TORA: Topological Representation Alignment for 3D Shape Assembly

Nahyuk Lee[1,*], Zhiang Chen[2,*], Marc Pollefeys[2], and Sunghwan Hong[2,3,†]

[1] Independent Researcher
[2] ETH Zurich
[3] ETH AI Center

Abstract. Flow-matching methods for 3D shape assembly learn pointwise velocity fields that transport parts toward assembled configurations, yet they receive no explicit guidance about which cross-part interactions should drive the motion. We introduce **TORA**, a topology-first representation alignment framework that distills relational structure from a frozen pretrained 3D encoder into the flow-matching backbone during training. We first realize this via simple instantiation, token-wise cosine matching, which injects the learned geometric descriptors from the teacher representation. We then extend to employ a Centered Kernel Alignment (CKA) loss to match the *similarity structure* between student and teacher representations for enhanced topological alignment. Through systematic probing of diverse 3D encoders, we show that geometry- and contact-centric teacher properties, not semantic classification ability, govern alignment effectiveness, and that alignment is most beneficial at later transformer layers where spatial structure naturally emerges. TORA introduces zero inference overhead while yielding two consistent benefits: faster convergence (up to 6.9×) and improved accuracy in-distribution, along with greater robustness under domain shift. Experiments on five benchmarks spanning geometric, semantic, and inter-object assembly demonstrate state-of-the-art performance, with particularly pronounced gains in zero-shot transfer to unseen real-world and synthetic datasets. Project page: `https://nahyuklee.github.io/tora`

Keywords: Topological Representation Alignment · Relational Knowledge Distillation · 3D Shape Assembly

1 Introduction

3D object assembly from unposed part point clouds is a fundamental geometric reasoning task with applications in archaeology [36, 41, 49], computer graphics [3, 22], and robotics [4, 13, 14, 38, 52]. The goal is to estimate per-part rigid transformations that reconstruct an object from its constituent parts. Importantly, assembly spans a broad spectrum: semantic assembly, which arranges

* Equal contribution
† Corresponding author

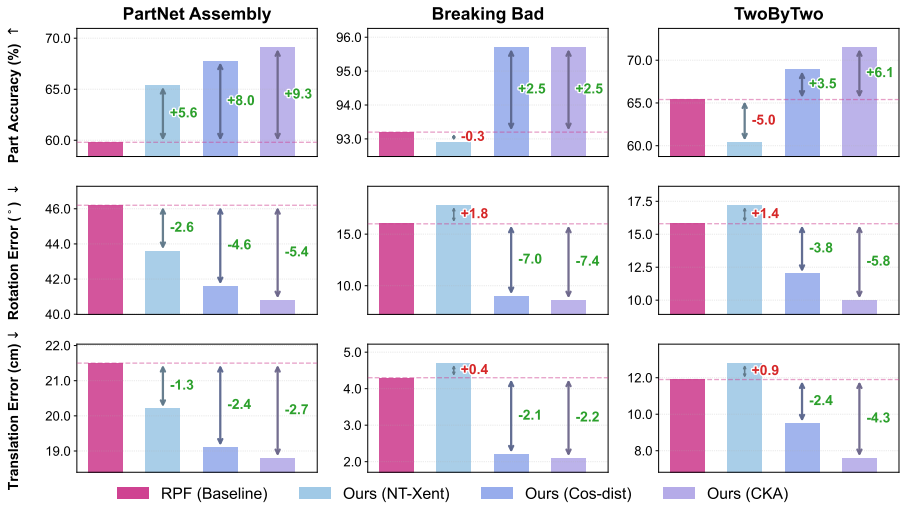

Fig. 1: Multi-part assembly results across regimes. We compare the RPF baseline with our alignment variants. We show that casting the training to a teacher-student distillation for injecting pretrained geometric priors consistently improves performance.

functionally meaningful intra-object parts (*e.g.*, chair legs and backrests), geometric assembly, which reconstructs physically fractured fragments based on surface complementarity, and inter-object assembly, where components from distinct objects must be mated under cross-object constraints (*e.g.*, peg-in-hole insertion). Across this spectrum, the dominant cues vary from part semantics to fine-grained surface compatibility, but a shared bottleneck is discovering *mating relations*—which regions should contact and co-move—robustly under ambiguity and domain shift.

Motivated by this, recent flow-matching methods [29, 42] have made significant progress in this direction. In particular, Rectified Point Flow (RPF) [42] learns a point-wise velocity field that transports noisy point clouds toward assembled configurations, followed by closed-form Procrustes/SVD recovery. While powerful, RPF is trained mainly with an endpoint loss on the final assembled geometry. As a result, the model must infer mating relations implicitly: decisive cues lie on sparse contact regions and are often ambiguous under symmetry, yet the loss does not indicate which cross-part interactions should drive the motion. This lack of explicit intermediate guidance can reduce robustness under distribution shift, motivating the injection of priors that highlight likely interactions.

A practical way to introduce such priors is to cast training in a teacher–student distillation framework, inspired by the 2D generative literature [51] but adapted to 3D point-flow assembly. In our setting, we first show that our simple instantiation, *e.g.,* token-wise cosine matching, is already a strong and effective alignment strategy: by encouraging each point token to match the teacher's feature content, it injects the teacher's learned geometric descriptors into the flow backbone and can implicitly transfer relational cues when the teacher repre-

sentation is well-structured. However, such independent cosine matching treats tokens independently; it does not explicitly constrain the *relational topology* [4] among points, and may therefore yield suboptimal distillation when assembly hinges on interaction structure across parts.

In this paper, we introduce **TORA**, a teacher–student alignment framework for 3D point-flow assembly that injects pretrained geometric priors into our flow matching model via a frozen 3D teacher encoder. We start from the alignment with standard token-wise objectives, including cosine matching and a contrastive NT-Xent variant [28], which aligns student tokens to teacher tokens either by directly matching feature content or by additionally enforcing per-point discriminability through negatives. Extending these baselines, we then proceed to introduce a topology-first objective based on Centered Kernel Alignment (CKA) [8] that directly matches the pairwise *similarity structure* among point tokens, aiming to explicitly preserve the information about who is similar to whom. Finally, we identify two practical choices that matter in this setting—aligning late-layer representations where global geometric structure emerges, and using the right teachers that encode interaction geometry more than category-level semantics.

We demonstrate the effectiveness of TORA on several benchmarks spanning semantic, geometric, and inter-object assembly regimes [38, 39, 48], highlighting the importance of transferring relational structure for robust 3D point-flow assembly. As shown in Fig. 1, TORA achieves state-of-the-art performance across these benchmarks, and further establishes a new state of the art in zero-shot assembly on additional real-world and synthetic datasets [24, 29]. Finally, we provide extensive ablations and analyses to validate our design choices and clarify when each component is most beneficial.

2 Related Work

3D shape assembly. Shape assembly aims to predict per-part rigid transformations that reconstruct a complete object from its constituent pieces [26, 27, 29, 30, 33, 34, 39, 42, 43, 47, 48, 53]. Early learning-based methods formulate this as a correspondence problem [1, 2, 5, 6, 11, 12, 15–20]: given point clouds of individual parts, they first establish point-wise correspondences [26, 27] across fragments, then extract poses via SVD or weighted averaging [26, 27, 34]. While effective for pairwise assembly, correspondence-based approaches face combinatorial challenges when scaling to multi-part scenarios, where establishing reliable correspondences across many irregularly shaped fragments becomes increasingly difficult.

