e.g., a chair seat symmetric under vertical flipping as in Figure 2), and (2) inter-part geometric similarity where parts with nearly identical shapes can assume interchangeable roles (e.g., chair legs). These symmetries imply that structurally equivalent assemblies may admit multiple plausible configurations. Instead of enforcing a unique canonical pose, we select the best-aligned symmetric variant during training, enabling robust and symmetry-consistent learning.

For self-symmetry, we consider all sign-flipped variants of the ground-truth pose $\{r_i^{\text{GT}}, t_i^{\text{GT}}\}$, denoted as $\{\tau \circ \{r_i^{\text{GT}}, t_i^{\text{GT}}\}\}_{\tau \in \mathcal{Z}_2^3}$, where $\tau \in \mathcal{Z}_2^3$ represents a flipping operation along x-, y- and z-axis having $2^3 = 8$ possible transformations. We select the flip $\tau_i^\star$ that yields the lowest pose error with respect to the predicted pose $\mathcal{F}(P_i)$:

$$\tau_i^* = \underset{\tau \in \mathcal{Z}_2^3}{\arg\min}\, L_{pose}\left(\mathcal{F}(P_i), \tau \circ \{r_i^{\text{GT}}, t_i^{\text{GT}}\}\right). \tag{13}$$

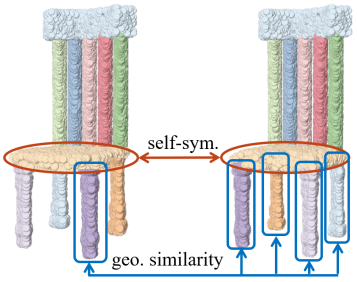

Figure 2: Parts with self-symmetry and geometric similarity in a chair.

To further account for assembly diversity arising from geometrically similar parts, we identify the set $\Omega(i)$ of ground-truth indices whose axis-aligned bounding box differences from P_i below a fixed threshold. The symmetry-aware loss encourages consistency between thr predicted pose $\mathcal{F}(P_i)$ and the most similar flipped ground-truth pose in this set, where each candidate j uses its own optimal self-symmetry transformation τ_j^* obtained from Eq. (13):

$$\mathcal{L}_{Sym} = \sum_{i=1}^{N} \min_{j \in \Omega(i)} L_{pose}(\mathcal{F}(P_i), \tau_j^* \circ \{r_j^{\text{GT}}, t_j^{\text{GT}}\}). \tag{14}$$

This formulation strengthens pose supervision by leveraging symmetry to accommodate diverse yet semantically valid configurations, while maintaining consistent with ground-truth transformations.

Total Training Objective We define the total training objective as a weighted combination of part loss, shape loss, and symmetry-aware loss as:

$$\mathcal{L} = \mathcal{L}_{Part} + \mathcal{L}_{Shape} + \lambda \mathcal{L}_{Sym}, \tag{15}$$

where λ is hyper-parameter balancing contribution of symmetry-aware loss in the overall optimization.

4 Experiment

Implementation Details We implement CFPA in PyTorch [51] using AdamW optimizer [52] with batch size 64. Following [6, 7, 15, 17], all the input parts are centralized and normalized by PCA and added with random noise during training. We adopt Min-of-N (MoN) strategy [53] for optimization. The model learns M=16 super-parts and uses 8-head attention in all MHA layers. Key hyper-parameters are set as ϵ=10^{-3}, γ=10, and λ=0.1. Code will be released if paper is accepted.

Table 1: Comparison for assembly accuracy evaluated by SCD, PA, and CA.

Methods	SCD(10^{-2}) ↓			PA(%) ↑			CA(%) ↑		
	Chair	Table	Lamp	Chair	Table	Lamp	Chair	Table	Lamp
B-Global [18, 54]	1.46	1.12	0.79	15.70	15.37	22.61	9.90	33.84	18.60
B-LSTM [55]	2.35	1.71	0.90	8.08	10.55	24.68	10.05	18.28	30.23
DGL [6]	0.91	0.50	0.93	39.00	49.51	33.33	23.87	39.96	41.70
Score-PA [15]	0.71	0.42	1.11	44.51	52.78	34.32	30.32	40.59	49.07
IET [8]	1.34	0.66	0.89	37.60	48.86	32.86	25.44	40.35	52.75
SPAFormer [38]	0.67	0.38	-	55.88	64.38	-	36.39	57.60	-
RGL [7]	0.98	0.40	1.05	48.85	55.13	35.54	30.68	41.41	50.09
3DHPA [17]	0.51	**0.32**	0.82	63.01	64.58	33.49	48.28	58.00	62.01
Ours	**0.49**	0.33	**0.77**	**69.24**	**68.48**	**36.35**	**49.20**	**58.51**	**63.32**

Table 2: Comparison for assembly accuracy with mPA, mCA metrics.

Methods	mPA(%) ↑			mCA(%) ↑		
	Chair	Table	Lamp	Chair	Table	Lamp
B-Global [18, 54]	28.01	44.55	37.03	29.05	44.27	42.34
B-LSTM [55]	14.49	15.92	43.75	18.91	31.77	43.21
DGL [6]	52.85	60.46	49.17	41.62	54.94	57.52
Score-PA [15]	58.80	65.61	49.06	43.96	53.72	62.55
IET [8]	49.29	59.34	48.29	42.51	55.27	64.47
RGL [7]	59.40	53.05	50.12	47.28	56.51	62.07
3DHPA [17]	77.06	75.06	48.47	64.47	71.00	75.81
Ours	**81.28**	**78.57**	**52.84**	**65.73**	**71.84**	**76.58**

Dataset In line with prior studies [6–8, 15, 18, 54, 55], we evaluate our method on the PartNet dataset [56], focusing on the three largest object categories: Chair (6,323 shapes), Table (8,218 shapes), and Lamp (2,207 shapes). We follow the official data splits for every categories, using 70% of the shapes for training, 10% for validation, and the remaining 20% for testing.

Evaluation Metrics Following prior work [6–8, 15, 18, 54, 55], we evaluate assembly performance using five metrics: (1) *Shape Chamfer Distance* (SCD), which measures the geometric similarity between assembled and ground-truth shapes; (2) *Part Accuracy* (PA), which assesses the correctness of predicted part transformations based on a predefined distance threshold; (3) *Connectivity Accuracy* (CA), which evaluates whether adjacent parts in the assembled shape are correctly connected within a specified threshold; (4) *Quality-Diversity Score* (QDS) and (5) *Weighted QDS* (WQDS) [15], both of which quantify the diversity and structural plausibility of generated assemblies. Detailed definitions of all metrics are provided in the Appendix.

4.1 Results and Comparisons

We evaluate assembly accuracy on Chair, Table, and Lamp categories using SCD, PA, and CA, with PA and CA computed under a Chamfer distance threshold of 0.01. As shown in Table 1, CFPA achieves the best results in 8 out of 9 cases and ranks second in the remaining one, demonstrating its effectiveness in both part assembly accuracy and structural consistency.

We also report PA and CA under varying thresholds (0.01-0.05), with performance curves in Figure 3 and average results in Table 2. CFPA consistently outperforms others across all settings. Notably, the work of 3DHPA [17], which employs geometry-based super-parts also performs well on CA, highlighting the benefit of hierarchical structural priors. Detailed values are in the Appendix.

We further evaluate assembly diversity using QDS and WQDS. As shown in Table 3, CFPA achieves the highest WQDS across all categories and the highest QDS on Chair, indicating its ability to generate shape diverse yet structurally valid assemblies.

