Imagine: Image-Guided 3D Part Assembly with Structure Knowledge Graph

Weihao Wang, Yu Lan, Mingyu You*, Bin He

College of Electronic and Information Engineering, Tongji University
Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University
State Key Laboratory of Intelligent Autonomous Systems, Frontiers Science Center for Intelligent Autonomous Systems,
Shanghai Key Laboratory of Intelligent Autonomous Systems
myyou@tongji.edu.cn

Abstract

3D part assembly is a promising task in 3D computer vision and robotics, focusing on assembling 3D parts together by predicting their 6-DoF poses. Like most 3D shape understanding tasks, existing methods primarily address this task by memorizing the poses of parts during the training process, leading to inaccuracies in complex assemblies and poor generalization to novel categories. In order to essentially improve the performance, structure knowledge of the target assembly is indispensable before assembling, which abstracts the potential part composition and their structural relationships. An image of the target assembly can serve as a common source for constructing this structure knowledge. Nevertheless, the image is far from enough, as its knowledge can be incomplete and ambiguous due to part occlusion and varying views. To tackle these issues, we propose **Imagine**, a novel **Imag**e-guided 3D part assembly framework with structure k**n**owl**e**dge graph. As a novel assembly prior, the structure knowledge graph originates from the image and is refined as understanding the 3D parts. It encodes robust part-aware structural and semantic information of the assembly, guides the 3D parts from a coarse super-structure to a fine assembly, and co-evolves progressively throughout the assembly process. Extensive experiments demonstrate the state-of-the-art performance of our framework, along with strong generalization to novel images and categories.

Introduction

Most 3D objects in our lives are composed of more fine-grained parts. Understanding the delicate relationships of parts is critical for several 3D shape understanding tasks, *e.g.*, assembling a set of 3D parts together by predicting their 6-DoF poses, known as 3D part assembly. Recently, several studies have attempted to address this task by memorizing the poses of parts using elaborate geometric modeling techniques, including graph neural networks (Zhan et al. 2020), attention mechanisms (Zhang et al. 2022; Du et al. 2024), and even the prevalent diffusion model (Cheng et al. 2023). Nevertheless, such rigid memorization exhibits poor accuracy in assembly with dense parts and complex structures, which typically have a large solution space. Moreover, these designs struggle to generalize to assemblies of unseen categories, due to the different structural priors from those they have learned. Even for experienced humans, determining the accurate assembly structure based solely on the geometries of parts without external guidance is challenging.

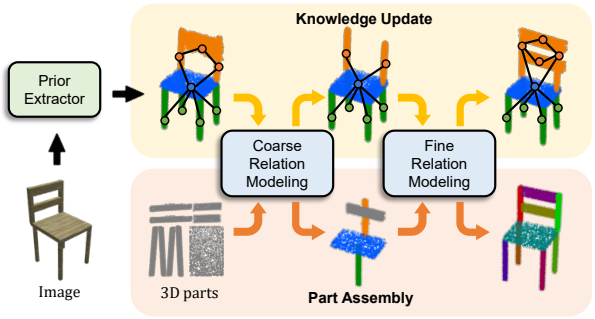

Figure 1: Pipeline of our framework. We derive the structure knowledge graph from the image, which is further refined in a knowledge-assembly co-evolution paradigm.

To address this dilemma, structural knowledge is indispensable. In this work, we refer to structural knowledge as a type of assembly prior that abstracts the potential part composition and approximate structure of the target assembly. This knowledge aids in determining the accurate structure of the assembly from all seemingly possible solutions. More importantly, it provides strong generalizable guidance for constructing unfamiliar structures of unseen categories, rather than blindly copying from memorized experiences. This knowledge can be obtained from various sources. An image of the target assembly serves as a common source which can be conveniently found in manuals or on selling websites.

Despite its convenience, the knowledge derived from an image can be insufficient for achieving accurate and generalizable assembly. First, such 2D structure information can be incomplete due to part occlusion. For example, consider the leg that is completely occluded by the seat in Figure 1. Additionally, parts that are partially occluded and only half-visible in the image can cause confusion among geometrically similar parts. Second, this information lacks robustness because it depends on the view of image. When the view changes, the image can exhibit quite different structure information (*e.g.*, comparing a front view with a back view), leading to confusion and inefficiencies during part assembly.

*Mingyu You is the corresponding author.

These drawbacks motivate us to seek a more robust structure knowledge according to the given 3D parts.

To tackle these issues, we propose **Imagine**, a novel **Imag**e-gu**i**ded 3D part assembly framework with structure k**n**owl**e**dge graph, as illustrated in Figure 1. Similar to how humans imagine the structure of an assembly in their mind, the structure knowledge graph originates from the image and is further refined with feedback from the part assembly process. This is achieved by our designed coarse-to-fine relation modeling, where the knowledge guides the assembly of parts from a super-structure to the final assembly, and co-evolves along this process. In our experiments, the proposed method demonstrates an impressive ability to construct assemblies with high consistency to the image, achieving state-of-the-art performance and robust generalization to novel images and categories.

In summary, our main contributions are:

- Propose a novel reconstruction-based assembly prior, termed the structure knowledge graph, to guide part assembly.
- Design a coarse-to-fine co-evolution framework, where the knowledge and part assembly progressively co-evolve toward the shared goal of accuracy.
- Achieve state-of-the-art performance on both synthetic and realistic datasets, demonstrating impressive generalization to novel images and categories.