Recent work has shifted toward generative formulations that sidestep explicit correspondence estimation [29, 42, 43]. Flow-matching-based methods [29, 42] learn continuous velocity fields or SE(3) trajectories that transport parts from arbitrary initial poses to their assembled configurations, naturally handling multi-part assembly and part symmetries within a unified probabilistic framework.

[4] We use "topological" to refer to relational inter-token similarity structure (*i.e.*, who is similar to whom), not algebraic topology in the sense of persistence or homology.

Rectified Point Flow (RPF) [42] achieves the current state-of-the-art by learning a point-wise flow conditioned on features from an overlap-aware encoder, demonstrating strong results across both pairwise registration and multi-part assembly benchmarks. However, RPF's encoder is pretrained solely on a binary overlap prediction, a relatively narrow geometric signal, and the flow model itself receives no explicit guidance about the broader spatial structure of the parts it assembles. We show that this geometric understanding can be substantially enriched by distilling representations from pretrained 3D point cloud encoders that have learned richer spatial features from large-scale shape data.

Distillation objectives for generative models. REPA [23, 25, 51] showed that aligning diffusion transformer [37] features to a frozen pretrained visual encoder can substantially accelerate training [50] and improve 2D image generation quality. Subsequent works have expanded this paradigm in several directions: REPA-E [28] enables end-to-end VAE tuning, HASTE [44] augments alignment with attention-based signals and staged termination, and REG [45] introduces discriminative tokens to strengthen denoising. iREPA [40] further improves spatial fidelity via convolutional projection and spatial normalization, while Geometry Forcing [46] adds geometric scale and angular constraints for 3D-consistent video diffusion in world-modeling settings. In contrast, our work studies representation alignment for 3D point-flow models in shape assembly, where the target signal is governed by geometric compatibility and sparse mating relations; this leads to different empirical behavior and motivates alignment objectives and teacher choices tailored to part-level 3D geometry.

3 Method

3.1 Problem Formulation

Given K unposed part point clouds $\{\mathbf{P}_k\}_{k=1}^K$ with $\mathbf{P}_k \in \mathbb{R}^{N_k \times 3}$, and N_k is the number of points in part $k \in \{1, \ldots, K\}$, 3D shape assembly seeks per-part rigid transformations $\{\mathbf{T}_k\}_{k=1}^K$ with $\mathbf{T}_k \in \mathrm{SE}(3)$ such that the transformed parts $\{\mathbf{T}_k \mathbf{P}_k\}$ reconstruct the original object. One part is conventionally fixed as an anchor ($\mathbf{T}_1 = \mathbf{I}$)[5], and the remaining poses are predicted relative to it.

3.2 Preliminary: Flow-matching based 3D shape assembly

Rectified point flow [42]. RPF reformulates this pose estimation problem as conditional generation in 3D Euclidean space. Rather than directly regressing $\{\mathbf{T}_k\}$, RPF learns a continuous flow that transports points from noise to their assembled positions, from which poses are recovered. Concretely, let $\mathbf{x}_k(0) \in \mathbb{R}^{N_k \times 3}$ denote the assembled point cloud for part k (sampled from the ground-truth object) and $\mathbf{x}_k(1) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. RPF defines a straight-line interpolation:

[5] We discuss the practicality and fairness of this choice and provide additional experiments in the supplementary material.

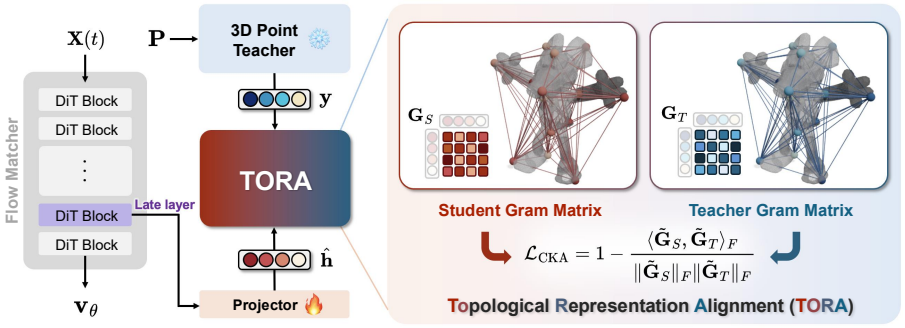

Fig. 2: Overview of the Topological Representation Alignment (TORA) framework. TORA distills relational geometric structures from a frozen 3D foundation teacher into a flow-matching student during training. By matching Gram-based similarity matrices via Centered Kernel Alignment (CKA), the student learns the pairwise "*who-is-similar-to-whom*" relational topology of parts. As detailed in Sec. 5, this structural distillation significantly accelerates convergence and enhances robustness under domain shift, while incurring strictly zero overhead during inference.

$$\mathbf{x}_k(t) = (1-t)\,\mathbf{x}_k(0) + t\,\mathbf{x}_k(1), \quad t \in [0,1], \tag{1}$$

with constant velocity $d\mathbf{x}_k(t)/dt = \mathbf{x}_k(1) - \mathbf{x}_k(0)$. A flow model V is trained to predict this velocity conditioned on the unposed parts $\{\mathbf{P}_k\}_{k=1}^K$, which are first encoded by a pretrained overlap-aware encoder. At inference, the learned flow is integrated from t=1 (noise) to t=0 to recover assembled positions $\hat{\mathbf{x}}_k(0)$, and per-part poses are extracted via Procrustes alignment:

$$\hat{\mathbf{T}}_k = \arg\min_{\mathbf{T}\in \mathrm{SE}(3)} \|\mathbf{T}\,\mathbf{P}_k - \hat{\mathbf{x}}_k(0)\|_F. \tag{2}$$

The training objective supervises the flow through the endpoint reconstruction of $\mathbf{x}_k(0)$, without explicitly specifying which cross-part correspondences or contact regions should drive the flow. Consequently, mating cues—typically sparse and sometimes ambiguous under symmetry—must be discovered implicitly from this global endpoint signal. The flow model processes concatenated point tokens from all parts through a sequence of transformer blocks, producing intermediate representations $\mathbf{h}^{(l)} \in \mathbb{R}^{N\times D}$ at each layer l (where $N = \sum_k N_k$), which we target for alignment.

3.3 TORA: Topological Representation Alignment

Figure 2 illustrates the overall architecture. Our methodology is built on RPF architecture [42], augmenting its flow matcher with a topological representation alignment branch.