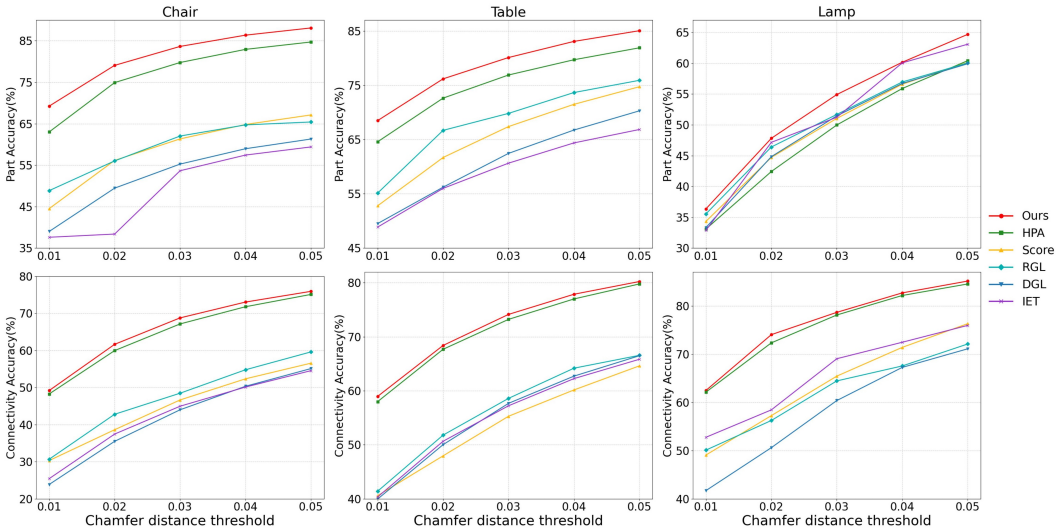

Figure 3: Performance curves of our CFPA and compared models on the Chair, Table and Lamp categories under Chamfer distance threshold ranging from 0.01 to 0.05. Best viewed in color.

Table 3: Comparison of assembled shape diversity evaluated using QDS and WQDS.

Methods	QDS(10^{-5}) ↑			WQDS(10^{-5}) ↑		
	Chair	Table	Lamp	Chair	Table	Lamp
B-Global [18, 54]	0.15	0.20	0.76	1.25	1.40	0.58
B-LSTM [55]	3.92	1.33	3.05	1.26	0.55	2.01
DGL [6]	1.69	3.05	1.84	1.35	2.97	1.73
Score-PA [15]	3.36	**9.17**	**6.83**	1.70	3.81	2.82
IET [8]	3.33	6.22	4.93	1.85	2.35	3.43
RGL [7]	5.85	7.55	6.37	2.09	3.51	3.15
3DHPA [17]	4.42	7.15	4.67	1.90	3.80	3.16
Ours	**6.71**	7.28	5.65	**2.75**	**3.92**	**3.74**

4.2 Ablation Study

Effectiveness of Super-Part. CFPA predicts coarse part poses with the help of semantic super-parts constructed via optimal transport (OT) in the feature space. We compare ith with three variants: 1) CFPA-w/o-SP, which removes super-part guidance; 2) CFPA-GE-SP, which builds super-parts based on geometric similarity; 3) CFPA-KM-SP, which uses K-means clustering [57] to build semantic super-parts based on basic part features. As shown in Table 4, super-part guidance improves performance, with OT-based construction achieving the best results over heuristic variants.

Table 4: Ablation study on super-parts.

Methods	SCD(10^{-2}) ↓	PA(%) ↑	CA(%) ↑
1) CFPA-w/o-SP	0.54	66.75	47.75
2) CFPA-GE-SP	0.53	67.60	47.97
3) CFPA-KM-SP	0.54	69.20	47.48
CFPA	**0.49**	**69.24**	**49.20**

Effectiveness of Dual-Range Feature Propagation. CFPA incorporates both short-range and long-range feature propagation to transfer information from super-parts to individual parts. We evaluate three ablated variants: 4) CFPA-w/o-SRFP, which removes short-range propagation; 5) CFPA-w/o-LRFP, which removes long-range propagation; 6) CFPA-w/o-MP, which omits feature refinement by message passing within geometrically similar parts as defined in

Table 5: Ablation study on designs in the coarse pose estimation stage and pose refinement stage.

Methods	SCD(10^{-2})↓	PA(%)↑	CA(%)↑
4) CFPA-w/o-SRFP	0.57	66.77	44.71
5) CFPA-w/o-LRFP	0.51	68.93	47.56
6) CFPA-w/o-MP	0.60	63.60	40.28
7) CFPA-w/o-CA	0.51	67.47	44.42
8) CFPA-w/o-IE	0.55	64.81	44.44
CFPA	**0.49**	**69.24**	**49.20**

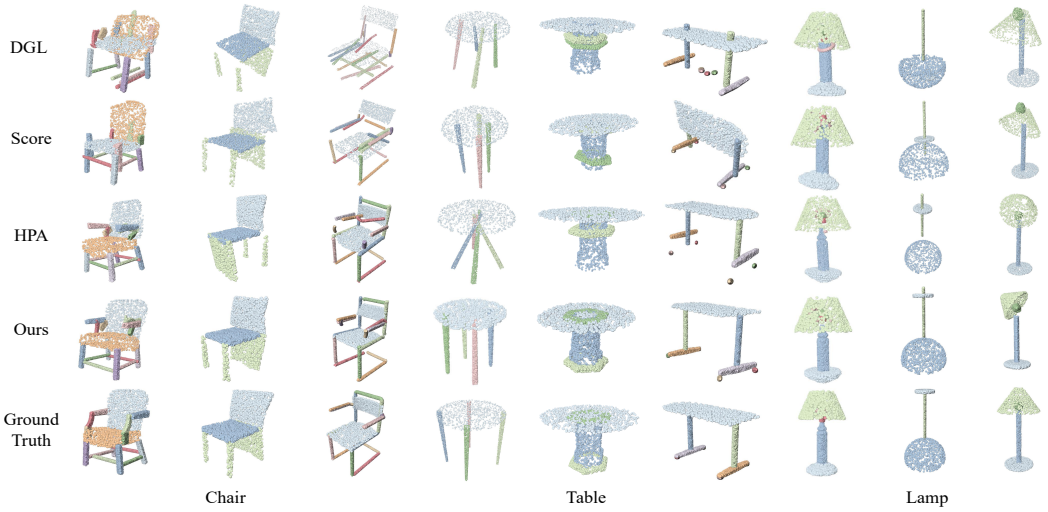

Figure 4: Qualitative results on the Chair, Table and Lamp categories.

Eq. (5). As shown in Table 5, all variants exhibit degraded performance, underscoring the importance of both propagation mechanisms.

Effectiveness of Operations in Pose Refinement Stage. We assess the impact of cross-stage attention and instance encoding by ablating each component with 7) CFPA-w/o-CA, which removes cross-stage attention; 8) CFPA-w/o-IE, which removes instance encoding. As shown in Table 5, both variants yield reduced performance, confirming the effectiveness of incorporating coarse-stage guidance and geometric relationships.

