DEFORMR: SINGLE-VIEW 3D SHAPE RETRIEVAL WITH IMAGE-GUIDED SHAPE ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-view 3D shape retrieval is a critical enabler for inverse graphics and detailed 3D scene representation, allowing generic 3D assets to serve as proxies for real-world instances. Recent feature-level analysis-by-synthesis approaches have improved robustness to occlusions and domain gaps through pose-conditioned feature alignment at test time. However, a rigid template may not match local geometry visible in an image. To address this mismatch, we propose DeformR, a framework that extends pose-aware 3D retrieval with local, image-guided shape adaptation. First, we formulate a 2D-to-3D feature-distillation and pose-refinement pipeline for shape retrieval and pose estimation. The pipeline supports multiple categories while retaining point-level representations. Second, with the selected template and pose fixed, DeformR is designed for conservative local adaptation rather than free-form reconstruction. Pose-aligned image features guide corrections to image-supported geometric residuals, while geometric regularizers preserve the retrieved template's global structure. Large geometric changes are neither expected nor desired. Experiments on Pascal3D and ImageNet3D show improved retrieval accuracy over the evaluated baselines and competitive pose-estimation performance. With the retrieved template and pose fixed, deformation improves alignment between the real-world image and the deformed shape.

1 INTRODUCTION

Single-view 3D shape retrieval selects a CAD model from a database to serve as a geometric proxy for an object in an image Tangelder & Veltkamp (2004); Xiao et al. (2020); Lin et al. (2021a); Song et al. (2024). Such proxies support downstream tasks in embodied AI, autonomous navigation, and inverse graphics Shi et al. (2025); Ma et al. (2025), but remain rigid and may not capture instance-specific geometry, limiting their application in the real world.

Foundation models have advanced 3D perception and shape retrieval. General 3D perception models such as SAM3D SAM 3D Team et al. (2025) demonstrate strong capabilities, though their inference cost is high and their robustness under severe occlusions remains underexplored. Concurrently, progress in efficient 3D retrieval has been driven by large-scale multimodal alignment and joint 2D–3D embedding spaces Liu et al. (2023); Zhou et al. (2024); Xue et al. (2023; 2024). While successful on synthetic benchmarks, these feed-forward methods can struggle with real-world complexities such as heavy occlusions and domain gaps, motivating recent work on occlusion-invariant retrieval and evaluation under unseen shapes and objects Chu et al. (2024); Wu et al. (2024). Recent analysis-by-synthesis approaches Shi et al. (2026) model pose-conditioned 3D representations and refine alignment at test time for 3D shape retrieval, which highly relies on the ability of the 2D foundation model, ignoring model 2D–3D correspondence. Yet rigid candidate matching cannot adjust image-supported local geometry without part-level 2D–3D correspondences.

To address this gap, we propose DeformR, a scalable and unified framework for robust 3D shape retrieval followed by instance-level geometric adaptation. Building upon general shape retrieval paradigms, we first formulate a comprehensive 2D-to-3D feature distillation pipeline. To resolve severe geometric ambiguities, such as distinguishing between identical parts at different locations (*e.g.,* the front and rear wheels of a bicycle), we introduce a dense spatial contrastive loss that enforces precise positional alignment. Furthermore, we design a novel hybrid 2D feature aggregator to seamlessly distill broad semantic context and fine-grained correspondence features into a single 3D encoder. This process yields a rich, explicit, and position-aware 3D object representation. Inspired