Related Work

3D Part Assembly

The emergence of PartNet (Mo et al. 2019b), a large-scale 3D shape dataset with part-level annotations, has significantly advanced several 3D-part-oriented research fields, including 3D part segmentation (Yu et al. 2019; Liu et al. 2023a; Zhou et al. 2023; Umam et al. 2024), 3D shape generation (Li, Liu, and Walder 2022; Tertikas et al. 2023; Li, Paschalidou, and Guibas 2024), and 3D part assembly. The goal of 3D part assembly is to assemble a set of parts by predicting their 6-DoF poses. DGL (Zhan et al. 2020) designs a dynamic graph learning framework to iteratively refine the poses of parts. RGL (Harish, Nagar, and Raman 2022) employs a similar graph representation in a recurrent learning framework to assemble parts sequentially. IET (Zhang et al. 2022) develops a transformer-based framework that incorporates instance encoding for better distinguishing geometrically similar parts. 3DHPA (Du et al. 2024) further extends this framework with hierarchical pose prediction from super-parts to finer parts. The diffusion model (Cheng et al. 2023) has also been applied to the part assembly task, which focuses on diversity of the assembly results.

Despite these increasingly sophisticated geometric modeling designs, these methods struggle to handle complex assemblies and generalize to novel categories. 3DPA (Li et al. 2020) addresses these limitations by using a single image to guide the assembly process, proposing a two-stage framework that first extracts 2D segmentation features conditioned on 3D parts and then integrates this structural information with the 3D parts for pose prediction. However, the segmentation-based structural information can be incomplete due to part occlusion and is also sensitive to changes in view. This limitation motivates us to seek a more robust and efficient structural knowledge in this paper. Beyond using images, Li et al. (Li et al. 2024) introduces joints as additional information to guide part assembly through connectivity modeling.

Single-View 3D Reconstruction

Reconstructing an assembly from its single-view image is a straightforward approach to extracting the structural knowledge of an assembly, which is a fundamental task in visual computing. Existing methods can be broadly classified into two categories: shape-aware and part-aware reconstruction. Shape-aware methods focus on reconstructing the entire object using techniques such as recurrent neural networks (Choy et al. 2016; Zou et al. 2017), encoder-decoder(Fan, Su, and Guibas 2017; Xie et al. 2019; Mittal et al. 2022), diffusion models (Nichol et al. 2022; Jun and Nichol 2023; Liu et al. 2024b). Part-aware methods, on the other hand, aim to reconstruct detailed part instances within the object. StructureNet (Mo et al. 2019a) decomposes an object into a hierarchical structure, containing a set of parts with associated semantics, geometries, and connective and symmetric relationships. This hierarchical structure is learned through an auto-encoding pipeline. VGSnet (Zhang et al. 2021) extends this method by incorporating part segmentation as additional input, and training a latent-aligned image encoder to reconstruct the structure from an image. Pq-net (Wu et al. 2020) introduces a sequential model to progressively generate and assemble the parts, which can be conditioned on an image as input. The recent work of Part123 (Liu et al. 2024a) leverages the powerful multi-view generation method (Liu et al. 2023b) and generalizable image segmentation model SAM (Kirillov et al. 2023) to reconstruct part-aware 3D objects within a neural-rendering-based framework (Mildenhall et al. 2021). Our proposed structure knowledge graph builds upon the hierarchical structure of StructureNet (Mo et al. 2019a), which provides rich knowledge, including semantic and structural information, of the assembly.

Method

Overview

In this work, we address the task of 3D part assembly with the aid of an additional image. Formally, given a set of 3D parts $P = \{p_1, ..., p_N\}$ and an image $I \in \mathbb{R}^{H\times W\times C}$ depicting the target assembly, the goal is to predict the 6-DoF poses $\{(t_1, r_1), ..., (t_N, r_N)\}$ for these parts. Here, N denotes the number of parts, and each part is represented by a point cloud $p \in \mathbb{R}^{1000\times 3}$. The 6-DoF pose for each part consists of a translation $t \in \mathbb{R}^3$ and a rotation $r \in \mathbb{R}^4$, represented by unit quaternion. Finally, the parts are assembled according to their poses into the complete assembly $\mathcal{A} = \cup_{i=1}^{N} \mathcal{T}_i(p_i)$, where $\mathcal{T}_i$ denotes the joint transformation of (t_i, r_i).

The pipeline of our framework, Imagine, is illustrated in Figure 1. In brief, Imagine first reconstructs a structure knowledge graph with the prior extractor, and then integrates this knowledge into the part assembly process through coarse-to-fine relation modeling. 3D parts are progressively assembled

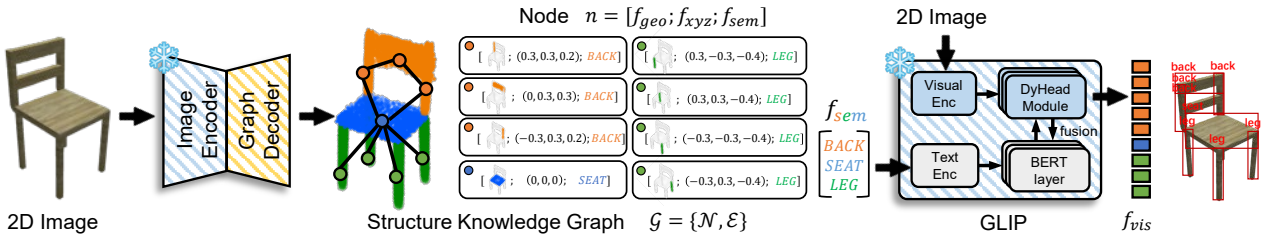

Figure 2: Illustration of the reconstruction-based prior extractor. The structure knowledge graph depicts the structure of assembly by a set of parts with geometries, semantics and spatial relationships, supplemented by semantic visual feature.

from a super-structure to the final assembly, with the knowledge being co-updated throughout this process. The following sections provide detailed explanations.