Effectiveness of Symmetry-Aware Loss. The symmetry-aware loss addresses both self-symmetry and geometric similarity among interchangeable parts. We ablate three variants: 9) CFPA-w/o-SS, which removes self-symmetry supervision; 10) CFPA-w/o-GS, which removes constraints on geometrically similar parts; 11) CFPA-w/o-SL, which disables the symmetry-aware loss entirely. As shown in Table 6, removing either component degrades performance, confirming the effectiveness of symmetry-aware supervision.

Table 6: Ablation study on symmetry-aware loss.

Methods	SCD(10^{-2}) ↓	PA(%) ↑	CA(%) ↑
9) CFPA-w/o-SS	0.51	67.86	47.24
10) CFPA-w/o-GS	0.51	68.98	47.49
11) CFPA-w/o-SL	0.52	67.51	47.17
CFPA	**0.49**	**69.24**	**49.20**

4.3 Qualitative Results

Figure 4 presents the qualitative results on Chair, Table and Lamp categories. Our method produces more reasonable and structurally coherent assemblies compared to previous methods. The generated shapes by our CFPA demonstrate assembly diversity arising from self-symmetric components and geometrically similar parts, while preserving overall structural validity.

Figure 5 illustrates the coarse-to-fine assembly process. Starting with PCA-normalized parts as inputs, the Coarse Pose Estimation stage estimates initial part poses (2nd column). In the Pose Refinement stage, part poses are refined to produce more accurate and structurally coherent assembly configurations (3rd column), leading to final shapes that closely match ground truth.

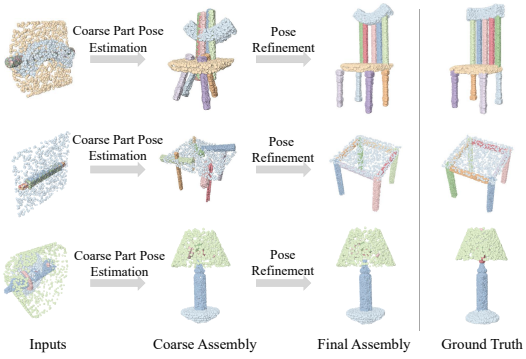

Figure 5: Visualization of coarse-to-fine progress.

More experimental details, ablation studies, and visualization are provided in the Appendix.

5 Conclusion

We propose CFPA, a coarse-to-fine framework for 3D part assembly that integrates semantic super-parts, dual-range feature propagation, and symmetry-aware supervision. By jointly modeling global structure and geometric symmetry, CFPA enables accurate and diverse part assemblies. Extensive experiments show that it achieves state-of-the-art performance in assembly accuracy, structural consistency, and diversity across multiple categories. In future work, we plan to extend CFPA to fragment reassembly and explore unsupervised learning to enhance robustness on unseen parts.

Limitation CFPA is designed for part-level assembly with semantically meaningful components and is not directly applicable to fragment reassembly involving irregular or incomplete parts. Extending the framework to handle such cases is a promising direction.

Impact Statement This work contributes to 3D part assembly by introducing a coarse-to-fine framework that unifies semantic abstraction and symmetry-aware reasoning. By explicitly modeling hierarchical structure and geometric symmetry, our method improves both accuracy and diversity in assembly tasks. The proposed approach can benefit downstream applications in 3D design, CAD modeling, and intelligent robotic assembly. This work poses no ethical concerns.

Acknowledgments and Disclosure of Funding

This work was supported by NSFC with grant numbers of 62306167, 12125104, 12426313, and project ZR2024QA161, ZR2023QG143 supported by Shandong Provincial Natural Science Foundation. It is also funded by National Social Science Fund Project (25CJY023).

References

[1] Thomas Funkhouser, Michael Kazhdan, Philip Shilane, Patrick Min, William Kiefer, Ayellet Tal, Szymon Rusinkiewicz, and David Dobkin. Modeling by example. *ACM Transactions on Graphics (TOG)*, 23(3):652–663, 2004.

[2] Yuval Litvak, Armin Biess, and Aharon Bar-Hillel. Learning pose estimation for high-precision robotic assembly using simulated depth images. In *International Conference on Robotics and Automation (ICRA)*, pages 3521–3527, 2019.

[3] Jianlan Luo, Eugen Solowjow, Chengtao Wen, Juan Aparicio Ojea, Alice M Agogino, Aviv Tamar, and Pieter Abbeel. Reinforcement learning on variable impedance controller for high-precision robotic assembly. In *International Conference on Robotics and Automation (ICRA)*, pages 3080–3087, 2019.

[4] Lin Shao, Toki Migimatsu, and Jeannette Bohg. Learning to scaffold the development of robotic manipulation skills. In *International Conference on Robotics and Automation (ICRA)*, pages 5671–5677, 2020.

[5] Kevin Zakka, Andy Zeng, Johnny Lee, and Shuran Song. Form2fit: Learning shape priors for generalizable assembly from disassembly. In *International Conference on Robotics and Automation (ICRA)*, pages 9404–9410, 2020.

[6] Guanqi Zhan, Qingnan Fan, Kaichun Mo, Lin Shao, Baoquan Chen, Leonidas J Guibas, Hao Dong, et al. Generative 3d part assembly via dynamic graph learning. *Advances in Neural Information Processing Systems*, 33:6315–6326, 2020.

[7] Abhinav Narayan, Rajendra Nagar, and Shanmuganathan Raman. Rgl-net: A recurrent graph learning framework for progressive part assembly. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 78–87, 2022.

[8] Rufeng Zhang, Tao Kong, Weihao Wang, Xuan Han, and Mingyu You. 3d part assembly generation with instance encoded transformer. *IEEE Robotics and Automation Letters*, 7(4):9051–9058, 2022.

[9] Ruiyuan Zhang, Jiaxiang Liu, Zexi Li, Hao Dong, Jie Fu, and Chao Wu. Scalable geometric fracture assembly via co-creation space among assemblers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7269–7277, 2024.

[10] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[11] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

[12] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[13] Jun Li, Kai Xu, Siddhartha Chaudhuri, Ersin Yumer, Hao Zhang, and Leonidas Guibas. Grass: Generative recursive autoencoders for shape structures. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017.

[14] Hao Wang, Nadav Schor, Ruizhen Hu, Haibin Huang, Daniel Cohen-Or, and Hui Huang. Global-to-local generative model for 3d shapes. *ACM Transactions on Graphics (TOG)*, 37(6):1–10, 2018.

[15] Junfeng Cheng, Mingdong Wu, Ruiyuan Zhang, Guanqi Zhan, Chao Wu, and Hao Dong. Score-pa: Score-based 3d part assembly. *British Machine Vision Conference (BMVC)*, 2023.

[16] Jae Eun Kim, Muhammad Zeeshan Arshad, Yoo Seong Jong, Je-Hyung Hong, Jinwook Kim, and Young Min Kim. 3d pots configuration system by optimizing over geometric constraints. In *International Conference on Pattern Recognition (ICPR)*, pages 2398–2405, 2021.

[17] Bi'an Du, Xiang Gao, Wei Hu, and Renjie Liao. Generative 3d part assembly via part-whole-hierarchy message passing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20850–20859, 2024.

[18] Yichen Li, Kaichun Mo, Lin Shao, Minhyuk Sung, and Leonidas Guibas. Learning 3d part assembly from a single image. In *European Conference on Computer Vision*, pages 664–682, 2020.

[19] Seyed-Mahdi Nasiri, Reshad Hosseini, and Hadi Moradi. Multiple-solutions ransac for finding axes of symmetry in fragments of objects. *Pattern Recognition*, 131:108805, 2022.