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

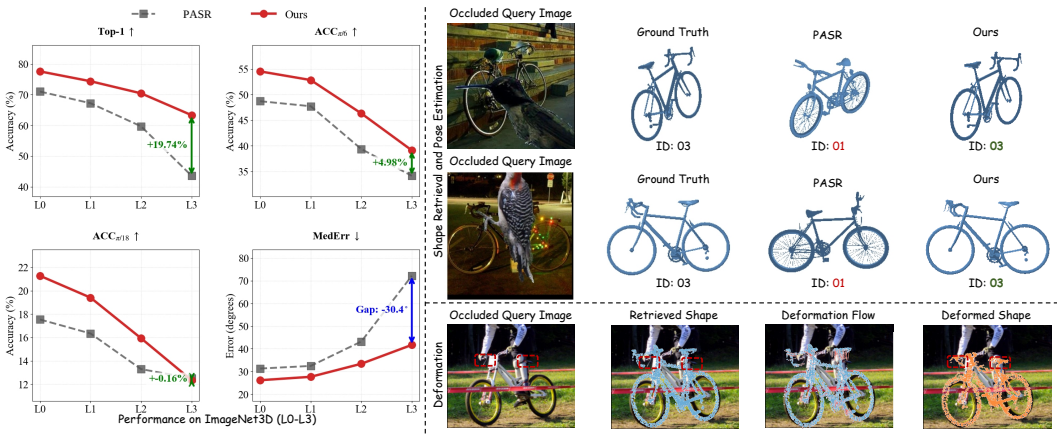

Figure 1: DeformR performs pose-aware retrieval followed by local shape adaptation. In the bicycle example, the input shows a saddle height that differs from the retrieved PASR template; DeformR locally adapts this region while regularization discourages unnecessary changes to the template's overall structure. This illustrates the intended operating regime: local image-supported correction rather than global shape reconstruction. Foundation-model alignment alone does not specify which 3D points to update; pose-aligned part-level correspondences condition the local updates. The plots compare DeformR and PASR across occlusion levels.

by approximate analysis-by-synthesis approaches Wang et al. (2021); Shi et al. (2026), at the inference stage, we further adopt test-time optimization. This process jointly searches for the candidate shape and 3D pose that best align rendered features with the query-image features. Because this synthesis-based inference is inherently sensitive to fine-grained geometry, the framework maintains remarkable robustness against partial occlusions, allowing it to simultaneously excel at both 3D shape retrieval and 3D pose estimation.

Second, to address the limits of rigid proxy matching, DeformR adds conservative image-guided adaptation after pose-aware retrieval and test-time pose refinement. Given a semantically and topologically compatible retrieved CAD template, it makes minimal image-supported local corrections while preserving template structure; retrieval ranking and pose estimates remain fixed. The goal is not to transform a mismatched template into the target, so large global geometry changes are neither expected nor desired. Generating a shape without a suitable template or changing its topology is a distinct reconstruction objective.

Experiments on real-world datasets, Pascal3D and ImageNet3D, show improved retrieval over baselines and competitive pose estimation. On Pascal3D, DeformR reaches 79.90% Top-1 accuracy and 66.14% Acc$_{\pi/18}$, lowering MedErr from 12.10 to 6.37 relative to PASR. On ImageNet3D L3, it exceeds the best retrieval baseline by 19.74 percentage points. With template and pose fixed, deformation improves Pascal3D RS-IoU@Opt against the dataset-provided object mask by 0.76–2.93 points. CAD-proxy CD is a complementary benchmark-shape measure.

Our contributions can be summarized as follows:

- We propose a new 3D shape retrieval framework that combines a hybrid 2D feature aggregator with dense spatial contrastive supervision to distill semantic- and part-level correspondence features into a 3D encoder.
- We introduce a 2D-guided module for local shape adaptation after template and pose selection. Pose-aligned image features condition pointwise updates, while ARAP and Laplacian regularizers discourage unnecessary changes to the retrieved asset's local structure.
- Experiments on Pascal3D and ImageNet3D show improved retrieval over the evaluated baselines. At fixed retrieval and pose, deformation improves overlap with the dataset-provided object mask on Pascal3D; CD to annotated CAD proxies provides a complementary shape-compatibility check.

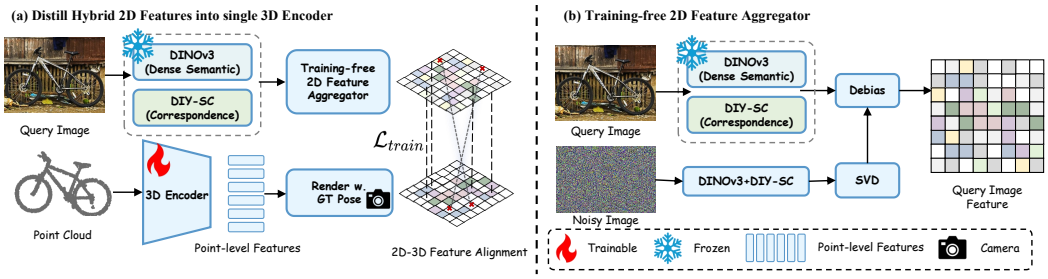

Figure 2: Overall 3D encoder training pipeline. Semantic and spatial cues from multiple 2D foundation models are distilled into one 3D point-cloud encoder.