Prior Extractor

With the advancements in computer vision, numerous off-the-shelf methods are available to extract structure knowledge from an image. Previous work (Li et al. 2020) extracts 2D part segmentation conditioned on 3D parts to guide assembly. However, as discussed earlier, this 2D structure knowledge can be inefficient due to issues like part occlusion and changing views. Recall the human assembly process, when given an image of the assembly, humans can imagine or mentally reconstruct its potential part composition and approximate structure, and then progressively arrange the parts according to this prior knowledge to achieve the final assembly. Inspired by this process, we propose a novel assembly prior called structure knowledge graph. The definition and construction of this graph are detailed in the following paragraphs.

Definition. The structure knowledge graph is defined as $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where $\mathcal{N}$ denotes the part nodes and $\mathcal{E}$ denotes their connections. Each part node $n \in \mathcal{N}$ consists of three components: (1) 3D geometry $f_{geo} \in \mathbb{R}^{1000\times3}$, represented by 1000 points; (2) Spatial center $f_{xyz} \in \mathbb{R}^3$, represented by a 3D coordinate, which is defined within the reconstruction coordinate system and captures the relative spatial relationships of the parts; (3) Semantic label $f_{sem} \in \mathbb{R}^1$, indicating the semantic category of the part, such as back, seat, or leg. Figure 2 visualizes an example.

Construction. The construction of structure knowledge graph builds upon StructureNet (Mo et al. 2019a). StructureNet designs a graph encoder to encode the hierarchical structure $\mathcal{S}$ of an assembly into a latent code $z_{\mathcal{S}}$, and a graph decoder to reconstruct $\mathcal{S}$ from $z_{\mathcal{S}}$, representing the process of auto-encoding. This hierarchical structure $\mathcal{S}$ decomposes the assembly into parts of different granularity, capturing their geometries, semantics, and connective and symmetric relationships. Since our focus is on the finest-grained parts, we retain the leaf nodes of $\mathcal{S}$ with their geometries and semantics to form our structure knowledge graph $\mathcal{G}$.

To reconstruct this hierarchical structure $\mathcal{S}$ from image, we pretrain a ResNet-based image encoder to extract a latent code z_I from image and align it with z_S by minimizing an $\mathcal{L}1$ loss. During inference, we decode z_I using the graph decoder of StructureNet to obtain $\mathcal{S}$. Due to page limitations, a detailed explanation of the pretraining process is provided in the supplementary material.

Visual feature. The reconstructed structure may be locally distorted, such as the horizontal back part shown in Figure 2. To address this, we incorporate the semantic visual features of potential parts in the image as supplementary information. Specifically, we employ the grounded language-to-image pretrained model GLIP (Li et al. 2022), which aligns detected objects with a set of language promotes. In our approach, we use the unique semantic labels in $\mathcal{G}$ as language promotes (*e.g.*, [BACK, SEAT, ARM, LEG] for a chair) to query the visual features $f_{vis} \in \mathbb{R}^{N_I\times256}$ of all potential parts in the image, where N_I denotes the number of detected parts. A brief architecture of GLIP is depicted in Figure 2. We use the feature from the last layer of the DyHeadModule as f_{vis}.

In summary, the structure knowledge graph provides an approximate target for assembly, containing part-level structural and semantic information. In the following section, we study how to leverage this powerful prior to guide part assembly.

Knowledge-Assembly Co-Evolution

As illustrated in Figure 1, we design a co-evolution process with knowledge update and part assembly, where the structure knowledge graph is refined throughout the assembly process. This is achieved by the proposed coarse-to-fine relation modeling. In coarse relation modeling, 3D parts are organized into a super-structure through geometry matching with the knowledge. In fine relation modeling, we further refine the poses of parts by modeling their geometric and semantic relationships. Details of these two modules are described as following.

Coarse relation modeling. Given a set of parts P, we design a bipartite part matching algorithm to search for the optimal subset $P^* \subseteq P$ whose geometries best matches with the part nodes $\mathcal{N}$ in the structure knowledge graph. This is achieved by minimizing the following cost function:

$$P^* = \arg\min_{p_i^* \in P} \sum_i \mathcal{L}_m\left(f_i, p_i^*\right), \tag{1}$$

where $\mathcal{L}_m$ denotes the pairwise matching cost that measures the geometric similarity between the 3D part p_i^* and the geometry of part node f_i in the graph. We implement $\mathcal{L}_m$ using the Chamfer distance, computed between two point clouds $\mathcal{X}$ and $\mathcal{Y}$:

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

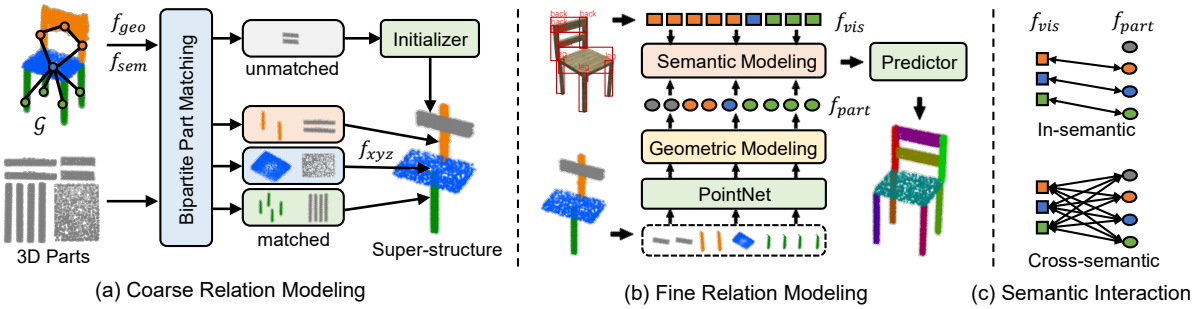

Figure 3: Illustrations of (a-b) the coarse-to-fine relation modeling and (c) the semantic interaction process.