[20] Je Hyeong Hong, Young Min Kim, Koang-Chul Wi, and Jinwook Kim. Potsac: A robust axis estimator for axially symmetric pot fragments. In *ICCV Workshops*, pages 1421–1428, 2019.

[21] Jiaxin Lu, Gang Hua, and Qixing Huang. Jigsaw++: Imagining complete shape priors for object reassembly. *arXiv preprint arXiv:2410.11816*, 2024.

[22] Arjun Jain, Thorsten Thormählen, Tobias Ritschel, and Hans-Peter Seidel. Exploring shape variations by 3d-model decomposition and part-based recombination. *Computer Graphics Forum*, 31:631–640, 2012.

[23] Evangelos Kalogerakis, Siddhartha Chaudhuri, Daphne Koller, and Vladlen Koltun. A probabilistic model for component-based shape synthesis. *ACM Transactions on Graphics (TOG)*, 31(4):1–11, 2012.

[24] Prakhar Jaiswal, Jinmiao Huang, and Rahul Rai. Assembly-based conceptual 3d modeling with unlabeled components using probabilistic factor graph. *Computer-Aided Design*, 74:45–54, 2016.

[25] Siddhartha Chaudhuri, Evangelos Kalogerakis, Leonidas Guibas, and Vladlen Koltun. Probabilistic reasoning for assembly-based 3d modeling. *ACM Transactions on Graphics (TOG)*, 30(4):1–10, 2011.

[26] Gianluca Scarpellini, Stefano Fiorini, Francesco Giuliari, Pietro Moreiro, and Alessio Del Bue. Diffassemble: A unified graph-diffusion model for 2d and 3d reassembly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28098–28108, 2024.

[27] Ruiyuan Zhang, Qi Wang, Jiaxiang Liu, Yu Zhang, Yuchi Huo, and Chao Wu. Leveraging pretrained diffusion models for zero-shot part assembly. *arXiv preprint arXiv:2505.00426*, 2025.

[28] Yichen Li, Kaichun Mo, Yueqi Duan, He Wang, Jiequan Zhang, and Lin Shao. Category-level multi-part multi-joint 3d shape assembly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3281–3291, 2024.

[29] Yulong Li, Andy Zeng, and Shuran Song. Rearrangement planning for general part assembly. In *Conference on Robot Learning*, 2023.

[30] Weihao Wang, Yu Lan, Mingyu You, and Bin He. Imagine: Image-guided 3d part assembly with structure knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7889–7897, 2025.

[31] Lintai Wu, Junhui Hou, Linqi Song, and Yong Xu. 3d shape completion on unseen categories: A weakly-supervised approach. *arXiv preprint arXiv:2401.10578*, 2024.

[32] G Sambit, YL Xian, B Aditya, S Soumik, and K Adarsh. Multi-level 3d cnn for learning multi-scale spatial features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[33] Mengran Gao, Ningjun Ruan, Junpeng Shi, and Wanli Zhou. Deep neural network for 3d shape classification based on mesh feature. *Sensors*, 22(18):7040, 2022.

[34] Francesco Milano, Antonio Loquercio, Antoni Rosinol, Davide Scaramuzza, and Luca Carlone. Primal-dual mesh convolutional neural networks. *Advances in Neural Information Processing Systems*, 33:952–963, 2020.

[35] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021.

[36] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. *ACM Transactions on Graphics (TOG), Siggraph Asia*, 38(6):Article 242, 2019.

[37] Kaichun Mo, He Wang, Xinchen Yan, and Leonidas Guibas. Pt2pc: Learning to generate 3d point cloud shapes from part tree conditions. In *European Conference on Computer Vision*, pages 683–701, 2020.

[38] Boshen Xu, Sipeng Zheng, and Qin Jin. SPAFormer: Sequential 3d part assembly with transformers. In *International Conference on 3D Vision*, 2025.

[39] Despoina Paschalidou, Luc van Gool, and Andreas Geiger. Learning unsupervised hierarchical part decomposition of 3d objects from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[40] Niloy J Mitra, Mark Pauly, Michael Wand, and Duygu Ceylan. Symmetry in 3d geometry: Extraction and applications. *Computer Graphics Forum*, 32:1–23, 2013.

[41] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Symmetry descriptors and 3d shape matching. In *Proceedings of Eurographics/ACM SIGGRAPH Symposium on Geometry processing*, pages 115–123, 2004.

[42] Yaron Lipman, Xiaobai Chen, Ingrid Daubechies, and Thomas Funkhouser. Symmetry factored embedding and distance. *ACM Transactions on Graphics (TOG)*, 29(4), 2010.

[43] Joshua Podolak, Philip Shilane, Aleksey Golovinskiy, Szymon Rusinkiewicz, and Thomas Funkhouser. A planar-reflective symmetry transform for 3d shapes. *ACM Transactions on Graphics (TOG)*, 25(3):549–559, 2006.

[44] Yanzhen Wang, Kai Xu, Jun Li, Hao Zhang, Ariel Shamir, Ligang Liu, Zhiquan Cheng, and Yueshan Xiong. Symmetry hierarchy of man-made objects. *Computer Graphics Forum*, 30:287–296, 2011.

[45] Zhijie Wu, Xiang Wang, Di Lin, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Sagnet: Structure-aware generative network for 3d-shape modeling. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

[46] Chiyu Jiang, Jingwei Huang, Andrea Tagliasacchi, and Leonidas J Guibas. Shapeflow: Learnable deformation flows among 3d shapes. *Advances in Neural Information Processing Systems*, 33:9745–9757, 2020.

[47] Lin Gao, Jie Yang, Tong Wu, Yu-Jie Yuan, Hongbo Fu, Yu-Kun Lai, and Hao Zhang. Sdm-net: Deep generative network for structured deformable mesh. *ACM Transactions on Graphics (TOG)*, 38(6):1–15, 2019.

[48] Camillo J Taylor and David J Kriegman. Minimization on the lie group so (3) and related manifolds. *Yale Center For Systems Science Technical Report*, 1994.

[49] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[50] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

[51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

[52] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[53] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 605–613, 2017.

[54] Nadav Schor, Oren Katzir, Hao Zhang, and Daniel Cohen-Or. Componet: Learning to generate the unseen by part synthesis and composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8759–8768, 2019.

[55] Rundi Wu, Yixin Zhuang, Kai Xu, Hao Zhang, and Baoquan Chen. Pq-net: A generative part seq2seq network for 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 829–838, 2020.

[56] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 909–918, 2019.

[57] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 5, pages 281–298, 1967.

[58] Carsten Moenning and Neil A Dodgson. Fast marching farthest point sampling. Technical report, University of Cambridge, Computer Laboratory, 2003.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims presented in the abstract and introductions are supported by empirical evidence in the experimental section (Section 4).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of this study are explicitly addressed in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental setups are described in Section 4 and Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code will be available if paper is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setups are described in Section 4 and Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The experiments in this study do not require statistical significance testing or error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experimental setups are described in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research respects NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Details can be found in Section 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper does not release new datasets or models, so this is not applicable.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This paper uses open-weight models and public datasets for experiments. The usage respects all the original licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There are no new assets introduced in the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components. There is no evidence that LLMs were used in a way that affects the core methodology, scientific rigorousness, or originality of the research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

A Model Details

Physical meaning & intuitive interpretation of super-part from OT Physically, as Eq. (1), each super-part is a weighted aggregation of basic part features learned via optimal transport, forming compact prototypes that encode multiple related parts. Intuitively, these super-parts group functionally or geometrically similar parts into semantic clusters, providing mid-level abstractions without predefined hierarchies.