Flow matcher. Given K unposed part point clouds $\{\mathbf{P}_k\}_{k=1}^K$, a frozen overlap-aware encoder extracts per-point conditioning features $\mathbf{c} \in \mathbb{R}^{N\times D}$ (see [42] for details). A DiT-based [37] transformer V_θ takes the noisy point positions $\mathbf{X}(t)$ at

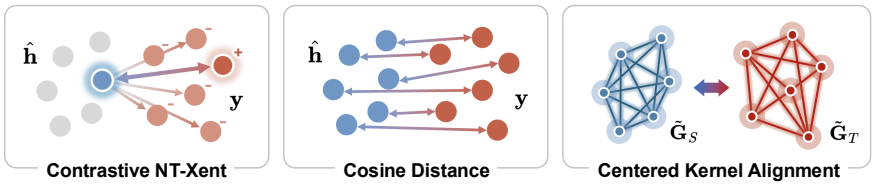

Fig. 3: Conceptual illustration of alignment objectives. Blue and red dots denote student tokens $\hat{\mathbf{h}}$ and teacher tokens $\mathbf{y}$, respectively. NT-Xent enforces per-point discriminability via positive/negative pairing, and cosine distance independently aligns each token pair. CKA objective matches the pairwise similarity structures (Gram matrices $\tilde{\mathbf{G}}_S$, $\tilde{\mathbf{G}}_T$), preserving relational topology rather than individual feature vectors.

timestep t together with $\mathbf{c}$, and predicts a per-point velocity field. The network consists of L transformer blocks; block l produces intermediate representations $\mathbf{h}^{(l)} \in \mathbb{R}^{N\times D}$, where $N = \sum_k N_k$ is the total number of points. Training minimizes the conditional flow matching objective:

$$\mathcal{L}_{\text{CFM}}(V_\theta) = \mathbb{E}_{t,\mathbf{X}}\left[\|V_\theta(t, \mathbf{X}(t) \mid \mathbf{X}) - \nabla_t \mathbf{X}(t)\|^2\right], \tag{3}$$

where $\mathbf{X}(t) = (1-t)\mathbf{x}_1 + t\,\mathbf{x}_0$ interpolates between target assembled positions $\mathbf{x}_1$ and noise $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, so that $\nabla_t \mathbf{X}(t) = \mathbf{x}_0 - \mathbf{x}_1$.

Alignment branch. We introduce a frozen *teacher encoder* f, selected based on the analysis in Sec. 4, which produces target representations from the clean (non-noisy) point clouds, $\mathbf{y} = f(\mathbf{P}) \in \mathbb{R}^{N\times D_f}$. A lightweight projector ϕ (a 3-layer MLP with SiLU activations [21]) maps the flow matcher's intermediate features at a selected layer l^* into the teacher feature space:

$$\hat{\mathbf{h}} = \phi(\mathbf{h}^{(l^*)}) \in \mathbb{R}^{N\times D_f}. \tag{4}$$

The total training loss augments the conditional flow matching objective with an alignment regularizer,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CFM}} + \lambda\,\mathcal{L}_{\text{align}}(\hat{\mathbf{h}}, \mathbf{y}), \tag{5}$$

where $\mathcal{L}_{\text{align}}$ admits multiple instantiations. We organize them along a single axis, namely the explicitness with which teacher relational structure is transferred, and study three objectives of increasing relational explicitness. First, token-wise cosine distance ($\mathcal{L}_{\text{cos-dist}}$) is the simplest baseline, transferring teacher relations implicitly through per-token matching. Second, a contrastive NT-Xent variant ($\mathcal{L}_{\text{NT-Xent}}$) additionally enforces per-token discriminability via negatives. Third, our primary objective, Centered Kernel Alignment ($\mathcal{L}_{\text{CKA}}$), transfers relations explicitly by matching the pairwise similarity (Gram) structure between student and teacher. We treat the two token-wise variants as baselines and ablations, and adopt $\mathcal{L}_{\text{CKA}}$ as the default TORA objective, since it consistently performs best across assembly regimes (Sec. 5). We provide a conceptual illustration of the three objectives in Fig. 3.

Token-wise alignment baselines (Cos-Dist, NT-Xent). Let $\tilde{\mathbf{h}}_i = \hat{\mathbf{h}}_i / \|\hat{\mathbf{h}}_i\|_2$ and $\tilde{\mathbf{y}}_i = \mathbf{y}_i / \|\mathbf{y}_i\|_2$ denote ℓ_2-normalized features for point token i, and let $\mathrm{sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b}$ denote cosine similarity [7] for normalized vectors. We consider (i) a contrastive NT-Xent objective that treats $(\tilde{\mathbf{h}}_i, \tilde{\mathbf{y}}_i)$ as a positive pair and all $(\tilde{\mathbf{h}}_i, \tilde{\mathbf{y}}_j)$ for $j \neq i$ as negatives with temperature τ,

$$\mathcal{L}_{\text{NT-Xent}}(\hat{\mathbf{h}}, \mathbf{y}) = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(\mathrm{sim}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{y}}_i)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\mathrm{sim}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{y}}_j)/\tau\right)}, \tag{6}$$

and (ii) token-wise cosine distance,

$$\mathcal{L}_{\text{cos-dist}}(\hat{\mathbf{h}}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \mathrm{sim}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{y}}_i)\right). \tag{7}$$

These objectives provide effective training-time guidance by transferring teacher structure at the level of individual point tokens. $\mathcal{L}_{\text{cos-dist}}$ matches per-point feature content, while $\mathcal{L}_{\text{NT-Xent}}$ additionally enforces per-point discriminability through negatives. Importantly, because assembly is governed by mating relations, token-wise alignment is most helpful when the teacher features implicitly induce a useful similarity structure among points. However, it does not explicitly constrain inter-point relationships and can be less reliable when those relations are subtle. To explicitly transfer such *relational* structure, we adopt a Centered Kernel Alignment (CKA) objective [8] that matches the pairwise similarity structure among tokens.

Our primary objective: relational alignment via CKA. Given student features $\hat{\mathbf{h}} \in \mathbb{R}^{N \times D_f}$ and teacher features $\mathbf{y} \in \mathbb{R}^{N \times D_f}$, we compute $N \times N$ Gram matrices between the features. However, computing exhaustive Gram matrices and using them for loss computations may introduce intractable overheads. We therefore randomly subsample a shared set of n token indices uniformly to keep the $n \times n$ Gram matrices tractable, and form

$$\mathbf{G}_S = \hat{\mathbf{h}}_{\mathcal{I}} \hat{\mathbf{h}}_{\mathcal{I}}^\top, \qquad \mathbf{G}_T = \mathbf{y}_{\mathcal{I}} \mathbf{y}_{\mathcal{I}}^\top \in \mathbb{R}^{n \times n}, \tag{8}$$

where $\mathcal{I} \subset \{1, \ldots, N\}$ with $|\mathcal{I}| = n \ll N$ and the subscript denotes row selection. These matrices encode all pairwise inner products between the sampled tokens. We then center them using

$$\mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^\top, \qquad \tilde{\mathbf{G}}_S = \mathbf{H}\mathbf{G}_S\mathbf{H}, \quad \tilde{\mathbf{G}}_T = \mathbf{H}\mathbf{G}_T\mathbf{H}. \tag{9}$$

Finally, we define the CKA loss as the negative normalized alignment between centered Gram matrices:

$$\mathcal{L}_{\text{CKA}}(\hat{\mathbf{h}}, \mathbf{y}) = 1 - \frac{\langle \tilde{\mathbf{G}}_S, \tilde{\mathbf{G}}_T \rangle_F}{\|\tilde{\mathbf{G}}_S\|_F \, \|\tilde{\mathbf{G}}_T\|_F}, \tag{10}$$

where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product. Unlike token-wise matching, $\mathcal{L}_{\text{CKA}}$ explicitly preserves "who is similar to whom" across all token pairs, and is invariant to isotropic scaling and orthogonal transformations of the feature space [8].