P^* denotes the matched 3D parts, each of which corresponds to a part node in the structure knowledge graph. In return, these correspondences provide each matched 3D part with additional information, including a spatial center f_{xyz} and a semantic label f_{sem} from the structure knowledge graph. To ensure the accuracy of correspondences, we impose two additional constrains on the matching result. First, geometry consistency: the Chamfer distance between each pair of matched parts $d_c(f_i, p_i^*)$ must be smaller than a threshold of 0.01. Second, semantic consistency: geometrically identical parts must share the same semantic label. Geometric identity among parts is identified using instance encoding from IET (Zhang et al. 2022). Any part that does not meet these constrains will be removed from P^*.

After matching, we initialize the matched parts P^* using their spatial centers. Since these centers are reconstructed within the reconstruction coordinate system, their scale may differ from that of the actual assembly. Therefore, we initialize geometrically identical parts in P^* with their average spatial center, as illustrated by the colorful parts of the super-structure in Figure 3(a). For the unmatched parts, we also design an initializer to predict a center for each group of geometrically identical parts (colored in gray). The initializer consists of a PointNet (Qi et al. 2017), two self-attention layers and a center predictor composed of two fully-connected layers and a linear projection layer.

Fine relation modeling. After coarse relation modeling, the 3D parts are organized into a super-structure. In fine relation modeling, we further refine the poses of parts by modeling their geometric and semantic relationships.

We use the PointNet to extract per-part geometric feature and design a geometry modeling module to capture their geometric relations through self-attention of $\text{Softmax}\left(\mathbf{Q}\mathbf{K}^T/\sqrt{d_k}\right)\mathbf{V}$, where $Q, K, V \in \mathbb{R}^{N\times 256}$ denote query, key and value vectors derived from the geometric features of the parts. To distinguish geometrically similar parts, we adopt instance encoding (Zhang et al. 2022) as the position encoding in the computation of attention mechanism. The part features after geometric modeling are referred to f_{part} in the figure.

In addition to geometry, semantic information from the image also provides in- and cross-semantic guidance for refining the detailed structure. For example, when assembling leg parts, the model should focus on the regions corresponding to legs in the image. To this end, we design a semantic modeling module, which consists of hierarchical cross-attention layers. In the first layer, the matched parts P^* compute in-semantic attention with visual features of the same semantic category, *e.g.*, leg parts interact with visual features of legs. This process is illustrated in Figure 3(c). In the subsequent layers, we compute standard attention across different semantic categories. In semantic modeling, we use geometric features as the query and visual features as the key and value.

Finally, we design a predictor to estimate the 6-DoF poses for the parts, which consists of two fully-connected layers and a linear projection layer. The predicted rotation is normalized into a unit quaternion.

Structure knowledge update. The structure knowledge graph may be inaccurate due to the limitations of the reconstruction process. Fortunately, our framework offers the possibility to enhance this knowledge. This enhancement can be achieved by removing unmatched nodes in coarse relation modeling and adding unmatched 3D parts with their predicted poses in fine relation modeling. We illustrate this process with examples in Figure 7.

Training and Losses

During the training of 3D part assembly, geometrically identical parts can have multiple correct poses, *e.g.*, the four identical legs of a chair can interchange their poses. Therefore, we perform bipartite part matching between the predicted poses and ground-truth poses for the geometrically identical parts, searching for the best matched prediction of translation t and rotation r.

After matching, we use Euclidean loss to supervise the predicted translation t of each part with its ground-truth translation $\hat{t}$, defined as:

$$\mathcal{L}_t = \sum_{i=1}^{N} \|t_i - \hat{t}_i\|_2^2. \tag{3}$$

We adopt the Chamfer distance defined in Equation 2 to supervise rigid rotation r:

$$\mathcal{L}_r = \sum_{i=1}^{N} d_c(r_i(p_i), \hat{r}_i(p_i)), \tag{4}$$

where p_i denotes the part point cloud, $r_i(p_i)$ denotes the rotated point cloud using the predicted rotation r and $\hat{r}_i(p_i)$ using the ground-truth rotation $\hat{r}$.