Details on Geometrically Similar Part Set and Normalized Affinity Weights In Eq. (5) of the main paper, we construct a graph to facilitate message passing among geometrically similar parts, thereby enhancing feature consistency across structurally related components. In this graph, each node represents an individual part, and edges are formed based on geometric similarity. This process involves two steps: (1) constructing the sets of geometrically similar parts, and (2) computing normalized affinity weights.

To determine geometric similarity, we follow [6] and compare the axis-aligned bounding box sizes of each part. Two parts are considered similar if the absolute difference in their bounding box sizes is below a predefined threshold (set to 0.1). These geometrically similar parts, such as chair arms or legs, are grouped into fully connected subgraphs to enable dense information exchange. Parts that do not belong to any similarity sets are treated as structurally unique and form isolated subgraphs; these subgraphs are fully connected internally to ensure message propagation. Figure A.1 visualizes an example of the graph resulting from a chair model.

For each subgraph, we assign uniform affinity weights to guide message passing. For a given part P_i and its geometrically similar part $P_{g\in\Omega(i)}$, we compute the normalized affinity weight as:

$$\beta_{ig} = \frac{1}{|\Omega(i)|}, \quad \forall g \in \Omega(i), \tag{16}$$

where $|\Omega(i)|$ denotes the number of parts in the geometrically similar set for part P_i. This uniform weighting avoids bias toward any individual part and ensures stable and balanced feature aggregation within each subgraph.

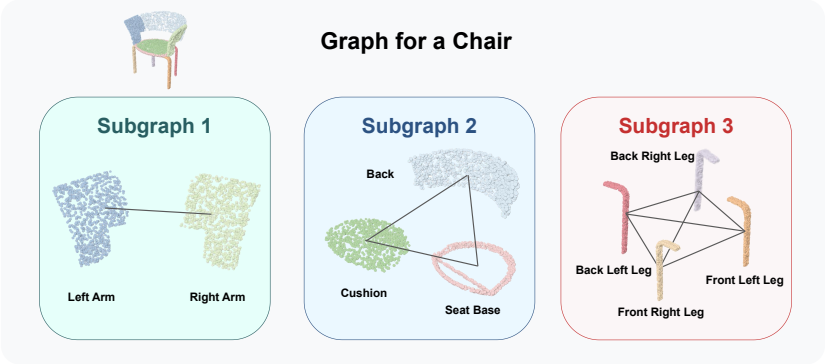

Figure A.1: Graph for a chair. Nodes represent individual parts, and edges denote connectivity based on geometric similarity. Subgraphs 1 and 3 contain geometrically similar parts (e.g., arms, legs) which are fully connected to facilitate message passing. Subgraph 2 consists of geometrically isolated parts (e.g., back, cushion, seat base) that do not belong to any geometrically similar part set, and they are fully connected to enable massage passing among structurally unique parts.

Details on Pose Refinement Stage The pose refinement stage enhances part-level representations by incorporating cross-stage attention and instance encoding, enabling more accurate and structurally consistent pose estimation.

The cross-stage attention mechanism utilizes coarse features $\{f_i^{Coarse}\}_{i=1}^N$ as keys and values, and part features $\{g_i\}_{i=1}^N$ as queries, where all feature vectors are of dimension 256. We implement

multi-head attention (MHA) with 8 heads operating in a 32-dimensional subspace using scaled dot-product attention.

For instance encoding, following [8], we employ a 40-dimensional embedding $\{e_i\}_{i=1}^{N}$ as a unique instance encoding vector for part i, formed by concatenating an inter-class encoding that uniquely identifies each individual part and an intra-class encoding that establishes relationships among geometrically-similar parts.

For pose prediction, we first pass the concatenated features $[g_i, \tilde{g}_i, e_i]$ through a MHA with 8-head to facilitate part-to-part information exchange. The output features are then processed by an multilayer perceptron (MLP) with dimensions $[1024, 512, 7]$ using ReLU activations, where the final layer predicts rotational quaternion and translation vector for each part.

Details on Lightweight PointNet The input part features are extracted using a shared PointNet [49] applied to the part point clouds. An MLP (with dimensions $[64, 256, 512, 1024]$) is first applied to transform the 3-dimensional point positions into 1024-dimensional features, and then a global max-pooling operation is applied to aggregate point-level features into a compact part-wise representation, which is further processed through three fully connected layers (with dimensions $[1024, 512, 256]$) to obtain a final 256-dimensional embedding for each part.

Computational cost of the proposed components To evaluate the computational cost of our proposed components, we report key metrics for a single forward pass with a batch size of 64, including the number of parameters (Million, M), GPU memory usage (GB), forward time (ms), and GFLOPS. We analyze the CFPA model and its variants, each excluding specific components as: 1)-2) CFPA-w/o-refine/coarse MHA that exclude the multi-head attention in the pose refinement/coarse pose estimation stage; 3) CFPA-w/o-OT that removes the OT-based super-part construction; 4)-5) CFPA-w/o-SRFP/LRFP that remove the short-/long-range feature propagation; 6) CFPA-w/o-MP that removes the message passing in long-range feature propagation; 7) CFPA-w/o-CA that removes the cross-stage attention; 8) CFPA-w/o-SL that removes the symmetry-aware loss. We also report the computational cost of the baseline method that removes all of the above components.

As shown in Table A.1, removing these components leads to a reduction in computational cost compared to the full CFPA model. Notably, some of these components are specifically designed to operate based on the relationships between parts and super-parts rather than directly on point clouds, with the maximum number of parts limited to 20 and super-parts to 16. This design inherently requires only a small computational cost. Additionally, certain components, such as OT, short-range feature propagation, and message passing, do not introduce any learnable parameters, further contributing to their relatively low computational cost.

Table A.1: Computational cost comparisons.

Method	#Para.(M)	GPU Memory(GB)	Forward time(ms)	GFLOPS
baseline	21.33	24.05	641.36	86.22
1) CFPA-w/o-refine MHA	22.90	24.28	674.80	86.26
2) CFPA-w/o-coarse MHA	25.14	26.72	772.09	86.30
3) CFPA-w/o-OT	25.89	26.80	657.87	86.34
4) CFPA-w/o-SRFP	25.89	26.36	804.91	86.34
5) CFPA-w/o-LRFP	25.76	24.68	778.68	86.33
6) CFPA-w/o-MP	25.89	24.60	791.31	86.34
7) CFPA-w/o-CA	25.63	25.71	698.05	86.33
8) CFPA-w/o-SL	25.89	26.60	802.72	86.34
CFPA	**25.89**	**27.04**	**806.80**	**86.34**

B Experiment Details

The model is trained for 500 epochs using a batch size of 64 across 4 NVIDIA RTX 4090 GPUs. We adopt the AdamW optimizer with an initial learning rate of 7.5×10^{-5} and a weight decay

of 1×10^{-4}. A cosine annealing learning rate schedule is applied with a decay factor of 100 to progressively reduce the learning rate over the training process.

Each input 3D part is represented as a point cloud containing 1,000 points, which are sampled by Farthest Point Sampling [58]. To ensure consistency and invariance to translation and rotation, all parts are pre-aligned to canonical coordinate system by Principal Component Analysis (PCA).