2 Related Works

3D shape retrieval. 3D shape retrieval from a single view aims to identify the best-matched 3D shape using single-view visual cues Chen et al. (2003); Lin et al. (2021b). Recent approaches primarily focus on learning a shared embedding space for retrieval by leveraging multi-view representations of 3D shapes and corresponding query images Lin et al. (2021b); Nguyen et al. (2022); Hu et al. (2023); Pal et al. (2024); Song et al. (2023; 2025). In addition, large-scale 3D representation models align global features between images and shapes Liu et al. (2023); Zhou et al. (2024); Xue et al. (2023; 2024). Recent retrieval methods also address occlusions and distribution shifts. OPEN learns occlusion-invariant image representations and image–shape correspondences, while Wu et al. evaluate retrieval under occlusion, truncation, and unseen-shape or unseen-object settings Chu et al. (2024). Most retrieval pipelines still select rigid templates and leave instance- and part-level geometric variations unmodeled.

Feature-level analysis-by-synthesis. Feature-level render-and-compare methods refine pose by aligning query-image features with features rendered from 3D representations. NeMo and later coarse-to-fine methods use this strategy for pose estimation Wang et al. (2021); Ma et al. (2022). PASR extends it to pose-aware retrieval by distilling DINOv3 features into a 3D encoder and optimizing candidate poses at test time Shi et al. (2026); DeformR adopts this framework and adds hybrid DINOv3/DIY-SC features with dense spatial contrastive supervision for point-level alignment. SDFit instead initializes a category-level morphable signed-distance field through retrieval and jointly optimizes shape and pose.

Deformation. Template-based deformation methods adapt a source shape toward target geometry Wen et al. (2022); Groueix et al. (2018), and retrieval-augmented methods select a template before deformation Di et al. (2023; 2024). Uy et al. Uy et al. (2021) jointly train retrieval and part-aware deformation so that the retrieval embedding favors candidates that can be adapted to the target. In contrast, DeformR ranks candidates by pose-conditioned feature alignment and uses dense 2D–3D correspondences to adapt the selected template after ranking.

3 Method

We present DeformR, a pose-aware 3D retrieval framework with local shape adaptation for an occluded single-view image. Its deformation stage is intentionally a conservative adaptation module rather than a reconstruction system. Given a semantically and topologically compatible template, pose-aligned image features guide pointwise corrections to local geometric residuals, while geometric regularizers preserve the template structure. Accordingly, large global geometry changes and correspondingly large IoU gains are neither expected nor desired. The pipeline has two stages: (1) pose-aware template retrieval and optimization (Sections 3.2 and 3.3), which selects a compatible template and estimates its pose; and (2) flow-matching shape deformation (Section 3.4), which adapts the selected template under image conditioning. Since deformation follows candidate selection, retrieval and pose accuracy are determined before adaptation and remain unchanged.

3.1 Preliminaries

Problem formulation. Given a real-world query image I, single-view 3D retrieval aims to find the most compatible template from a shape repository $\mathcal{D} = \{\mathcal{T}_1, \mathcal{T}_2, \dots\}$. Each template is represented

as a point cloud $\mathcal{P} = \{p_i\}_{i=1}^{C} \in \mathbb{R}^{C\times 3}$. Because the camera pose is unknown, matching must reason about both geometry and viewpoint. Following analysis-by-synthesis, we compare query features F_q with the pose-conditioned feature maps F_p rendered from candidate shapes. Let $\mathcal{M}$ denote point-level features extracted by a 3D encoder, and let $\mathcal{R}$ be a differentiable renderer that projects $\mathcal{M}$ to 2D under camera pose α. We extract $F_q \in \mathbb{R}^{D\times H\times W}$ using a hybrid 2D encoder.

During training, the known pose α_{gt} supervises alignment between rendered and query features. During inference, we optimize the pose for each candidate shape to minimize feature misalignment. This formulation compares pose-consistent geometric correspondence rather than pixel-level texture and can therefore accommodate appearance variation and clutter.

3.2 Training Setup

Hybrid 2D Feature Extraction and Debiasing. During training, our primary objective is to optimize a 3D encoder that extracts discriminative shape representations through cross-modal knowledge distillation. DINOv3 Siméoni et al. (2025) provides strong semantic features but can exhibit positional bias and limited cross-view correspondence. In contrast, DIY-SC Dünkel et al. (2025) captures fine geometric cues but has less semantic information. We combine their features with a hybrid 2D aggregator as follows.

$$F_q = \lambda_{\text{DINO}}\ \text{DINOv3}(I) + \lambda_{\text{DIY}}\ \text{DIY}(I), \tag{1}$$

The coefficients λ_{DINO} and λ_{DIY} set the relative contributions of DINOv3 and DIY-SC features.