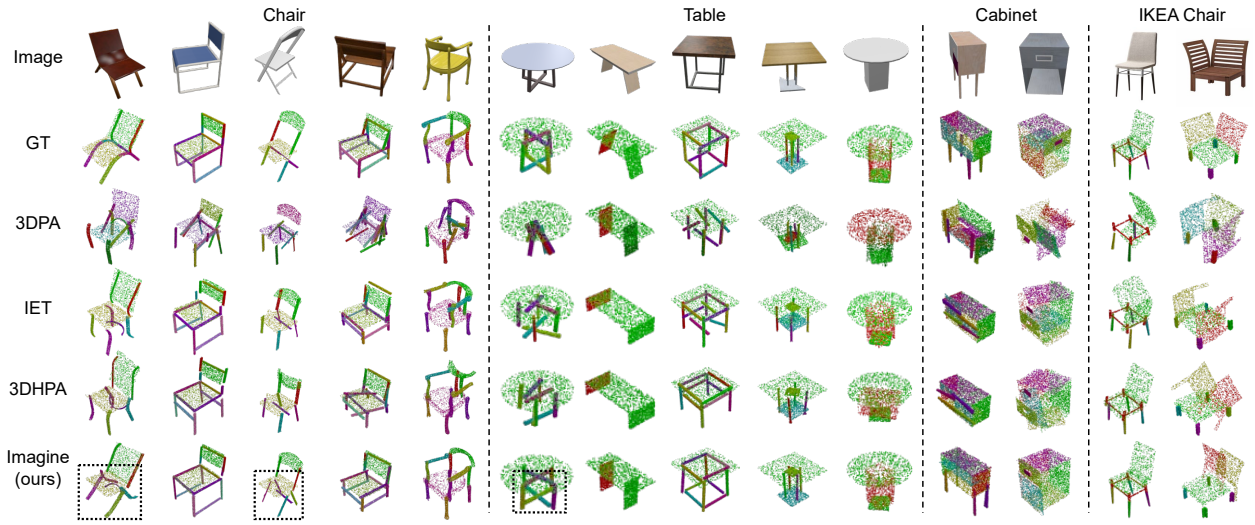

Figure 4: Visualization results on three main furniture categories of PartNet dataset. The assembly results of Imagine are more consistent with the input images. Different colors indicate different parts. Zoom in for better visualization.

Method	SCD(10^{-2}) ↓				PA(%) ↑				CA(%) ↑			
	Chair	Table	Cabinet	IChair	Chair	Table	Cabinet	IChair	Chair	Table	Cabinet	IChair
3DPA (Li et al. 2020)	0.832	0.481	0.599	1.730	46.62	59.27	43.72	12.94	29.74	46.57	32.26	18.64
IET (Zhang et al. 2022)	0.552	0.365	0.389	1.097	62.78	61.18	54.95	10.26	45.44	55.67	50.25	20.34
3DHPA (Du et al. 2024)	0.534	0.343	0.368	1.299	63.23	64.80	55.13	6.27	47.92	59.09	54.01	14.26
Imagine (ours)	**0.421**	**0.252**	**0.235**	**0.854**	**65.88**	**65.13**	**56.35**	**16.81**	**49.81**	**59.89**	**54.80**	**25.67**

Table 1: Part assembly results on PartNet and IKEA-Manual. 'IChair' denotes evaluation on the chairs from IKEA-Manual.

Then we compute the shape Chamfer distance to supervise the entire assembly:

$$\mathcal{L}_s = d_c(\mathcal{A}, \hat{\mathcal{A}}), \tag{5}$$

where $\mathcal{A}$ and $\hat{\mathcal{A}}$ denote the predicted assembly and the ground-truth assembly, respectively.

In coarse relation modeling, we predict a center c for each group of geometrically identical unmatched parts, and we supervise this center with the ground-truth center $\hat{c}$ by:

$$\mathcal{L}_{sp} = \sum_i \|c_i - \hat{c}_i\|_2^2. \tag{6}$$

The overall loss is defined as a weighted combination of these loss items:

$$\mathcal{L} = \lambda_t \mathcal{L}_t + \lambda_r \mathcal{L}_r + \lambda_s \mathcal{L}_s + \lambda_{sp} \mathcal{L}_{sp}. \tag{7}$$

Experiments

Dataset

We conduct experiments on PartNet (Mo et al. 2019b), which contains 26,671 3D objects with 573,585 part instances annotated, covering 24 different categories. Following (Li et al. 2020), we choose the major categories of furniture, chair, table and cabinet, for experiment. Data splits are set to 70%/20%/10% for train/val/test. The original parts are represented by triangle meshes. We sample 1000 points for each part using Farthest Point Sampling (Moenning and Dodgson 2003) and apply PCA to transform the parts into the canonical space as input. We adopt the RGB images from (Uy et al. 2021), where each object is rendered with a resolution of 224×224 from 24 different views.

Implementation Details

We set a maximum number of parts to 20, and randomly select an image as input from 24 views. The model is trained on each category for 500 epochs with a batch size of 64 on 4 V100 GPUs. We adopt the AdamW optimizer with an initial learning rate of 0.00015 and a weight decay of 0.0001. The number of attention layers in geometric and semantic modeling is set to 3. The loss weights are set by $\lambda_t = \lambda_s = \lambda_{sp} = 1$, $\lambda_r = 10$. We use the part-aware GLIP model finetuned by PartSLIP (Liu et al. 2023a). Weights of image encoder, StructureNet decoder and GLIP model are frozen in training. Details of these pretrained models can be found in the supplementary material.

