TORA: Topological Representation Alignment for 3D Shape Assembly

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

[1]Independent Researcher [2]ETH Zurich [3]ETH AI Center [*] Equal contribution

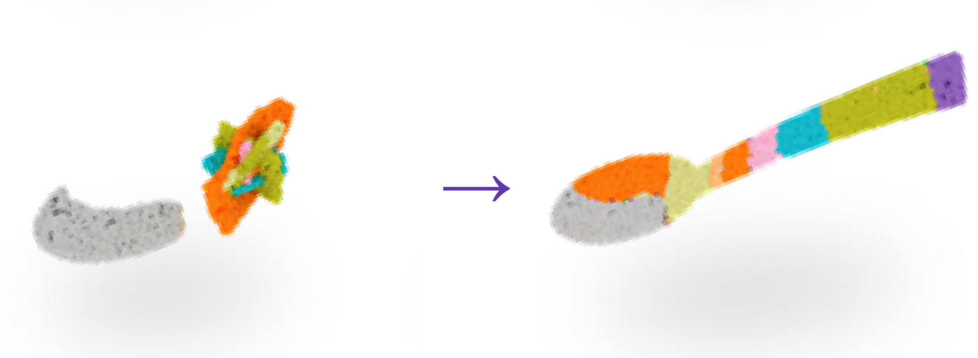
Project page

ETH AI CENTER

ECCV 2026 | MALMÖ | SEPT 8–12

Task: 3D Shape Assembly

Given K **unposed** part point clouds $\{P_k\}$, predict per-part rigid transforms $T_k \in SE(3)$ that reconstruct the original object.



Unposed parts → *Assembled object*

Three regimes with one shared bottleneck:

Semantic	Geometric	Inter-object
PartNet-Assembly	Breaking Bad	TwoByTwo

Discovering **mating relations**: *which regions should contact and co-move*, robustly under symmetry and domain shift.

Motivation: what flow matching misses

- **Rectified Point Flow (RPF)** transports noisy points to assembled positions, then recovers poses in closed form.
  - Trained with **endpoint loss only** ⇒ no signal tells the model *which cross-part interactions* should drive the motion.
  - Mating cues are **sparse and ambiguous** under symmetry ⇒ brittle under distribution shift.

Our idea. Cast training as **teacher–student distillation**: inject relational geometric priors from a **frozen 3D foundation encoder** at *training time*.

Main contributions

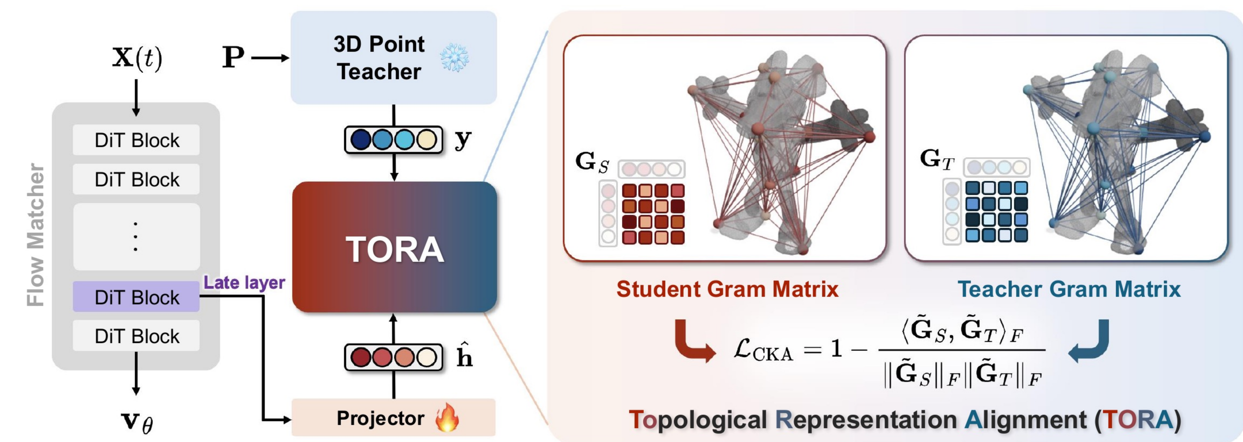
Topology-first **representation alignment** for 3D point-flow assembly using a **Centered Kernel Alignment (CKA)** loss.
Systematic **teacher and layer analysis**: **Geometry-/contact-centric properties**, not category semantics, predict transfer; alignment pays off at **late layers**.
SOTA results on both seen and unseen benchmarks.