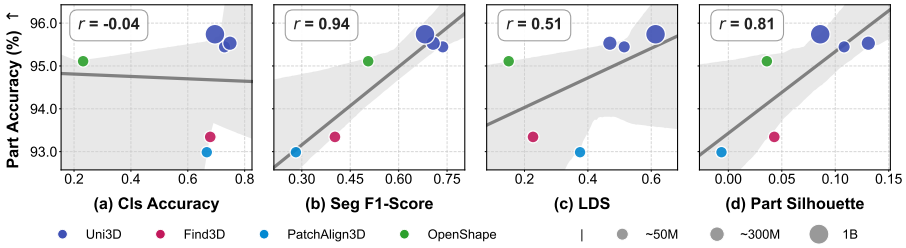

Fig. 4: Correlation Analysis of Teacher Representations. We analyze the relationship between representation properties and assembly performance. **(a-b)** Global semantic understanding (object classification) exhibits a much weaker correlation with final shape assembly accuracy in comparison to spatial structure awareness (mating surface segmentation). **(c-d)** Local spatial structure metrics such as LDS struggle to identify performant teachers, whereas measures highlighting particular geometric properties depict much clearer trends in assembly performance. Overall, geometry- and contact-centric teacher properties are more indicative of downstream assembly quality, motivating our structural distillation objective.

4 Understanding Alignment for 3D Assembly

4.1 What Makes a Good Teacher for 3D Assembly?

Having defined the alignment objectives above, in this section, we proceed to investigate what makes a good teacher for 3D shape assembly. A key ingredient in our training-time alignment is the choice of a teacher encoder. While several pretrained 3D point cloud encoders are available, it is unclear *a priori* which teacher properties are most relevant for 3D part assembly, where success is governed by geometric compatibility and sparse mating interactions. To ground this choice, we study a range of pretrained 3D encoders as teachers and ask: which aspects of a teacher representation predict downstream improvements when used for alignment?

To this end, we evaluate six pretrained 3D encoders as teachers [10,31,35,54]. For each teacher model, we compute lightweight linear-probe metrics on frozen features that capture complementary properties: (i) **classification accuracy** as a proxy for global semantic content, (ii) **mating-surface segmentation F1** as a proxy for contact-awareness and interaction geometry, (iii) **Local-vs-Distant Similarity (LDS)** as a geometry-sensitivity measure, and (iv) **Part Silhouette** as a proxy for part-level geometric priors.[6] We then correlate each probe score with downstream assembly performance (Part Accuracy) obtained when the teacher is used for alignment.

Figure 4 shows a consistent trend across the teacher models we evaluate. **Semantic classification accuracy shows little predictive value** for assembly transfer, exhibiting near-zero correlation with downstream Part Accuracy. In contrast, **geometry- and interaction-centric probes are more predictive**

[6] We describe probe setups and protocols in the supplementary material.

of downstream gains: mating-surface segmentation F1 shows the strongest association with Part Accuracy, Part Silhouette prediction is also strongly aligned, and LDS exhibits a moderate relationship. Taken together, these results support a simple takeaway: effective teacher supervision for 3D assembly depends more on encoding *interaction geometry*—potential contact regions and shared geometric context across parts—than on category-level semantics.

These observations suggest a practical guideline for teacher selection in 3D shape assembly: teacher quality is better assessed using **geometry/contact probes** than global recognition metrics. Guided by this, we adopt Uni3D [54] as our default teacher, as it consistently achieves strong geometry/contact probe performance and yields the best downstream assembly accuracy across alignment objectives. We use Uni3D throughout the paper unless stated otherwise, and further justify this choice in the following section.

4.2 Teacher Choice for Distillation

Figure 5 evaluates how the choice of teacher encoder affects downstream assembly when used for alignment. We compare several pretrained 3D foundation models as teachers while keeping the student backbone, training protocol, and alignment formulation fixed, and report Part Accuracy for both token-wise cosine matching ($\mathcal{L}_{\text{cos-dist}}$) and relational topology alignment ($\mathcal{L}_{\text{CKA}}$). Overall, stronger teachers consistently translate into better assembly performance.

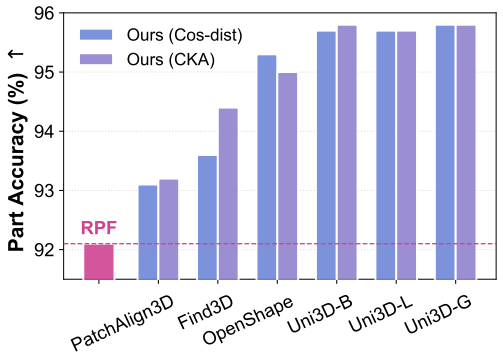

Fig. 5: Impact of different teachers on distillation. We compare the Part Accuracy of TORA across $\mathcal{L}_{\text{cos-dist}}$ and $\mathcal{L}_{\text{CKA}}$ across various 3D foundation models as teachers on Breaking Bad dataset [39]. The dashed line indicates the RPF baseline [42].

Specifically, across teachers, we find that **Uni3D** yields the most reliable gains under both objectives and achieves the highest Part Accuracy, outperforming PatchAlign3D, Find3D, and OpenShape by a clear margin. Notably, this advantage holds across Uni3D variants, indicating that teacher choice is not merely a matter of model scale but of how well the representation encodes assembly-relevant geometric structure. These results align with our correlation analysis: teachers that better capture geometry/contact cues provide more effective supervision for point-flow assembly. We refer the readers to the supplementary material for additional visualization that corroborates this.

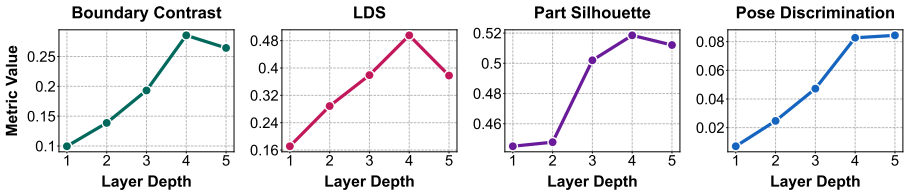

Fig. 6: Spatial structure emerges in later layers. We measure four spatial metrics across layers of the RPF flow model *without* alignment. All metrics increase with depth, indicating that the model progressively resolves spatial part structure in its later layers.

4.3 Emergent Spatial Structure in Later Layers

In addition to the teacher selection choice, it also remains an important exploration to choose an effective alignment target. To this end, we analyze how spatial structure emerges across layers of the *unaligned* RPF flow backbone. We extract intermediate features from each transformer layer and evaluate four complementary metrics on frozen representations: **Boundary Contrast**, which measures how sharply features change across inter-part boundaries; **LDS**, which measures whether features preserve local geometric neighborhoods relative to distant points; **Part Silhouette**, which measures how well features cluster by part identity; and **Pose Discrimination**, which measures sensitivity to rigid part motions by comparing features from the assembled configuration to those from a pose-deformed version (per-part rotated/translated). We provide precise definitions and implementation details for all metrics in the supplementary material.