B.1 Evaluation Metrics

We evaluate our method using five commonly adopted metrics in part assembly tasks: Shape Chamfer Distance (SCD), Part Accuracy (PA), Connectivity Accuracy (CA), Quality-Diversity Score (QDS), and Weighted Quality-Diversity Score (WQDS). The definitions of these metrics are provided below.

Shape Chamfer Distance (SCD) The Shape Chamfer Distance (SCD) quantifies the geometric discrepancy between two point clouds by computing the average nearest-neighbor distance between points in each set. Formally, given two point sets $\mathcal{X}$ and $\mathcal{Y}$, the Chamfer distance is defined as:

$$d_c(\mathcal{X}, \mathcal{Y}) = \sum_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} \|x - y\|_2^2 + \sum_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \|x - y\|_2^2 \tag{17}$$

In our case, SCD is computed between the predicted assembled shape S^* and the ground truth shape S^{GT}, i.e., $d_c(S^*, S^{\text{GT}})$.

Part Accuracy (PA) It evaluates the proportion of parts whose transformed point clouds, using the predicted transformations, yield accurate geometric alignment with their ground-truth counterparts, as measured by the Chamfer Distance. Specifically, it is defined as:

$$\text{PA} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(d_c\left(\mathcal{T}\left(P_i\right), \mathcal{T}^{GT}(P_i)\right) < \tau_p\right), \tag{18}$$

where $\mathbf{1}(\cdot)$ is indicator function. $\mathcal{T}(P_i)$ and $\mathcal{T}^{GT}(P_i)$ represent the transformed part point clouds using the predicted and ground-truth transformations, respectively. d_c represents the Chamfer distance, and τ_p is the predefined threshold.

Connectivity Accuracy (CA) It measures the structural correctness of pairwise part connections in the assembled shape. For each contact point pair $\{c_{ij}^*, c_{ij}^*\}$ in the ground-truth object space, where c_{ij}^* is the point on part i closest to part j and c_{ji}^* is the corresponding nearest point on part j to part i. we identify their corresponding coordinates in the canonical part spaces (normalized using PCA) as $\{c_{ij}, c_{ij}\}$. The predicted transformations $\mathcal{T}_i, \mathcal{T}_j$ are then applied to these canonical points, and their distance is measured in the predicted assembled space. The connectivity accuracy is defined as:

$$\text{CA} = \frac{1}{|\mathcal{C}|} \sum_{\{c_{ij}, c_{ji}\} \in \mathcal{C}} \mathbf{1}\left(d\left(\mathcal{T}_i\left(c_{ij}\right), \mathcal{T}_j\left(c_{ji}\right)\right) < \tau_c\right), \tag{19}$$

where $\mathcal{C}$ denotes the set of all contact point pairs in the assembly, d is the Euclidean distance, and τ_c is the predefined threshold.

Meanwhile, following [17], we apply mean part accuracy (mPA) and mean connectivity accuracy (mCA) for more comprehensive evaluation as follows:

$$\text{mPA} = \frac{1}{|\Xi|} \sum_{\tau_p \in \Xi} \text{PA}, \quad \text{mCA} = \frac{1}{|\Xi|} \sum_{\tau_c \in \Xi} \text{CA}, \tag{20}$$

where $\Xi = \{0.01, 0.02, 0.03, 0.04, 0.05\}$ represents a predefined set of Chamfer distance thresholds.

Quality-Diversity Score (QDS) Following [15, 17], we use QDS to jointly evaluate the structural quality and geometric diversity of generated assemblies, which is defined as:

$$\text{QDS} = \frac{1}{N^2} \sum_{i,j=1}^{N} d_c\left(S_i^*, S_j^*\right) \mathbf{1}\left(\text{CA}(S_i^*) > \tau_q\right) \mathbf{1}\left(\text{CA}(S_j^*) > \tau_q\right), \tag{21}$$

where $\mathbf{1}(\cdot)$ is indicator function, and $d_c(\cdot)$ is Chamfer distance. The threshold τ_q is set to 0.5. QDS incorporates quality constraints by excluding pairs with low assembly quality, simultaneously evaluating both diversity and quality of the generated shapes.

Weighted Quality-Diversity Score WQDS is the weighted QDS with the connectivity accuracy, prioritizing assembled shapes with high-quality connections between parts:

$$\mathrm{WQDS} = \frac{1}{N^2} \sum_{i,j=1}^{N} d_c\left(S_i^*, S_j^*\right) \mathrm{CA}\left(S_i^*\right) \mathrm{CA}\left(S_j^*\right) \mathbf{1}\left(\mathrm{CA}\left(S_i^*\right) > \tau_q\right) \mathbf{1}\left(\mathrm{CA}\left(S_j^*\right) > \tau_q\right). \quad (22)$$

B.2 Experimental Results under Multiple Thresholds

Tables B.1–B.3 report detailed numerical results corresponding to Figure 2 in the main paper, presenting PA and CA for our CFPA and several state-of-the-art baselines across a range of Chamfer distance thresholds (0.01–0.05) on the Chair, Table, and Lamp categories. The consistently superior performance of CFPA across all thresholds and object classes highlights its robustness and effectiveness in achieving accurate and structurally coherent part assemblies.

Table B.1: Comparisons on Chair category of PA / CA at Chamfer distance thresholds of 0.01 to 0.05.

Method	PA(%) ↑						CA(%) ↑					
	0.01	0.02	0.03	0.04	0.05	AVG	0.01	0.02	0.03	0.04	0.05	AVG
B-Global [18, 54]	15.70	22.02	28.72	34.74	38.86	28.01	9.90	23.04	31.06	38.00	43.26	29.05
B-LSTM [55]	8.08	12.36	15.46	17.39	19.14	14.49	10.04	15.29	19.58	23.20	26.44	18.91
DGL [6]	39.00	49.45	55.25	58.96	61.30	52.85	23.87	35.50	44.01	50.35	55.12	41.62
RGL [7]	48.85	56.02	62.02	64.71	65.41	59.40	30.68	42.80	48.48	54.80	59.62	47.28
Score-PA [15]	44.51	56.14	61.33	64.81	67.11	58.80	30.32	38.63	46.66	52.36	56.58	43.96
IET [8]	37.60	38.37	53.65	57.43	59.39	49.29	25.44	37.46	44.97	50.14	54.56	42.51
3DHPA [17]	63.01	74.91	79.76	82.93	84.71	77.06	48.28	59.97	67.16	71.80	75.12	64.47
Ours	**69.24**	**79.07**	**83.64**	**86.36**	**88.10**	**81.28**	**49.20**	**61.66**	**68.79**	**73.06**	**75.95**	**65.73**

Table B.2: Comparisons on Table category of PA / CA at Chamfer distance thresholds of 0.01 to 0.05.