6.9×	95.7%	0	94.4%
Faster convergence	Assembly part accuracy	Online inference cost	Zero-shot transfer

Takeaway

Distilling relational topology, not per-token features, from a geometry-aware 3D teacher makes point-flow assembly train faster, generalize further, and cost nothing at inference.

TORA: Topological Representation Alignment



*A **frozen** 3D teacher encodes the assembled geometry; a projector φ maps a late DiT layer of the flow matcher into the teacher space, where relational structure is matched. The projector is discarded after training.*

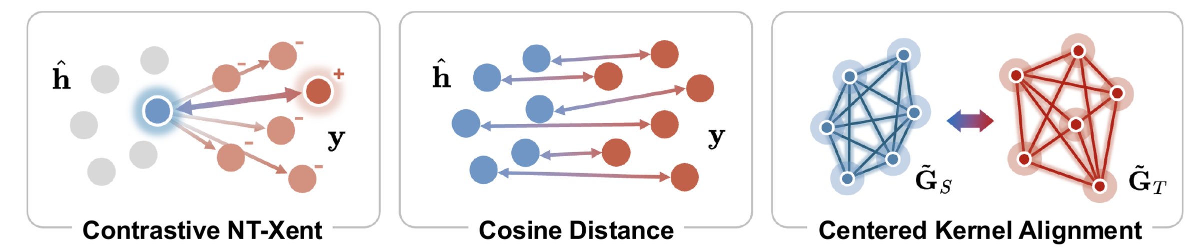
$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CFM}} + \lambda\,\mathcal{L}_{\text{align}}(\hat{\mathbf{h}}, \mathbf{y}) \qquad \hat{\mathbf{h}} = \phi(\mathbf{h}^{(l^*)})$$

Centered Kernel Alignment (CKA)

$$\mathbf{G}_T = \mathbf{y}_{\mathcal{I}}\mathbf{y}_{\mathcal{I}}^\top \qquad \mathbf{G}_S = \hat{\mathbf{h}}_{\mathcal{I}}\hat{\mathbf{h}}_{\mathcal{I}}^\top \qquad \tilde{\mathbf{G}} = \mathbf{HGH}$$

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

Three alignment objectives along one axis: **how explicitly teacher relational structure is transferred**



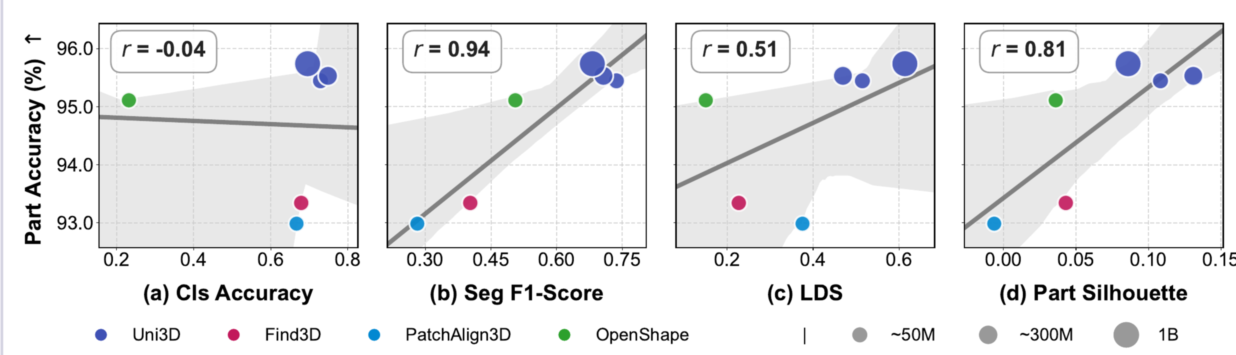
Contrastive NT-Xent	Cosine Distance	Centered Kernel Alignment
$-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\mathrm{sim}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{y}}_i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\mathrm{sim}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{y}}_j)/\tau\right)}$	$\frac{1}{N}\sum_{i=1}^{N}\left(1 - \mathrm{sim}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{y}}_i)\right)$	$1 - \frac{\langle \tilde{\mathbf{G}}_S, \tilde{\mathbf{G}}_T \rangle_F}{\|\tilde{\mathbf{G}}_S\|_F \|\tilde{\mathbf{G}}_T\|_F}$
Contrastive per-token discriminability	Per-token feature matching	**Pairwise similarity** structure matching

Efficient relational alignment via CKA: subsample a shared set *I* of *n* tokens and compare *centered* Gram matrices

- Preserves **"who is similar to whom"** across all token pairs
- Invariant to isotropic scaling and orthogonal transforms.
- Online teacher pass costs only 1.4G/27ms (offline: 0.18G/3ms).