As shown in Fig. 6, all four metrics increase with layer depth, indicating that later layers progressively organize features into assembly-relevant spatial structure. Boundary transitions become sharper, local geometry is more coherently represented, and part-level clusters become more separable. The increase in pose discrimination further suggests that pose-awareness—necessary for 6-DoF recovery—emerges predominantly in later layers where global context is integrated.Motivated by these trends, we apply alignment at late representations, where the model is actively forming the global structure and interactions that govern mating. Figure 7 corroborates this choice: aligning at deeper layers consistently improves Part Accuracy, with the best performance achieved at l^*=5, outperforming both early-layer alignment and the unaligned baseline. Based on these analyses, in the following, we evaluate the effectiveness of the proposed framework and compare it against existing methods.

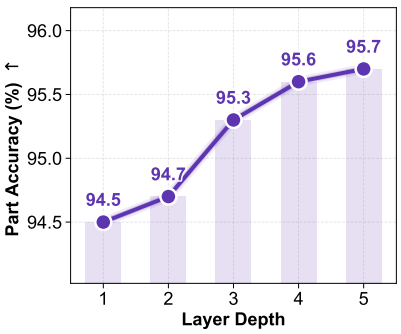

Fig. 7: Ablation study on alignment layer depth. Part Accuracy consistently improves when applying alignment to later layers.

5 Experimental Results

5.1 Experimental Setup

Datasets. We evaluate our method on six benchmark datasets, such as Breaking Bad [39], TwoByTwo [38], PartNet-Assembly [48], Fractura [29], Fantastic Breaks [24], Breaking Bad-*Artifact* [39], that span diverse assembly complexities, ranging from irregular geometric fractures to semantically meaningful multi-part furniture and functional pairwise interactions. All models are trained across all object categories within each benchmark. We refer the readers to the appendix for more details of the datasets.

Evaluation Metrics. Following the rigorous evaluation protocols established in recent literature [42], we quantitatively assess assembly performance using three primary metrics:

- *Part Accuracy (PA):* The primary indicator of overall assembly success. It is defined as the percentage of parts whose Chamfer Distance to their ground-truth assembled positions is strictly below a predefined threshold $\tau = 0.01$.
- *Rotation Error (RE):* We measure the geodesic distance (in degrees) between the predicted and ground-truth rotation matrices using the Rodrigues formula. As pointed out by Sun *et al.* [42], this provides a mathematically proper distance metric on the SO(3) manifold, avoiding the representation singularities inherent to the RMSE of Euler angles used in older benchmarks.
- *Translation Error (TE):* The Root Mean Square Error (RMSE) of the predicted translation vectors, measured in centimeters (cm).

5.2 Implementation Details

We implement our method in PyTorch Lightning [9] and train on 8 NVIDIA GH200 GPUs with a total batch size of 256 for 2,000 epochs. We use AdamW [32] with a learning rate of 5×10^{-4}, halved every 200 epochs after 1,000 epochs. For the overlap-aware encoder, we use the official pretrained checkpoint from RPF [42]. Unless otherwise stated, all experiments use Uni3D-L [54] as the teacher encoder, CKA loss ($\mathcal{L}_{\text{CKA}}$) as the alignment objective with n=1,024 subsampled tokens. The alignment weight is set to λ=0.5, and for the NT-Xent variant, we use a temperature of τ=0.07.

5.3 Experimental Results

Multi-part Assembly. Table 1 summarizes results on three complementary multi-part assembly regimes: *geometric* reassembly (Breaking Bad), *semantic* part assembly (PartNet-Assembly), and *inter-object* assembly under stronger distribution shift (TwoByTwo). We report Part Accuracy together with rotation and translation errors. Across all benchmarks, casting training as teacher–student distillation consistently improves the strong RPF baseline, confirming

Table 1: Quantitative comparison on 3D shape assembly benchmarks. The **best** and second best results are highlighted. For fair comparison, we re-evaluate all methods under a unified evaluation protocol. See supplementary material for details.

(a) Breaking Bad [39]: Multi-part *geometric* shape assembly benchmark. Results are grouped by the number of parts per object: 2 to 20 (left) and 21 to 33 (right).

	Breaking Bad - Everyday [2,20]			**Breaking Bad - Everyday** [21,33]		
Methods	PA (%) ↑	RE (°) ↓	TE (cm) ↓	PA (%) ↑	RE (°) ↓	TE (cm) ↓
Jigsaw [34]	69.7	30.2	8.9	12.0	92.1	20.7
PMTR [26]	70.6	24.9	14.9	8.4	72.2	20.1
CMNet [27]	80.4	19.7	11.6	24.7	67.7	20.8
GARF [29]	92.4	**7.4**	2.7	25.1	68.0	32.9
RPF [42]	93.2	16.0	4.3	62.1	77.3	15.2
Ours$_{\text{NT-Xent}}$	92.9	17.8	4.7	62.6	77.0	15.2
Ours$_{\text{Cos-dist}}$	**95.7**	9.0	2.2	**72.4**	**64.3**	**12.3**
Ours$_{\text{CKA}}$	**95.7**	8.6	**2.1**	71.7	64.8	12.5

(b) PartNet-Assembly [48]: Multi-part *semantic* shape assembly, and **TwoByTwo** [38]: *inter-object* assembly benchmark.

	PartNet-Assembly			**TwoByTwo**		
Methods	PA (%) ↑	RE (°) ↓	TE (cm) ↓	PA (%) ↑	RE (°) ↓	TE (cm) ↓
RPF [42]	59.8	46.2	21.5	65.4	15.8	11.9
Ours$_{\text{NT-Xent}}$	65.4	43.6	20.2	60.4	17.2	12.8
Ours$_{\text{Cos-dist}}$	67.8	41.6	19.1	68.9	12.0	9.5
Ours$_{\text{CKA}}$	**69.1**	**40.8**	**18.8**	**71.5**	**10.0**	**7.6**

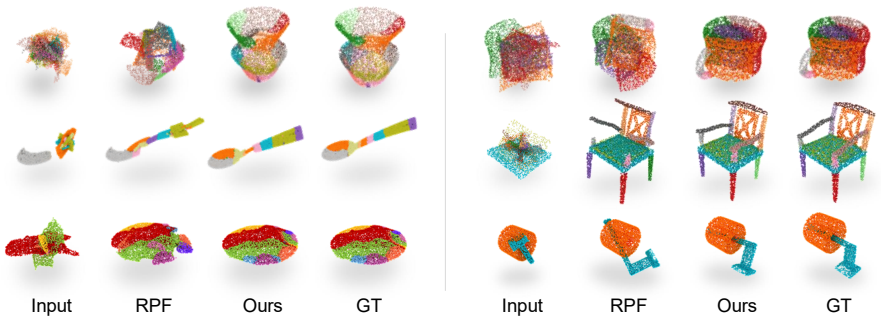

Fig. 8: Qualitative Comparison. While the baseline RPF often struggles with precise part positioning and fails to resolve complex inter-part relations, ours consistently produces structurally coherent assemblies that closely match the ground truth.

that injecting pretrained geometric priors into the flow backbone is broadly beneficial for 3D assembly.