Method	PA(%) ↑						CA(%) ↑					
	0.01	0.02	0.03	0.04	0.05	AVG	0.01	0.02	0.03	0.04	0.05	AVG
B-Global [18, 54]	15.37	43.56	49.77	54.69	59.36	44.55	33.84	35.98	44.10	50.58	56.86	44.27
B-LSTM [55]	10.55	13.80	16.10	18.55	20.58	15.92	18.28	26.27	32.64	38.21	43.43	31.77
DGL [6]	49.51	56.21	62.44	66.74	70.28	60.46	39.96	50.03	57.68	62.72	66.50	54.94
RGL [7]	55.13	66.68	69.81	73.65	75.90	53.05	41.41	51.80	58.60	64.19	66.57	56.51
Score-PA [15]	52.78	61.68	67.38	71.48	74.72	65.61	40.59	47.96	55.29	60.16	64.62	53.72
IET [8]	48.86	55.99	60.64	64.37	66.84	59.34	40.35	50.66	57.23	62.26	65.84	55.27
3DHPA [17]	64.58	72.62	76.88	79.69	81.89	75.06	58.00	67.69	73.23	77.01	79.79	71.00
Ours	**68.48**	**76.17**	**80.10**	**83.07**	**85.04**	**78.57**	**58.51**	**68.40**	**74.15**	**77.88**	**80.24**	**71.84**

Table B.3: Comparisons on Lamp category of PA / CA at Chamfer distance thresholds 0.01 to 0.05.

Method	PA(%) ↑						CA(%) ↑					
	0.01	0.02	0.03	0.04	0.05	AVG	0.01	0.02	0.03	0.04	0.05	AVG
B-Global [18, 54]	22.61	31.38	38.85	42.56	49.77	37.03	18.60	35.79	46.13	52.42	58.74	42.34
B-LSTM [55]	29.59	38.80	45.78	50.13	54.47	43.75	28.31	38.83	44.85	49.45	54.60	43.21
DGL [6]	33.33	44.83	51.45	56.71	59.91	49.17	41.70	50.60	60.36	67.23	71.08	57.52
RGL [7]	35.54	46.40	51.68	56.95	60.05	50.12	50.09	56.23	64.43	67.53	72.08	62.07
Score-PA [15]	34.32	44.71	51.04	56.58	60.14	49.06	49.07	57.24	65.42	71.38	76.29	62.55
IET [8]	32.86	43.15	49.56	56.14	59.73	48.29	52.75	57.65	66.52	70.87	74.56	64.47
3DHPA [17]	33.67	42.42	49.97	55.89	60.41	48.47	62.01	72.32	78.09	82.13	84.52	75.81
Ours	**36.35**	**47.81**	**54.91**	**60.15**	**64.68**	**52.84**	**62.45**	**74.02**	**78.64**	**82.67**	**85.12**	**76.58**

B.3 Details on Ablation Study in the Main Paper

Details for CFPA-w/o-SP In Table 4 of the main paper, we conduct an ablation study to assess the contribution of semantic super-parts. In the CFPA-w/o-SP variant, the coarse pose estimation stage is performed without leveraging super-part. Instead, the dual-range feature propagation is applied directly to the raw part-level features extracted from the input point clouds by PointNet [49].

Details for CFPA-GE-SP In the experiment of CFPA-GE-SP, we construct the super-part based on geometric similarity, and the parts with similar axis-aligned bounding boxes are considered to belong to the same super-part. The feature of each super-part is obtained by applying max-pooling over the features of its constituent parts which are extracted by PointNet [49].

Details for CFPA-KM-SP For the model of CFPA-KM-SP, we employ K-means clustering [57] on all the part features extracted by PointNet [49]. The process begins by randomly initializing M clustering centroids and then assigning each part feature to its nearest centroid with Euclidean distance. Each centroid is then iteratively updated by averaging the features assigned to it.

Details for CFPA-w/o-SS In Table 6 of the main paper, we perform an ablation study to evaluate the impact of the symmetry-aware loss, particularly its treatment of self-symmetry and inter-part geometric similarity. In the CFPA-w/o-SS variant, we disable the self-symmetry supervision and retain only the constraint on geometrical similar parts. Specifically, the symmetry-aware loss is modified to compare the predicted pose of each part against the ground-truth poses of geometrically similar counterparts, defined as:

$$\mathcal{L}_{Sym}^{\text{w/o-SS}} = \sum_{i=1}^{N} \min_{j \in \Omega(i)} L_{pose}(\mathcal{F}(P_i), \{r_j^{\text{GT}}, t_j^{\text{GT}}\}). \tag{23}$$

where $\Omega(i)$ denotes the set of parts geometrically similar to P_i, and $\mathcal{F}(P_i)$ is the predicted pose of part P_i. This variant allows us to isolate and assess the contribution of self-symmetry modeling within the overall symmetry-aware supervision.

Details for CFPA-w/o-GS In the CFPA-w/o-GS variant, we ablate the geometric similarity component of the symmetry-aware loss while only retaining supervision on self-symmetry. Specifically, only symmetric transformations for ground-truth part are considered during training. The loss is:

$$\mathcal{L}_{Sym}^{\text{w/o-GS}} = \sum_{i=1}^{N} \min_{\tau \in \mathcal{Z}_2^3} L_{pose}(\mathcal{F}(P_i), \tau \circ \{r_i^{\text{GT}}, t_i^{\text{GT}}\}). \tag{24}$$

where $\mathcal{Z}_2^3$ represents the set of eight possible axis-flipping transformations, and $\tau \circ \{r_i^{\text{GT}}, t_i^{\text{GT}}\}$ denotes the flipped version of ground-truth part under transformation τ. This formulation evaluates the effect of modeling self-symmetry in isolation, without considering inter-part geometric similarity.

C More Ablation Studies

Effectiveness of Super-Part Numbers. The number of semantic super-parts (M) is a critical factor in the performance of CFPA. A small M may group functionally distinct parts, resulting in the loss of structural detail, and a large M may overemphasize local features while compromising global coherence. The results on Chair category (Figure C.1) indicate that the optimal performance is achieved at $M = 16$.

Effectiveness of Hyper-parameter in Loss Function. The hyper-parameter λ in the overall loss modulates the influence of the symmetry-aware loss term relative to the part-level and shape-level supervision objectives. To investigate its effect, we conduct an ablation study on the Chair category by varying λ and reporting the corresponding performance in Table C.1. The results indicate that setting $\lambda = 0.1$ yields the best performance.

Table C.1: Ablation study on hyper-parameter in loss.

λ	SCD(10^{-2}) $\downarrow$	PA(%) $\uparrow$	CA(%) $\uparrow$
0.1	**0.49**	69.24	**49.20**
1	0.51	68.86	47.85
10	0.52	**69.45**	47.08

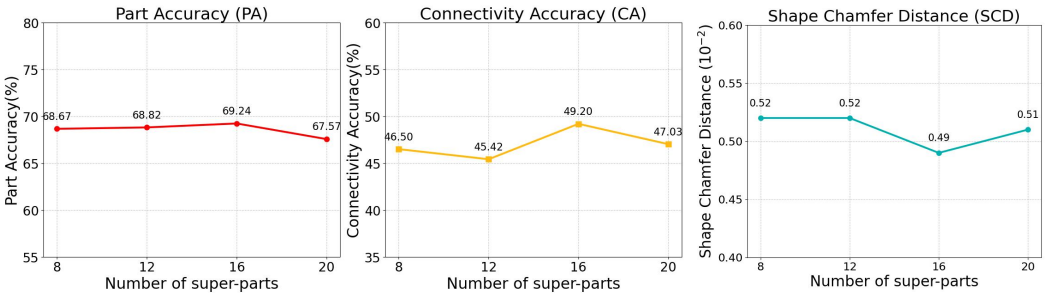

Figure C.1: Performance curves in PA, CA and SCD for CFPA with varying super-parts numbers.