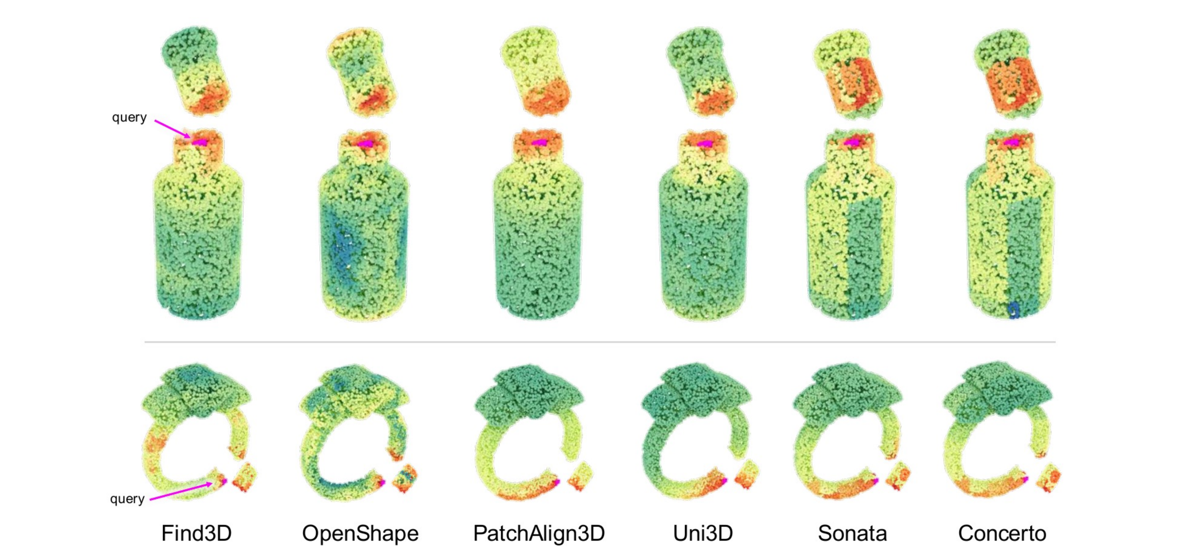
What makes a good teacher?

We probe six frozen 3D encoders and correlate each probe score with downstream Part Accuracy.



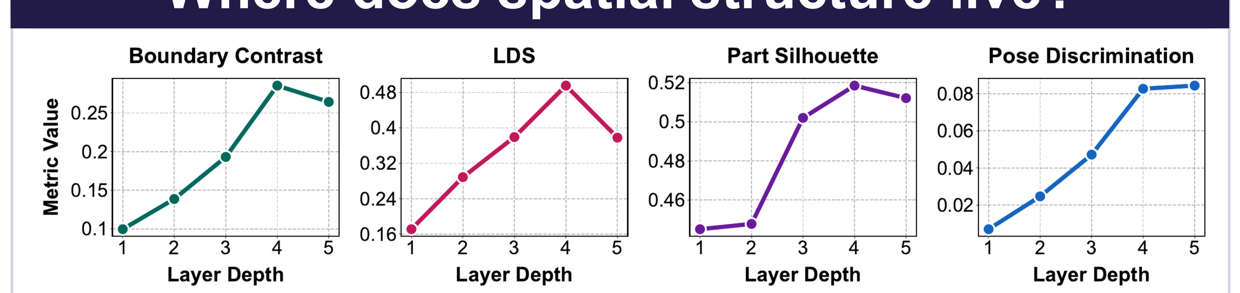
Mating-surface **Seg F1** correlates **strongest** (r = 0.94).
Part Silhouette follows (r = 0.81); **LDS** is moderate (r = 0.51).
Semantic **classification accuracy** is **non-predictive** (r =−0.04).

Effective supervision comes from **interaction geometry** i.e. potential contact regions and shared geometric context, and *not* from category-level semantics.



Similarity to a query point (pink) on a mating surface in each teacher's frozen feature space. Uni3D-L localizes the opposing contact region most sharply.