Table 3: Zero-shot Evaluation on Breaking Bad - Artifact [39], FRACTURA [29] and Fantastic Breaks [24] datasets.

	Breaking Bad - Artifact			FRACTURA			Fantastic Breaks		
Methods	PA (%) ↑	RE (°) ↓	TE (cm) ↓	PA (%) ↑	RE (°) ↓	TE (cm) ↓	PA (%) ↑	RE (°) ↓	TE (cm) ↓
GARF [29]	91.4	8.7	3.0	44.2	37.8	26.9	88.3	8.2	3.0
RPF [42]	88.3	20.9	5.3	68.1	50.1	11.2	96.9	6.3	1.5
Ours$_{\text{NT-Xent}}$	87.0	22.9	5.9	71.7	48.9	10.6	97.6	6.3	1.5
Ours$_{\text{Cos-dist}}$	93.2	11.3	2.8	74.9	**35.5**	7.9	**97.7**	4.5	1.1
Ours$_{\text{CKA}}$	**94.4**	**8.0**	**2.1**	**76.0**	36.4	**7.7**	97.2	**3.5**	**0.9**

On Breaking Bad (2 to 20 parts), even our simplest instantiation—token-wise cosine alignment—already yields substantial gains over RPF in Part Accuracy and pose errors, demonstrating that representation alignment transfers effectively to the 3D point-flow setting. Topology alignment (CKA) matches or further improves these results, achieving the best overall pose accuracy. The many-part setting (21 to 33 parts) highlights the scalability challenges faced by existing methods: correspondence-based approaches such as Jigsaw, CMNet, and PMTR degrade significantly as the combinatorial complexity of multi-fragment matching grows, while GARF, despite being competitive on the few-part split, drops sharply to lower Part Accuracy. In contrast, RPF maintains reasonable performance in this regime, and our alignment further amplifies its scalability, delivering massive gains in both Part Accuracy and pose errors. Figure 8 provides qualitative examples where RPF struggles with precise part positioning, while our method produces assemblies closely matching the ground truth.

On PartNet-Assembly and TwoByTwo, CKA delivers the greatest improvements, suggesting that explicitly preserving relational structure is particularly valuable when assembly involves structured part interactions and domain shift. We also note that NT-Xent degrades performance on TwoByTwo, indicating that enforcing per-point discriminability can be counterproductive in inter-object settings where the teacher already separates distinct objects (see supplementary).

Loss combination ablation. The comparisons above evaluate the three alignment objectives individually, where $\mathcal{L}_{\text{CKA}}$ consistently performs best. A natural follow-up is whether *combining* them provides complementary gains over $\mathcal{L}_{\text{CKA}}$ alone. We study this on TwoByTwo, as its inter-object setting most strongly exposes the differences among objectives (Tab. 1), making it the most discriminative regime for isolating each loss's contribution.

Table 2 reports Part Accuracy on TwoByTwo when augmenting $\mathcal{L}_{\text{CKA}}$ with token-wise objectives. In all cases, combining multiple losses degrades performance, with the full combination performing worst. This indicates that $\mathcal{L}_{\text{CKA}}$'s relational signal already captures the relevant structural information, and token-wise objectives introduce competing gradients that dilute the alignment.

$\mathcal{L}_{\text{align}}$	PA ↑
$\mathcal{L}_{\text{CKA}}$ **(Ours)**	**71.5**
$\mathcal{L}_{\text{CKA}} + \mathcal{L}_{\text{cos-dist}}$	70.0
$\mathcal{L}_{\text{CKA}} + \mathcal{L}_{\text{NT-Xent}}$	68.5
$\mathcal{L}_{\text{CKA}} + \mathcal{L}_{\text{cos-dist}} + \mathcal{L}_{\text{NT-Xent}}$	67.7

Table 2: Loss combination ablation on TwoByTwo.

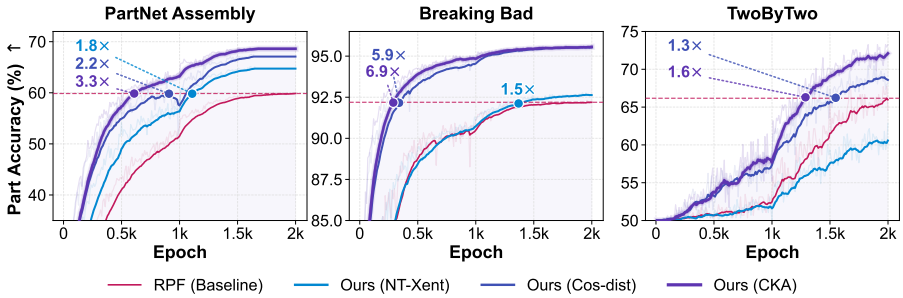

Fig. 9: Convergence comparison. We monitor the validation Part Accuracy of Ours (CKA) (—) against the RPF (—) and other alignment strategies including Ours (NT-Xent) (—) and Ours (Cos-dist) (—) over training epochs across three datasets. The dashed horizontal line (- -) represents the peak accuracy of the baseline. The annotated multipliers indicate the convergence speedup relative to the baseline to reach its peak performance.

Zero-shot Transfer to Unseen Datasets. To assess robustness under domain shift, we evaluate models trained on the Breaking Bad everyday split in a zero-shot manner on three unseen datasets: Breaking Bad artifact split [39] (synthetic, unseen object categories), FRACTURA [29] (mixed synthetic and real fractures across scientific domains), and Fantastic Breaks [24] (real-world scanned objects). Table 3 summarizes the results without any fine-tuning. Our alignment transfers substantially better than the RPF baseline across all three datasets. In particular, ours with a CKA objective achieves the best overall performance, attaining the lowest pose errors on both FRACTURA and Fantastic Breaks. Token-wise cosine alignment consistently improves over RPF, whereas the contrastive NT-Xent objective is less reliable under shift, often degrading pose accuracy. Overall, these results confirm that topological alignment yields strong generalization to unseen distributions, supporting our claim that transferring relational structure is particularly beneficial for 3D assembly.

Convergence Analysis. Figure 9 compares validation Part Accuracy as training progresses on PartNet-Assembly, Breaking Bad, and TwoByTwo. Across all datasets, alignment accelerates optimization relative to the RPF baseline, and the proposed topology alignment provides the most consistent speedup. On Breaking Bad, TORA reaches the baseline's peak performance substantially earlier (about 6.9× faster), and also converges to a higher final accuracy; token-wise cosine alignment also speeds up training but with a smaller gain. On PartNet-Assembly, TORA again yields the largest acceleration, reaching the baseline peak 3.3× sooner, compared to 2.2× for cosine and 1.8× for NT-Xent. The effect is most pronounced under domain shift on TwoByTwo, where TORA improves both convergence and final accuracy, while token-wise alignment provides more limited acceleration and NT-Xent remains less reliable. Overall, these results indicate that explicitly matching relational structure provides a richer and more task-aligned training signal, enabling faster and more stable optimization across assembly regimes.