To further reduce positional bias, we apply a lightweight, training-free debiasing step inspired by Cuttano et al. (2026). We feed a noisy image $I_{\mathcal{N}}$ into the hybrid encoder to obtain bias features F_{bias}. Since this bias concentrates in a stable low-rank subspace, we compute its singular value decomposition, $F_{\text{bias}} = U\Sigma V^{\top}$, and take the top-$s$ right singular vectors $B = [v_1, \ldots, v_s]$ as a positional basis. We then remove this component from the query features to obtain debiased features $\hat{F}_q$.

Dense Spatial Contrastive Alignment. To distill knowledge from the 2D foundation models into the 3D encoder Φ_{3D}, we use a differentiable point-based renderer $\mathcal{R}$. Given the ground-truth camera pose α_{gt}, it projects 3D point features into a 2D map $F_p \in \mathbb{R}^{D\times H\times W}$ and aligns rendered features with their query-image locations through $F_p = \mathcal{R}(\mathcal{P}, \mathcal{M}, \alpha_{\text{gt}})$.

We then enforce global correspondence by aligning the projected features F_p with the query image features $\hat{F}_q$. To focus solely on the object, we use a foreground mask $M \in \{0,1\}^{H\times W}$ generated by the renderer. With $\mathcal{S}(\cdot,\cdot)$ defined as cosine similarity, the masked alignment loss is defined as

$$\mathcal{L}_{\text{align}} = \frac{\sum_{u,v}\left(1 - \mathcal{S}\big(\hat{F}_q^{(u,v)}, F_p^{(u,v)}\big)\right) M^{(u,v)}}{\sum_{u,v} M^{(u,v)}}. \tag{2}$$

While $\mathcal{L}_{\text{align}}$ captures overall structure, masked alignment alone may not disambiguate visually similar, repetitive parts, such as a bicycle's front and rear wheels. To improve part-level discrimination, we introduce a dense contrastive loss. For N_{loc} valid spatial locations sampled within the mask M, we compute the similarity matrix between matched 2D–3D pairs as $S_{ij} = \frac{\left(\hat{F}_q^{i}\right)^{\top} F_p^{j}}{\tau_{\text{c}}}$, where τ_{c} is the temperature. We combine the unidirectional InfoNCE loss with the alignment loss as shown below.

$$\mathcal{L}_{\text{contra}} = -\frac{1}{N_{\text{loc}}}\sum_{i=1}^{N_{\text{loc}}} \log \frac{\exp(S_{ii})}{\sum_{j=1}^{N_{\text{loc}}}\exp(S_{ij})}, \quad \mathcal{L}_{\text{train}} = \gamma_1\mathcal{L}_{\text{align}} + \gamma_2\mathcal{L}_{\text{contra}}, \tag{3}$$

where, γ_1 and γ_2 weight the alignment and contrastive losses, respectively.

3.3 Pose-Aware Retrieval and Alignment

Initial search. Since the ground-truth pose α_{gt} is unknown during inference, we first perform a coarse search over candidate shapes and poses to initialize pose estimates and rank templates. We compare the debiased query features $\hat{F}_q$ with 2D features rendered under $N_{\text{pose}} = N_e N_a N_\theta$ poses, uniformly sampled across elevation, azimuth, and in-plane rotation. For each candidate shape $\mathcal{T}$, the

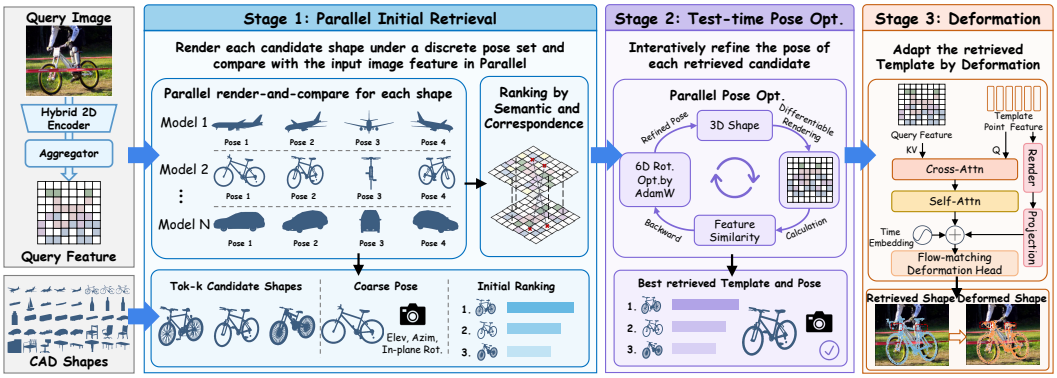

Figure 3: DeformR inference pipeline, including initial top-k retrieval, optional test-time pose optimization, and flow-matching adaptation of the selected template.