Metrics

Following (Zhang et al. 2022), we use Shape Chamfer Distance (SCD), Part Chamfer Distance (PCD), Part Accuracy (PA) and Connectivity Accuracy (CA) to evaluate the quality of assembly. SCD and PCD are defined in Equation 5, with SCD computed per shape and PCD per part. PA measures the

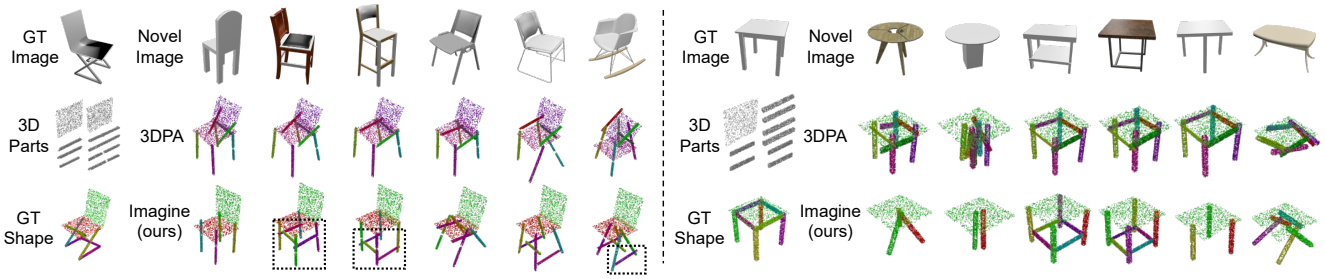

Figure 5: Visualization of in-category generalization. Given a set of parts, we show its original assembly and image in the left column. The right columns demonstrate the generated assemblies with different images. Zoom in for better visualization.

Method	Refrigerator				Dishwasher				Microwave			
	SCD↓	PCD↓	PA↑	CA↑	SCD↓	PCD↓	PA↑	CA↑	SCD↓	PCD↓	PA↑	CA↑
IET-A	1.234	19.69	1.74	20.17	0.719	27.23	1.86	16.25	0.732	12.21	2.04	10.67
3DPA-A	1.302	20.49	1.89	21.82	0.787	22.65	4.13	14.23	**0.511**	11.35	4.72	**26.17**
Imagine-A (ours)	**0.397**	**3.454**	**18.85**	**24.65**	**0.420**	**6.125**	**14.51**	**21.65**	0.515	**4.114**	**8.03**	17.10

Table 2: Evaluation on unseen categories. Suffix 'A' indicates the model trained on all three categories of chair, table and cabinet.

Method	Chair → Table			Chair → Cabinet		
	SCD↓	PCD↓	PA↑	SCD↓	PCD↓	PA↑
3DPA	2.103	8.121	15.16	3.982	14.17	8.28
IET	1.839	14.68	9.52	4.692	35.55	1.84
3DHPA	2.621	18.44	5.71	4.533	37.31	1.08
Imagine	**1.465**	**1.347**	**16.56**	**3.386**	**2.466**	**9.35**

Table 3: Evaluation on unseen categories. We evaluate model trained on category a on another category b, short for $a \rightarrow b$. SCD and PCD are scaled by 10^{-2}.

percentage of correctly predicted parts with a PCD of less than 0.01. Similarly, CA measures the percentage of correctly matched contact points in the predicted assembly.

Comparison with the State-of-the-art Methods

Results on PartNet. We compare our method, Imagine, with several state-of-the-art methods focused on 3D part assembly with (3DPA) and without image (IET, 3DHPA). The quantitative results are summarized in Table 1. As shown, Imagine outperforms all baseline methods, especially in SCD across each category, indicating a higher fidelity of assembly. To support this analysis, we provide additional visualization results in Figure 4. Most cases involve either dense parts or complex structures. As shown, Imagine is excels at restoring the local structure of assembly according to the image, such as the crossed legs marked by black dotted boxes.

Results on realistic dataset. IKEA-Manual (Wang et al. 2022) contains 102 assemblies collected from real IKEA furniture. We further collect realistic photos of these furniture from the Internet, and evaluate chair models on 57 chairs from this real-world dataset. The results refer to 'IChair' in Table 1 and Figure 4. In comparison, our method demonstrates robust performance when using real furniture and realistic photos as input, even with rare structures, such as the chair in the last column.

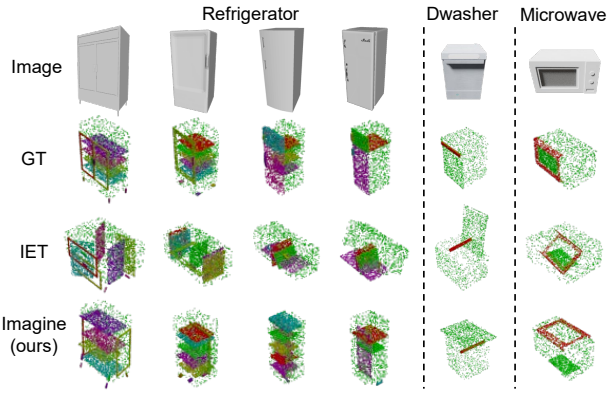

Figure 6: Visualization of unseen-category generalization.

In-category generalization. In image-guided assembly, the model must understand the assembly structure according to the image, rather than simply relying on memorizing poses. To evaluate this, we design an in-category generalization experiment. Given a fixed set of parts, we evaluate our model and 3DPA, another image-guided baseline, by changing the input image. The results are visualized in Figure 5. As shown, our method better adapts to changes in the image, generating structures that are consistent with the description of image. For example, the spatial layout of the vertical legs and horizontal bars of the visualized chair demonstrates flexible assembly plans, while 3DPA generates homogeneous results and fails to generalize in this scenario.