Efficiency Analysis. We profile the training overhead of TORA against the baseline RPF in Tab. 4. Crucially, because the teacher encoder remains completely frozen, its representations can be practically precomputed and cached *offline*. Under this standard feature-caching setting, TORA incurs near-zero practical overhead, requiring only an additional 0.18 GB peak VRAM (+5.5%) and ∼3 ms per step (+2.8%) compared to the baseline—negligible on modern hardware. However, when executing an *online* forward pass, the computational footprint is shown to increase. Nevertheless, it remains highly manageable, as only approximately 1.4 GB and 27 ms are additionally induced. We highlight that this is achieved via our implementation to efficiently compute Gram matrices through random subsampling, as discussed in Sec. 3.3. Ultimately, this confirms that TORA is highly scalable, preserving the massive convergence speedup without compromising training throughput, while adding strictly zero overhead during inference.

Table 4: Per-step Training Overhead. Measured on a single NVIDIA GH200 GPU with a batch size of 1. Statistics are averaged following a 50-step warmup.

Methods	$\mathcal{L}_{\text{align}}$	**Memory** (GB)	**Time** (ms)	**Throughput** (steps/s)
Ours (Online Teacher)	NT-Xent	4.79	124.85	8.01
	cos-dist	4.85	122.02	8.20
	CKA	4.72	128.39	7.79
Ours (Offline Teacher)	NT-Xent	3.48	104.19	9.60
	cos-dist	3.48	104.69	9.55
	CKA	3.48	104.22	9.59
RPF [42]	-	3.30	101.40	9.86

6 Conclusion

In this work, we have introduced TORA, a topology-first teacher–student alignment framework for robust 3D point-flow assembly. By distilling inter-point relational structure from a frozen pretrained 3D encoder into a flow-matching assembly model, TORA injects interaction-aware geometric priors while preserving the original inference pipeline and incurring zero test-time overhead. Extensive experiments across semantic, geometric, and inter-object assembly benchmarks demonstrate consistent improvements over strong flow-based baselines, with particularly pronounced gains under domain shift, where relational topology transfer is most beneficial. Our analysis further clarifies what makes an effective teacher for 3D assembly—geometry- and contact-centric signals rather than category semantics—and shows that topology alignment accelerates convergence and improves final accuracy. We hope these findings encourage broader use of relational distillation objectives for 3D generative transport models and enable more robust assembly in real-world robotics and graphics applications.

Acknowledgements

This research was supported by the ETH AI Center through an ETH AI Center postdoctoral fellowship to Sunghwan Hong, Swiss AI Initiative (a144), and ELLIOT (Grant Agreement 101214398).

References

1. An, H., Jung, J., Kim, M., Hong, S., Kim, C., Fukuda, K., Jeon, M., Han, J., Narihira, T., Ko, H., et al.: C3g: Learning compact 3d representations with 2k gaussians. arXiv preprint arXiv:2512.04021 (2025)
2. An, H., Kim, J.H., Park, S., Jung, J., Han, J., Hong, S., Kim, S.: Cross-view completion models are zero-shot correspondence estimators. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1103–1115 (2025)
3. Chaudhuri, S., Kalogerakis, E., Guibas, L., Koltun, V.: Probabilistic reasoning for assembly-based 3d modeling. ACM Transactions on Graphics (TOG) **30**(4), 1–10 (2011)
4. Chen, Z., Lee, N., Sun, B., Kwon, T., Pollefeys, M., Bauer, Z., Hong, S.: Prose: Training-free egocentric scene registration with vision-language models. arXiv preprint arXiv:2606.16569 (2026)
5. Cho, S., Hong, S., Jeon, S., Lee, Y., Sohn, K., Kim, S.: Cats: Cost aggregation transformers for visual correspondence. Advances in Neural Information Processing Systems **34**, 9011–9023 (2021)
6. Cho, S., Hong, S., Kim, S.: Cats++: Boosting cost aggregation with convolutions and transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 7174–7194 (2022)
7. Cho, S., Shin, H., Hong, S., Arnab, A., Seo, P.H., Kim, S.: Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4113–4123 (2024)
8. Cortes, C., Mohri, M., Rostamizadeh, A.: Algorithms for learning kernels based on centered alignment. The Journal of Machine Learning Research **13**, 795–828 (2012)
9. Falcon, W., team, T.P.L.: Pytorch lightning (2019), `https://github.com/PyTorchLightning/pytorch-lightning`
10. Hadgi, S., Gong, B., Sundararaman, R., Pierson, E., Li, L., Wonka, P., Ovsjanikov, M.: Patchalign3d: Local feature alignment for dense 3d shape understanding. arXiv preprint arXiv:2601.02457 (2026)
11. Han, J., An, H., Jung, J., Narihira, T., Seo, J., Fukuda, K., Kim, C., Hong, S., Mitsufuji, Y., Kim, S.: Dˆ 2ust3r: Enhancing 3d reconstruction with 4d pointmaps for dynamic scenes. arXiv preprint arXiv:2504.06264 (2025)
12. Han, J., Hong, S., Jung, J., Jang, W., An, H., Wang, Q., Kim, S., Feng, C.: Emergent outlier view rejection in visual geometry grounded transformers. arXiv preprint arXiv:2512.04012 (2025)
13. Han, J., Jeon, S., Jung, J., Zurbrügg, R., An, H., Portela, T., Hutter, M., Pollefeys, M., Kim, S., Hong, S.: Geometric action model for robot policy learning. arXiv preprint arXiv:2606.17046 (2026)
14. Harada, K., Nagata, K., Rojas, J., Ramirez-Alpizar, I.G., Wan, W., Onda, H., Tsuji, T.: Proposal of a shape adaptive gripper for robotic assembly tasks. Advanced Robotics **30**(17-18), 1186–1198 (2016)
15. Hong, S., Cho, S., Kim, S., Lin, S.: Unifying feature and cost aggregation with transformers for semantic and visual correspondence. arXiv preprint arXiv:2403.11120 (2024)
16. Hong, S., Cho, S., Nam, J., Lin, S., Kim, S.: Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In: European Conference on Computer Vision. pp. 108–126. Springer (2022)