Table 1: Pascal3D performance comparison across different models and occlusion levels.

Model	Occ.	Top-1 ↑	Top-5 ↑	$\mathrm{CD}_{\mathrm{CAD}} \times 10^3$ ↓	RS-IoU@Opt ↑	$\mathrm{Acc}_{\pi/6}$ ↑	$\mathrm{Acc}_{\pi/18}$ ↑	MedErr ↓
OpenShape	L0	55.80	96.06	8.88	-	-	-	-
	L1	49.80	94.87	11.09	-	-	-	-
	L2	47.99	93.34	13.89	-	-	-	-
	L3	42.56	89.47	15.97	-	-	-	-
Uni3D	L0	54.01	93.91	8.83	-	-	-	-
	L1	50.25	92.83	10.62	-	-	-	-
	L2	46.73	91.76	11.92	-	-	-	-
	L3	45.84	90.59	13.97	-	-	-	-
SC-IBSR	L0	66.42	93.25	-	-	-	-	-
	L1	58.65	88.40	-	-	-	-	-
	L2	50.58	82.34	-	-	-	-	-
	L3	38.12	68.89	-	-	-	-	-
SAM3D	L0	-	-	-	-	72.17	21.01	17.90
	L1	-	-	-	-	70.86	18.58	18.94
	L2	-	-	-	-	69.37	17.62	19.79
	L3	-	-	-	-	64.85	15.44	21.72
PASR	L0	76.43	98.70	5.50	**76.92**	80.2	40.3	12.1
	L1	73.21	98.57	6.45	44.76	76.2	36.7	13.3
	L2	71.49	97.91	7.78	46.38	68.2	34.3	15.4
	L3	63.05	96.08	10.96	44.38	55.2	25.7	23.9
DeformR	L0	**79.90**	**98.91**	**4.19**	74.33	**88.31**	**66.14**	**6.37**
	L1	**77.55**	**98.76**	**4.42**	**72.60**	**85.75**	**61.76**	**7.13**
	L2	**75.09**	**98.62**	**4.88**	**72.56**	**81.52**	**55.90**	**8.42**
	L3	**70.36**	**97.49**	**6.17**	**69.35**	**72.57**	**46.10**	**11.46**

trained encoder Φ extracts point features $\mathcal{M}$, which the renderer projects as $F_p^{(n)} = \mathcal{R}(\mathcal{P}, \mathcal{M}, \alpha_n)$ for each pose α_n. We select the pose α_{init} that minimizes the alignment loss $\mathcal{L}_{\text{align}}(\hat{F}_q, F_p^{(n)})$. This minimum serves as the candidate's initial ranking score, as shown in panel 3(a).

Test-time pose optimization for top-k candidates. As shown in panel 3(b), we refine the poses of the top-k candidates from the initial search by minimizing $\mathcal{L}_{\text{align}}$ with AdamW. We use a 6D rotation parameterization, which performed best in our pose-encoding ablation.

After pose optimization, we compute the alignment loss for each candidate $\mathcal{T}_i \in \{\mathcal{T}_1, \ldots, \mathcal{T}_k\}$ using its optimized rendered feature map F_p^i. We select the candidate with the lowest loss, defined below.

$$\mathcal{L}(\mathcal{T}_i, I) = \mathcal{L}_{\text{align}}(\hat{F}_q, F_p^i), \qquad \mathcal{T}^* = \arg\min_{\mathcal{T}_i} \mathcal{L}(\mathcal{T}_i, I). \tag{4}$$

Candidate scoring ends when the template $\mathcal{T}^*$ and pose α^* are selected. Deformation then adapts this fixed result without rescoring candidates or re-optimizing pose, so the reported retrieval and pose accuracies remain unchanged after deformation.

3.4 FLOW-MATCHING SHAPE DEFORMATION

Even when semantically and topologically compatible, the retrieved CAD template may differ from the image-visible local geometry. Our lightweight, 2D-guided, pose-conditioned module uses projected image features to condition pointwise updates while retaining the selected template as a geometric anchor.