Unseen-category generalization. We further evaluate the generalization ability of our method on unseen categories. Initially, we use the chair model to assemble tables and cabinets. As shown in Table 3, Imagine performs robustly on tables but experiences a significant drop in cabinet. This is

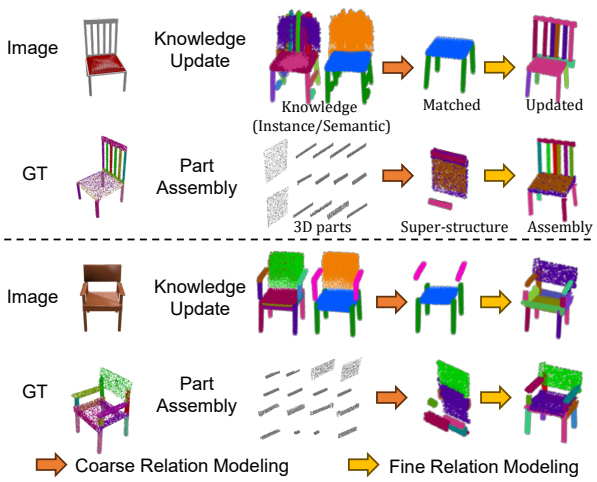

Figure 7: Visualization of the progressive knowledge update and part assembly in coarse-to-fine relation modeling.

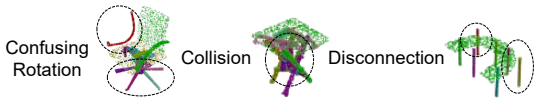

Figure 8: Analysis of failed cases.

because the knowledge learned from only one category is too limited to generalize effectively to novel categories with different structural priors. To address this, we subsequently train a model on all three categories and introduce additional novel categories for evaluation. The results, shown in Table 2 and Figure 6, demonstrate that by learning from more heterogeneous structural knowledge, our model exhibits decent generalization on these unseen categories. In contrast, IET blindly copies its memorized poses to novel cases, resulting in table-like fridges and a chair-like dishwasher.

Analysis of Knowledge-Assembly Co-Evolution

We assemble the 3D parts and update the structural knowledge graph progressively, moving from coarse to fine modeling. To illustrate this procedure, we provide two cases with dense parts in Figure 7. In coarse relation modeling, the seat and leg parts achieve better matching. The chair above provides a clear demonstration of knowledge updating, where the initially noisy knowledge, particularly regarding the back parts, becomes more accurate through this process. During fine relation modeling, the parts differentiate into different positions and construct the final assembly, which compensates the knowledge graph with the unmatched parts in return, *e.g.*, update of the back parts of both chairs in the knowledge.

Ablation Study

Loss Components. We investigate the impact of each loss component in Equation 7 by removing them individually. As shown in Table 4, the translation loss $\mathcal{L}_t$ is crucial for learning the spatial layout of parts, as its absence leads to a significant drop in performance. The rotation loss $\mathcal{L}_r$ and shape Chamfer distance loss $\mathcal{L}_{scd}$ both play important roles in refining the local and global structure of the assembly. Additionally, center supervision of the unmatched parts $\mathcal{L}_{sp}$ during coarse relation modeling is also beneficial for constructing more accurate assemblies.

Ablation	SCD(10^{-2}) ↓	PA(%) ↑	CA(%) ↑
w/o $\mathcal{L}_t$	0.655	18.92	18.91
w/o $\mathcal{L}_r$	0.431	63.09	44.68
w/o $\mathcal{L}_{scd}$	0.481	63.59	46.06
w/o $\mathcal{L}_{sp}$	0.433	64.25	45.57
w/o $\mathcal{G}$	0.453	61.27	45.26
w/o f_{vis}	0.467	62.05	46.31
Imagine	**0.421**	**65.88**	**49.81**

Table 4: Function of the losses and designed modules.

f_{vis}	SCD(10^{-2}) ↓	PA(%) ↑	CA(%) ↑
Pretrained ViT	0.436	64.72	47.99
2D Segmentation	0.449	62.52	46.13
GLIP	**0.421**	**65.88**	**49.81**

Table 5: Effect of different visual features.

Visual feature. We evaluate the function of the semantic-grounded visual feature f_{vis} by removing it during fine relation modeling, referring to w/o f_{vis} in Table 4. As shown, part accuracy decreases by 3.83%. To further demonstrate its efficiency, we replace the visual feature with those from a pretrained tiny ViT (Dosovitskiy et al. 2020) and the 2D part segmentation feature used by 3DPA. As shown in Table 5, the model achieves a competitive performance with the general visual feature from ViT , but shows little improvement with the segmentation feature, as this instance-aware segmentation lacks the semantic guidance of parts.

Conclusion

In this work, we emphasize the importance of structure knowledge in 3D part assembly. To this end, we extract a novel structure knowledge graph from a single image, which guides part assembly within a coarse-to-fine learning paradigm. Notably, this design enables the model to construct assemblies with high structural consistency relative to the image and to better generalize to novel images and categories. We hope this approach will inspire related research fields, such as image-guided 3D shape segmentation, retrieval, and deformation.

Limitation and future work. We observe several cases where Imagine fails to determine the correct rotations of irregular parts, and presents part collision and disconnection, as demonstrated in Figure 8. Additionally, the structure knowledge knowledge derived from a single-view image may be insufficient for the model to generalize to novel categories with significant semantic and structural differences, such as cars and airplanes. In future work, we plan to introduce stronger external guidance, such as sequential manuals and instructional videos, and to construct collision-free and well-connected assemblies.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62473290, 62073244 , 61825303, 62088101, and the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100) and the Fundamental Research Funds for the Central Universities.

References

Cheng, J.; Wu, M.; Zhang, R.; Zhan, G.; Wu, C.; and Dong, H. 2023. Score-pa: Score-based 3d part assembly. *arXiv preprint arXiv:2309.04220*.