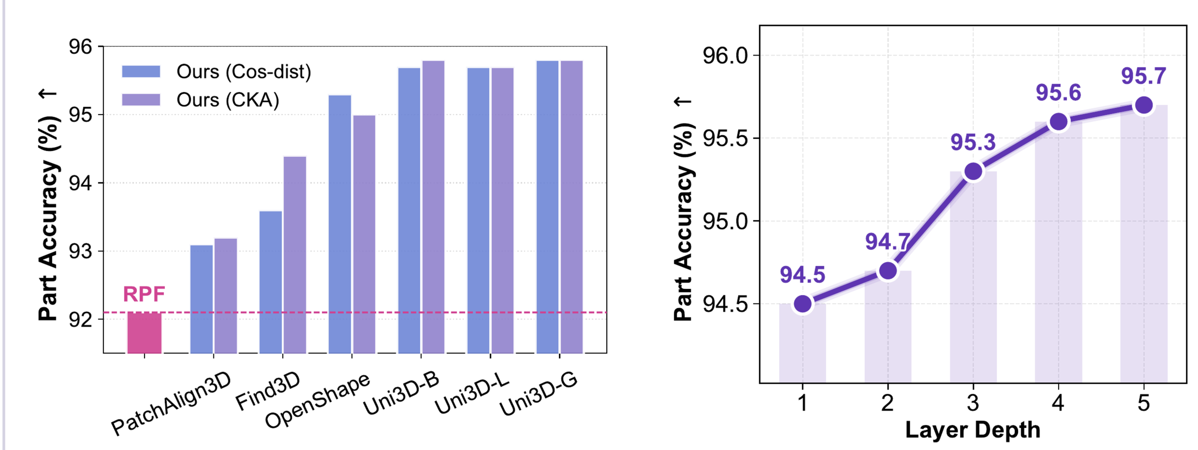
Where does spatial structure live?



Four spatial metrics across layers of the unaligned RPF flow backbone.

All metrics grow **monotonically with depth**, so we align at **later layers** where global structure governing mating is formed.