Deformation architecture. Given the query image I, retrieved point cloud $\mathcal{P}_{\text{src}}$, and estimated pose α^*, we predict a pointwise velocity field for geometric adaptation. We fuse the debiased 2D query features $\hat{F}_q$ and 3D template features $\mathcal{M}$ using cross-attention. Specifically, both are mapped to a shared space, yielding F_q^* and F_m^*; point features serve as queries, and image features provide the cross-attention keys and values.

$$\mathcal{F}_{\text{attn}} = \text{softmax}\left(\frac{(W^Q F_m^*)(W^K F_q^*)^\top}{\sqrt{d}}\right)(W^V F_q^*). \tag{5}$$

Here, W^{Q}, W^{K}, and W^{V} are learned projection matrices, and d is the query/key feature dimension.

A self-attention block and feed-forward layer refine the fused features. We then augment them with pose-conditioned local features F_p^* and a sinusoidal time embedding, which conditions the prediction on interpolation time. An MLP head predicts one 3D velocity per source point, $v_\theta \in \mathbb{R}^{N\times 3}$, where N is the number of source points, as shown in panel 3(c). Integration preserves source-point order without resampling or reordering, so each deformed point retains its index correspondence to the retrieved template.

Flow-matching deformation objective. Rather than directly regressing a single set of offsets, we use conditional flow matching Lipman et al. (2023) to model the template deformation as a continuous trajectory. Here, $\mathcal{P}_{\text{gt}}$ denotes the annotated CAD proxy associated with the training image. During training, $\mathcal{P}_{\text{src}}$ is sampled uniformly from category-matched CAD records in the training split, including the annotated proxy, independently of the retrieval network. At inference, it is the selected Top-1 template. For each source point $x_0 \in \mathcal{P}_{\text{src}}$, we find its nearest target point $x_1 \in \mathcal{P}_{\text{gt}}$ and train the network to predict the corresponding constant velocity.

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,x_0,x_1}\left\|v_\theta(x_t,t,c) - (x_1 - x_0)\right\|_2^2, \tag{6}$$

where, $x_t = (1-t)x_0 + tx_1$ is the interpolated state at time $t \in [0,1]$, sampled uniformly during training. The current state is encoded as a point cloud and projected under the training pose, perturbed by up to 30°, to sample per-point image features; these features, the query feature map, and the template features form the conditioning c. At inference, we integrate the field from x_0 with ten uniform Euler steps, re-encoding the current point cloud and resampling its pose-aligned image features at each step while keeping the query feature map fixed. We apply the auxiliary geometric and projection losses to the single-step endpoint estimate $\hat{x}_1 = x_t + (1-t)v_\theta(x_t,t,c)$.

Geometric and structural regularization. The symmetric Chamfer Distance term ($\mathcal{L}_{\text{CD}}$) regularizes the deformed geometry toward the annotated CAD proxy. It provides a benchmark-level geometric prior for shape compatibility. ARAP ($\mathcal{L}_{\text{ARAP}}$) and Laplacian ($\mathcal{L}_{\text{Lap}}$) regularizers encourage preservation of the retrieved template's local structure and discourage unnecessary geometric changes. We implement the ARAP regularizer as a local edge-vector preservation term on a source-shape k-nearest-neighbor graph with $k = 8$.

We use a differentiable soft intersection-over-union loss between the projected deformed shape and the annotated CAD proxy under the training camera pose. The prediction and target masks, m_p and m_t, are soft occupancy maps rendered from the endpoint point cloud $\hat{\mathcal{P}}_1 = \{\hat{x}_{1,i}\}_{i=1}^N$ and annotated CAD proxy $\mathcal{P}_{\text{gt}}$, respectively. Thus, m_t is the rendered projection of the annotated CAD proxy. Dataset-provided query object masks are used only to evaluate RS-IoU.

$$\mathcal{L}_{\text{proj}} = 1 - \frac{\sum_{u,v} m_p m_t}{\sum_{u,v}(m_p + m_t - m_p m_t) + \epsilon}, \tag{7}$$

Table 2: ImageNet3D performance comparison across different models and occlusion levels