Choy, C. B.; Xu, D.; Gwak, J.; Chen, K.; and Savarese, S. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, 628–644. Springer.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, B.; Gao, X.; Hu, W.; and Liao, R. 2024. Generative 3D Part Assembly via Part-Whole-Hierarchy Message Passing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20850–20859.

Fan, H.; Su, H.; and Guibas, L. J. 2017. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 605–613.

Harish, A. N.; Nagar, R.; and Raman, S. 2022. Rgl-net: A recurrent graph learning framework for progressive part assembly. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 647–656. IEEE.

Jun, H.; and Nichol, A. 2023. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

Li, S.; Liu, M.; and Walder, C. 2022. EditVAE: Unsupervised parts-aware controllable 3D point cloud shape generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1386–1394.

Li, S.; Paschalidou, D.; and Guibas, L. 2024. PASTA: Controllable Part-Aware Shape Generation with Autoregressive Transformers. *arXiv preprint arXiv:2407.13677*.

Li, Y.; Mo, K.; Duan, Y.; Wang, H.; Zhang, J.; and Shao, L. 2024. Category-level multi-part multi-joint 3d shape assembly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3281–3291.

Li, Y.; Mo, K.; Shao, L.; Sung, M.; and Guibas, L. 2020. Learning 3d part assembly from a single image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 664–682. Springer.

Liu, A.; Lin, C.; Liu, Y.; Long, X.; Dou, Z.; Guo, H.-X.; Luo, P.; and Wang, W. 2024a. Part123: Part-aware 3D Reconstruction from a Single-view Image. In *ACM SIGGRAPH 2024 Conference Papers*, 1–12.

Liu, M.; Xu, C.; Jin, H.; Chen, L.; Varma T, M.; Xu, Z.; and Su, H. 2024b. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36.

Liu, M.; Zhu, Y.; Cai, H.; Han, S.; Ling, Z.; Porikli, F.; and Su, H. 2023a. Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21736–21746.

Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2023b. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Mittal, P.; Cheng, Y.-C.; Singh, M.; and Tulsiani, S. 2022. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 306–315.

Mo, K.; Guerrero, P.; Yi, L.; Su, H.; Wonka, P.; Mitra, N.; and Guibas, L. J. 2019a. Structurenet: Hierarchical graph networks for 3d shape generation. *arXiv preprint arXiv:1908.00575*.

Mo, K.; Zhu, S.; Chang, A. X.; Yi, L.; Tripathi, S.; Guibas, L. J.; and Su, H. 2019b. 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*, 909–918.

Moenning, C.; and Dodgson, N. A. 2003. Fast marching farthest point sampling. Technical report, University of Cambridge, Computer Laboratory.

Nichol, A.; Jun, H.; Dhariwal, P.; Mishkin, P.; and Chen, M. 2022. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Tertikas, K.; Paschalidou, D.; Pan, B.; Park, J. J.; Uy, M. A.; Emiris, I.; Avrithis, Y.; and Guibas, L. 2023. Generating part-aware editable 3D shapes without 3D supervision. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4466–4478.

Umam, A.; Yang, C.-K.; Chen, M.-H.; Chuang, J.-H.; and Lin, Y.-Y. 2024. PartDistill: 3D Shape Part Segmentation by Vision-Language Model Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3470–3479.

Uy, M. A.; Kim, V. G.; Sung, M.; Aigerman, N.; Chaudhuri, S.; and Guibas, L. J. 2021. Joint learning of 3d shape retrieval and deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11713–11722.

Wang, R.; Zhang, Y.; Mao, J.; Zhang, R.; Cheng, C.-Y.; and Wu, J. 2022. Ikea-manual: Seeing shape assembly step by step. *Advances in Neural Information Processing Systems*, 35: 28428–28440.

Wu, R.; Zhuang, Y.; Xu, K.; Zhang, H.; and Chen, B. 2020. Pq-net: A generative part seq2seq network for 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 829–838.

Xie, H.; Yao, H.; Sun, X.; Zhou, S.; and Zhang, S. 2019. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2690–2698.

Yu, F.; Liu, K.; Zhang, Y.; Zhu, C.; and Xu, K. 2019. Partnet: A recursive part decomposition network for fine-grained and hierarchical shape segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9491–9500.

Zhan, G.; Fan, Q.; Mo, K.; Shao, L.; Chen, B.; Guibas, L. J.; Dong, H.; et al. 2020. Generative 3d part assembly via dynamic graph learning. *Advances in Neural Information Processing Systems*, 33: 6315–6326.

Zhang, R.; Kong, T.; Wang, W.; Han, X.; and You, M. 2022. 3d part assembly generation with instance encoded transformer. *IEEE Robotics and Automation Letters*, 7(4): 9051–9058.

Zhang, X.; Ma, R.; Zou, C.; Zhang, M.; Zhao, X.; and Gao, Y. 2021. View-aware geometry-structure joint learning for single-view 3D shape reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6546–6561.

Zhou, Y.; Gu, J.; Li, X.; Liu, M.; Fang, Y.; and Su, H. 2023. PartSLIP++: Enhancing Low-Shot 3D Part Segmentation via Multi-View Instance Segmentation and Maximum Likelihood Estimation. *arXiv preprint arXiv:2312.03015*.

Zou, C.; Yumer, E.; Yang, J.; Ceylan, D.; and Hoiem, D. 2017. 3d-prnn: Generating shape primitives with recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 900–909.