Effectiveness of Graph Construction based on Geometric Similarity In Eq. (5), message passing is performed on a part graph constructed according to geometric similarity, as described in Section A. Parts with similar axis-aligned bounding boxes are grouped into subgraphs, while geometrically unique parts without similar counterparts are also grouped into a fully connected subgraph. To evaluate the effectiveness of this graph construction strategy, we compare two alternative designs: 12) CFPA-w/o-UPC, where geometrically unique parts remain isolated and do not communicate with other parts; 13) CFPA-FC-UP, where each unique part is fully connected to all other parts. The results presented in Table C.2 validate the effectiveness of our proposed balanced graph construction. By preserving geometric similarity through subgraphs and modeling functional relationships among unique parts, CFPA achieves the best performance.

Table C.2: Ablation study on graph construction.

Methods	SCD(10^{-2}) ↓	PA(%) ↑	CA(%) ↑
12) CFPA-w/o-UPC	0.54	68.05	46.77
13) CFPA-FC-UP	0.53	68.78	47.53
CFPA	**0.53**	**68.48**	**48.48**

Performance of the proposed symmetry-aware loss on other baselines To evaluate the generalization ability of the proposed symmetry-aware loss, we conducted experiments by integrating it into various baseline methods. As shown in Table C.3, the models equipped with our symmetry-aware loss (* -w/-SL, where * represents the baselines) achieve improved performance.

These results demonstrate that the symmetry-aware loss is not only effective within our framework but also generalizable across diverse baseline methods. This further validates its robustness and applicability in 3D part assembly tasks.

Table C.3: Comparison of different baselines with our symmetry-aware loss.

Method	SCD↓	PA(%)↑	CA(%)↑	QDS(10^{-5})↑	WQDS(10^{-5})↑
B-Global [18, 54]	1.46	15.70	9.90	0.15	**1.25**
B-Global-w/-SL	**1.21**	**18.10**	**12.92**	**0.43**	1.21
B-LSTM [55]	**2.35**	8.08	10.05	3.92	1.26
B-LSTM-w/-SL	**2.35**	**9.34**	**11.25**	**4.03**	**1.33**
DGL [6]	0.91	39.00	23.87	**1.69**	1.35
DGL-w/-SL	**0.87**	**41.77**	**29.51**	1.66	**1.73**
RGL [7]	0.98	48.85	30.68	5.85	**2.09**
RGL-w/-SL	**0.92**	**49.15**	**33.27**	**5.91**	1.55
Score-PA [15]	0.71	**44.51**	30.32	3.36	**1.70**
Score-PA-w/-SL	**0.65**	43.77	**31.32**	**5.11**	1.41
IET [8]	1.34	37.60	25.44	**3.33**	1.85
IET-w/-SL	**1.21**	**38.53**	**27.21**	**3.33**	**1.93**
3DHPA [17]	**0.51**	63.01	**48.28**	4.42	1.90
3DHPA-w/-SL	**0.51**	**66.34**	47.21	**4.82**	**2.01**

D More Visualization Results

We provide more visualization results for instances from the Chair, Table, and Lamp categories in Figure D.1, Figure D.2, and Figure D.3, respectively. Compared to previous methods, our model generates assembly results that align more closely with ground-truth configurations, indicating superior accuracy. In particular, our approach effectively captures diverse assembly variations (e.g., different arrangements of chair and table legs), showcasing its robustness and adaptability in the 3D part assembly task.

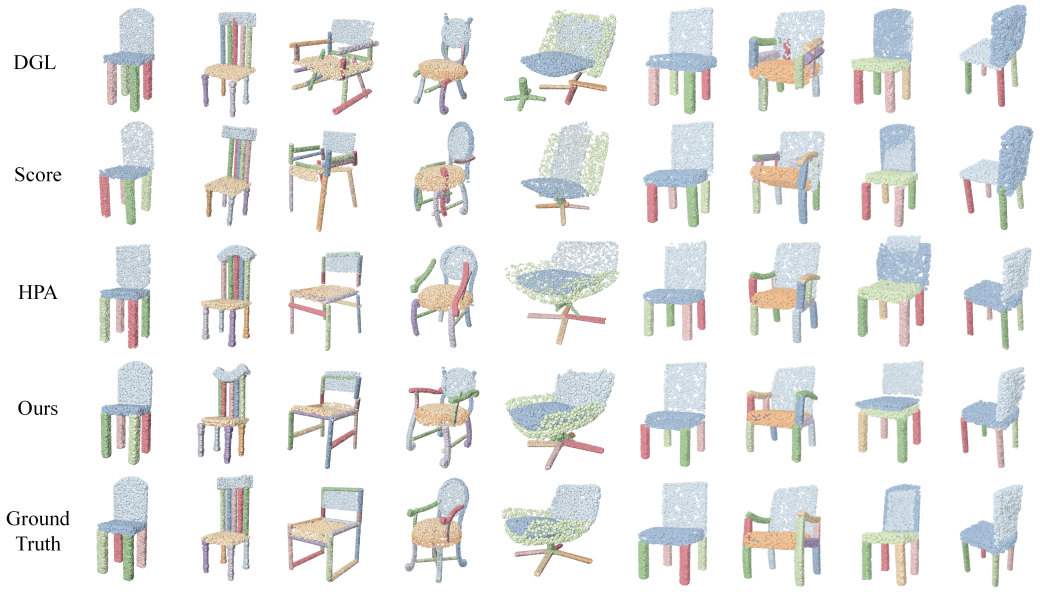

Figure D.1: Qualitative results on the Chair category.

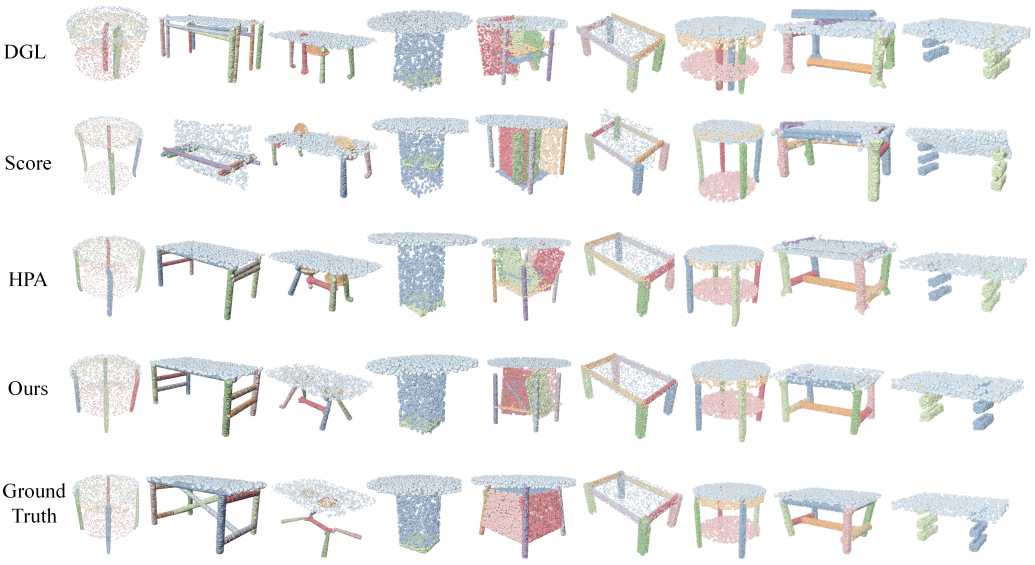

Figure D.2: Qualitative results on the Table category.

Figure D.3: Qualitative results on the Lamp category.