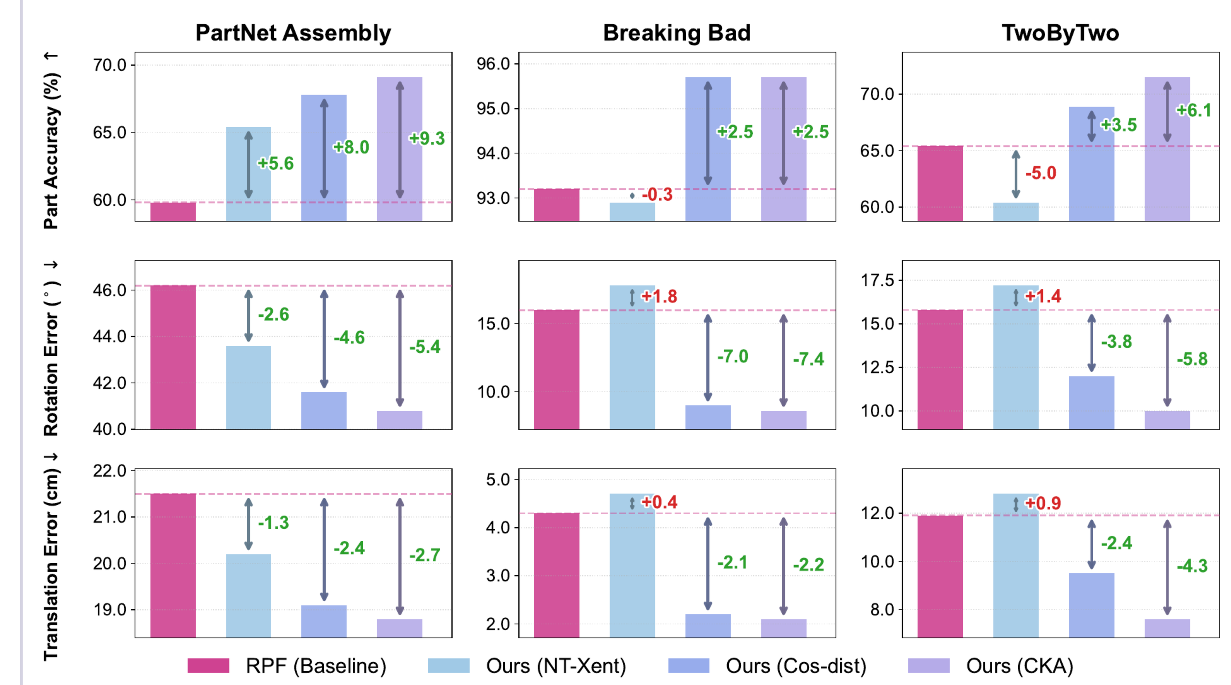
Design choices, validated



Teacher encoder (Breaking Bad) *Alignment layer depth*

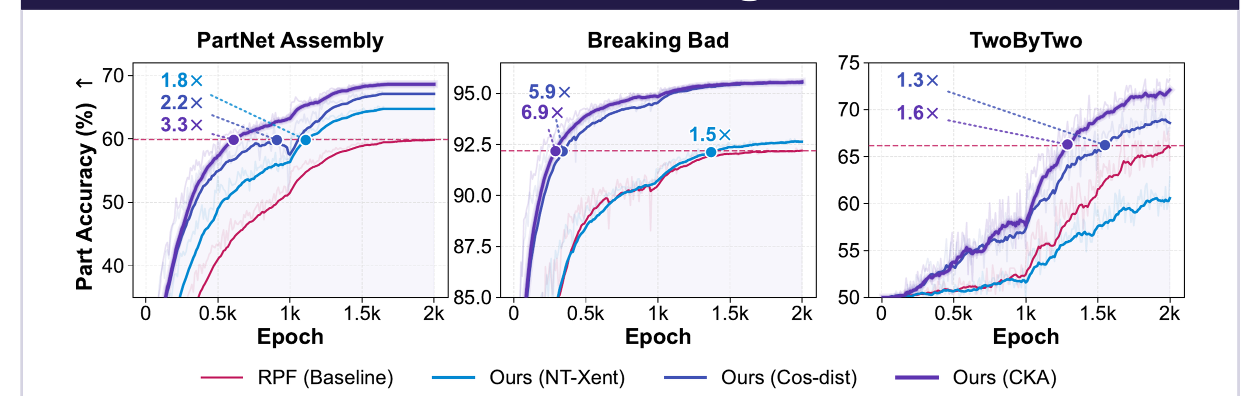
Uni3D wins under both Cos-dist and CKA, and the advantage holds across B/L/G scales — teacher *quality*, not size.
Part Accuracy rises monotonically with depth; best at l^* = 5.

Multi-part assembly results



CKA is best overall; margins are largest under structured part interactions (PartNet) and domain shift (TwoByTwo).

Faster convergence



TORA reaches the baseline's *peak* accuracy **6.9×** (Breaking Bad), **3.3×** (PartNet), **1.6×** (TwoByTwo) sooner and converges higher.

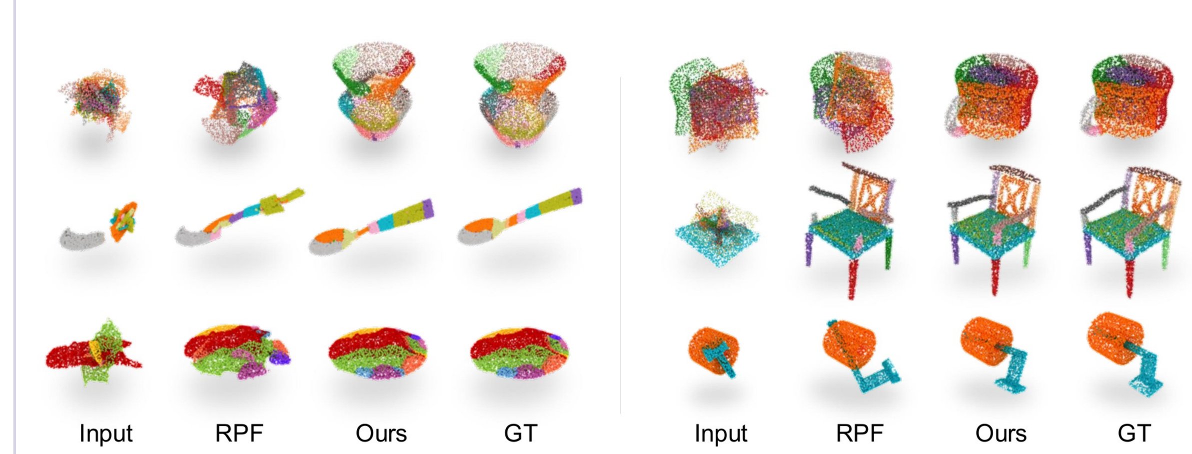
Zero-shot transfer

	BB-Artifact		FRACTURA		Fantastic Breaks	
Method	**PA ↑**	**RE ↓**	**PA ↑**	**RE ↓**	**PA ↑**	**RE ↓**
GARF	91.4	8.7	44.2	37.8	88.3	8.2
RPF	88.3	20.9	68.1	50.1	96.9	6.3
Ours$_{\text{Cos-dist}}$	93.2	11.3	74.9	**35.5**	**97.7**	4.5
Ours$_{\text{CKA}}$	**94.4**	**8.0**	**76.0**	36.4	97.2	**3.5**

Trained on BBad-Everyday, evaluated on unseen synthetic and real-world data.

TORA helps **most where domain shift is largest**.

Qualitative comparison



RPF struggles with precise part positioning; TORA produces structurally coherent assemblies.