17. Hong, S., Jung, J., Shin, H., Han, J., Yang, J., Luo, C., Kim, S.: Pf3plat: Pose-free feed-forward 3d gaussian splatting. arXiv preprint arXiv:2410.22128 (2024)
18. Hong, S., Jung, J., Shin, H., Yang, J., Kim, S., Luo, C.: Unifying correspondence pose and nerf for generalized pose-free novel view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20196–20206 (2024)
19. Hong, S., Kim, S.: Deep matching prior: Test-time optimization for dense correspondence. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9907–9917 (2021)
20. Hong, S., Nam, J., Cho, S., Hong, S., Jeon, S., Min, D., Kim, S.: Neural matching fields: Implicit representation of matching fields for visual correspondence. Advances in Neural Information Processing Systems **35**, 13512–13526 (2022)
21. Jocher, G., Stoken, A., Borovec, J., Changyu, L., Hogan, A., Chaurasia, A., Diaconu, L., Ingham, F., Colmagro, A., Ye, H., et al.: ultralytics/yolov5: v4. 0-nn. silu () activations, weights & biases logging, pytorch hub integration. Zenodo (2021)
22. Jones, R.K., Barton, T., Xu, X., Wang, K., Jiang, E., Guerrero, P., Mitra, N.J., Ritchie, D.: Shapeassembly: Learning to generate programs for 3d shape structure synthesis. ACM Transactions on Graphics (TOG) **39**(6), 1–20 (2020)
23. Kim, C., Shin, H., Hong, E., Yoon, H., Arnab, A., Seo, P.H., Hong, S., Kim, S.: Seg4diff: Unveiling open-vocabulary segmentation in text-to-image diffusion transformers. arXiv preprint arXiv:2509.18096 (2025)
24. Lamb, N., Palmer, C., Molloy, B., Banerjee, S., Banerjee, N.K.: Fantastic breaks: A dataset of paired 3d scans of real-world broken objects and their complete counterparts. In: CVPR (2023)
25. Lee, J., Jung, J., Han, J., Narihira, T., Fukuda, K., Seo, J., Hong, S., Mitsufuji, Y., Kim, S.: 3d scene prompting for scene-consistent camera-controllable video generation. arXiv preprint arXiv:2510.14945 (2025)
26. Lee, N., Min, J., Lee, J., Kim, S., Lee, K., Park, J., Cho, M.: 3d geometric shape assembly via efficient point cloud matching. In: Proceedings of the International Conference on Machine Learning (ICML) (2024)
27. Lee, N., Min, J., Lee, J., Park, C., Cho, M.: Combinative matching for geometric shape assembly. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9540–9549 (2025)
28. Leng, X., Singh, J., Hou, Y., Xing, Z., Xie, S., Zheng, L.: Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18262–18272 (2025)
29. Li, S., Jiang, Z., Chen, G., Xu, C., Tan, S., Wang, X., Fang, I., Zyskowski, K., McPherron, S.P., Iovita, R., et al.: Garf: Learning generalizable 3d reassembly for real-world fractures. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5711–5721 (2025)
30. Li, Y., Mo, K., Duan, Y., Wang, H., Zhang, J., Shao, L.: Category-level multi-part multi-joint 3d shape assembly. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3281–3291 (2024)
31. Liu, M., Shi, R., Kuang, K., Zhu, Y., Li, X., Han, S., Cai, H., Porikli, F., Su, H.: Openshape: Scaling up 3d shape representation towards open-world understanding. Advances in neural information processing systems **36**, 44860–44879 (2023)
32. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
33. Lu, J., Liang, Y., Han, H., Hua, J., Jiang, J., Li, X., Huang, Q.: A survey on computational solutions for reconstructing complete objects by reassembling their fractured parts. In: Computer Graphics Forum. vol. 44, p. e70081. Wiley Online Library (2025)

34. Lu, J., Sun, Y., Huang, Q.: Jigsaw: Learning to assemble multiple fractured objects. Advances in Neural Information Processing Systems **36**, 14969–14986 (2023)
35. Ma, Z., Yue, Y., Gkioxari, G.: Find any part in 3d. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7818–7827 (2025)
36. McBride, J.C., Kimia, B.B.: Archaeological fragment reconstruction using curve-matching. In: CVPRW (2003)
37. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
38. Qi, Y., Ju, Y., Wei, T., Chu, C., Wong, L.L., Xu, H.: Two by two: Learning multi-task pairwise objects assembly for generalizable robot manipulation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17383–17393 (2025)
39. Sellán, S., Chen, Y.C., Wu, Z., Garg, A., Jacobson, A.: Breaking bad: A dataset for geometric fracture and reassembly. Advances in Neural Information Processing Systems **35**, 38885–38898 (2022)
40. Singh, J., Leng, X., Wu, Z., Zheng, L., Zhang, R., Shechtman, E., Xie, S.: What matters for representation alignment: Global information or spatial structure? arXiv preprint arXiv:2512.10794 (2025)
41. Son, K., Almeida, E.B., Cooper, D.B.: Axially symmetric 3d pots configuration system using axis of symmetry and break curve. In: CVPR (2013)
42. Sun, T., Zhu, L., Huang, S., Song, S., Armeni, I.: Rectified point flow: Generic point cloud pose estimation. arXiv preprint arXiv:2506.05282 (2025)
43. Wang, Z., Chen, J., Furukawa, Y.: Puzzlefusion++: Auto-agglomerative 3d fracture assembly by denoise and verify. In: ICLR (2025)
44. Wang, Z., Zhao, W., Zhou, Y., Li, Z., Liang, Z., Shi, M., Zhao, X., Zhou, P., Zhang, K., Wang, Z., et al.: Repa works until it doesn't: Early-stopped, holistic alignment supercharges diffusion training. arXiv preprint arXiv:2505.16792 (2025)
45. Wu, G., Zhang, S., Shi, R., Gao, S., Chen, Z., Wang, L., Chen, Z., Gao, H., Tang, Y., Yang, J., et al.: Representation entanglement for generation: Training diffusion transformers is much easier than you think. arXiv preprint arXiv:2507.01467 (2025)
46. Wu, H., Wu, D., He, T., Guo, J., Ye, Y., Duan, Y., Bian, J.: Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. arXiv preprint arXiv:2507.07982 (2025)
47. Wu, R., Tie, C., Du, Y., Zhao, Y., Dong, H.: Leveraging se (3) equivariance for learning 3d geometric shape assembly. In: ICCV (2023)
48. Xu, B., Zheng, S., Jin, Q.: Spaformer: Sequential 3d part assembly with transformers. In: 2025 International Conference on 3D Vision (3DV). pp. 1317–1327. IEEE (2025)
49. Yoo, S.J., Liu, S., Arshad, M.Z., Kim, J., Kim, Y.M., Aloimonos, Y., Fermuller, C., Joo, K., Kim, J., Hong, J.H.: Structure-from-sherds++: Robust incremental 3d reassembly of axially symmetric pots from unordered and mixed fragment collections. arXiv preprint arXiv:2502.13986 (2025)
50. Yoon, H., Jung, J., Kim, J., Choi, H., Shin, H., Lim, S., An, H., Kim, C., Han, J., Kim, D., et al.: Visual representation alignment for multimodal large language models. arXiv preprint arXiv:2509.07979 (2025)
51. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. arXiv preprint arXiv:2410.06940 (2024)
52. Zakka, K., Zeng, A., Lee, J., Song, S.: Form2fit: Learning shape priors for generalizable assembly from disassembly. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 9404–9410. IEEE (2020)

53. Zhan, G., Fan, Q., Mo, K., Shao, L., Chen, B., Guibas, L.J., Dong, H., et al.: Generative 3d part assembly via dynamic graph learning. Advances in Neural Information Processing Systems **33**, 6315–6326 (2020)
54. Zhou, J., Wang, J., Ma, B., Liu, Y.S., Huang, T., Wang, X.: Uni3d: Exploring unified 3d representation at scale. arXiv preprint arXiv:2310.06773 (2023)