Model	Occ.	Top-1 ↑	Top-5 ↑	$\mathrm{CD}_{\mathrm{CAD}} \times 10^3$ ↓	RS-IoU@Opt ↑	$\mathrm{Acc}_{\pi/6}$ ↑	$\mathrm{Acc}_{\pi/18}$ ↑	MedErr ↓
OpenShape	L0	51.73	87.43	47.98	-	-	-	-
	L1	43.80	82.28	60.04	-	-	-	-
	L2	41.66	78.14	66.72	-	-	-	-
	L3	35.88	72.87	75.71	-	-	-	-
Uni3D	L0	55.17	90.21	44.15	-	-	-	-
	L1	49.91	86.29	28.68	-	-	-	-
	L2	46.05	84.09	36.42	-	-	-	-
	L3	43.56	80.56	44.10	-	-	-	-
SAM3D	L0	-	-	-	-	46.77	16.31	33.88
	L1	-	-	-	-	37.76	14.28	68.60
	L2	-	-	-	-	36.05	13.29	75.68
	L3	-	-	-	-	34.12	**12.50**	83.96
PASR	L0	70.99	87.56	34.02	63.87	48.71	17.54	31.30
	L1	67.19	83.73	38.40	43.50	47.73	16.34	32.46
	L2	59.67	77.69	45.32	40.71	39.36	11.83	43.20
	L3	42.66	61.58	62.83	33.60	26.21	7.07	72.16
DeformR	L0	**77.55**	**94.50**	**15.72**	**64.86**	**54.54**	**21.28**	**26.17**
	L1	**74.35**	**91.61**	**21.72**	**62.20**	**52.85**	**19.41**	**27.67**
	L2	**70.41**	**89.58**	**28.32**	**59.45**	**46.33**	**15.93**	**33.41**
	L3	**63.30**	**83.35**	**43.88**	**54.43**	**39.10**	12.34	**41.73**

where, $\epsilon = 10^{-6}$ stabilizes the denominator and prevents division by zero. The displacement regularizer is $\mathcal{L}_{\mathrm{reg}} = \frac{1}{N}\sum_i \|\hat{x}_{1,i} - x_{0,i}\|_2^2$.

ARAP, Laplacian, and displacement regularizers are not included merely for optimization stability; they encode our design principle of explaining image-supported residuals while minimally disturbing the retrieved CAD geometry. In particular, $\mathcal{L}_{\mathrm{reg}}$ directly penalizes point displacement and biases updates toward minimal corrections.

We combine these terms in the total objective:

$$\mathcal{L}_{\text{deform}} = \gamma_{\text{flow}}\mathcal{L}_{\text{flow}} + \gamma_{\text{proj}}\mathcal{L}_{\text{proj}} + \gamma_{\text{CD}}\mathcal{L}_{\text{CD}} + \gamma_{\text{ARAP}}\mathcal{L}_{\text{ARAP}} + \gamma_{\text{Lap}}\mathcal{L}_{\text{Lap}} + \gamma_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{8}$$

Each γ coefficient is a scalar hyperparameter that weights its corresponding loss term.

4 Experiments

4.1 Setup

Datasets. We evaluate our method on Pascal3D Xiang et al. (2014) and ImageNet3D Ma et al. (2024). We use all categories from Pascal3D and 80 high-quality categories from ImageNet3D. We follow the preprocessing pipeline in PASR Shi et al. (2026) to generate occluded images. We define three levels of occlusion, occluding 2-10% (L1), 10-20% (L2), and 20-40% (L3) of the object's area. The retrieval and pose-estimation methods in the main comparison are not given ground-truth query masks; SAM3D receives a SAM3-predicted mask. The separate reconstruction comparison provides OpenLRM with a ground-truth foreground mask, as described in Appendix Table 7.

Implementation details. For the hybrid 2D encoder, we set $\lambda_{\mathrm{DINO}} = \lambda_{\mathrm{DIY}} = 1$ and $s = 500$. For feature alignment, we set $\gamma_1 = \gamma_2 = 1$ and $\tau_{\mathrm{c}} = 0.07$. For deformation, we set $\gamma_{\mathrm{flow}} = \gamma_{\mathrm{proj}} = \gamma_{\mathrm{CD}} = 1$, $\gamma_{\mathrm{ARAP}} = \gamma_{\mathrm{Lap}} = 0.1$, and $\gamma_{\mathrm{reg}} = 0.01$; the projection term is enabled after a 10-epoch warm-up. Appendix A gives training details; Appendices B and C describe search initialization and pose optimization.

Baselines. We report retrieval results for PASR Shi et al. (2026), SC-IBSR Song et al. (2024), OpenShape Liu et al. (2023), and Uni3D Zhou et al. (2024), and pose results for PASR and SAM3D SAM 3D Team et al. (2025). SAM3D requires a mask, so we provide one predicted by SAM3 Carion et al. (2025); the other reported methods do not use predicted query masks. Appendix I summarizes baseline input and inference protocols.

Metrics. Top-1 and Top-5 retrieval accuracies, pose accuracies, and RS-IoU@Opt are reported as percentages. All reported CD values are multiplied by 10^3. $\mathrm{CD}_{\mathrm{CAD}}$ compares each predicted shape with the normalized CAD proxy annotated to the image and measures benchmark-proxy compatibility. The deformation objective also includes a CD term to this proxy, so post-deformation CD

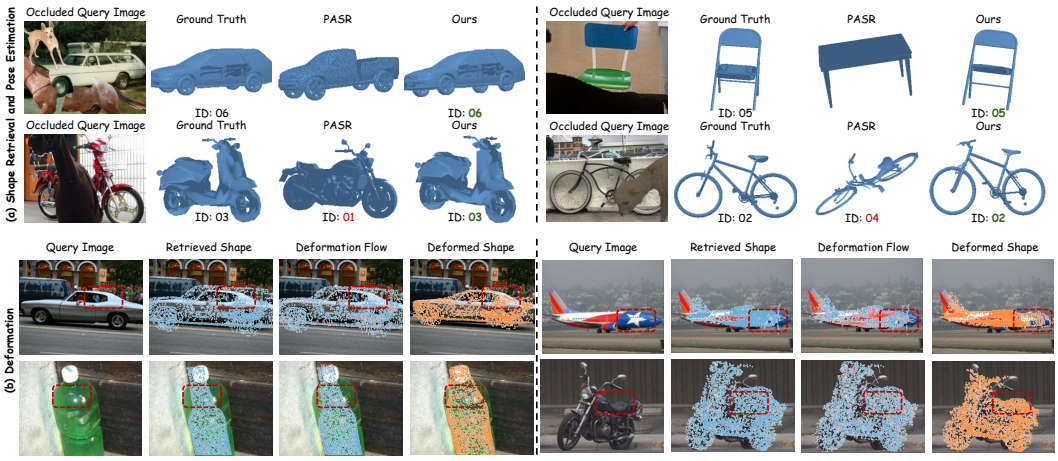

Figure 4: (a) DeformR retrieval and pose estimates compared with PASR on occluded real-world images. (b) Selected templates after local adaptation at the optimized poses.

is reported as a complementary shape measure. RS-IoU@Opt compares the rendered shape with the dataset-provided object mask for each query. This evaluation mask is distinct from the rendered CAD-proxy target used by the differentiable projection loss during training. For DeformR, $\mathrm{CD}_{\mathrm{CAD}}$ and RS-IoU@Opt are reported after deformation, whereas retrieval and pose metrics are determined before deformation and remain unchanged. We evaluate pose with $\mathrm{Acc}_{\pi/6}$, $\mathrm{Acc}_{\pi/18}$, and MedErr.

4.2 Quantitative Results

Pascal3D and Occluded Pascal3D. DeformR achieves the highest Top-1/Top-5 retrieval accuracy and pose accuracy, and the lowest MedErr across all occlusion levels.

ImageNet3D and Occluded ImageNet3D. Table 2 shows that DeformR leads in Top-1/Top-5 retrieval and RS-IoU@Opt and attains the lowest MedErr at each level. At L3, Top-1 exceeds the best retrieval baseline by 19.74 percentage points.

4.3 Qualitative Results

Qualitative results on real-world data. Figure 4 shows selected retrieval, pose, and adaptation cases on occluded images. In these examples, DeformR's templates and poses align more closely with visible regions than PASR, while local adaptation retains the templates' overall structure.

Single-view reconstruction baselines. Table 7 gives a contextual Pascal3D L0 comparison of geometry; selected qualitative comparisons are shown in Appendix Figure 11.

Failure cases. Appendix Figure 10 illustrates representative failure modes under L3 occlusion.

Zero-shot semantics and correspondences. Figure 6 shows qualitative part-clustering and image–shape correspondence examples.

4.4 Ablation Study

Retrieval components. Table 3 reports the feature and pose component ablations on Pascal3D. Dense contrastive alignment raises $\mathrm{Acc}_{\pi/18}$ from 40.30% to 48.08% and lowers MedErr from 12.10 to 10.47. In the hybrid setting, 6D rotation raises Top-1 from 78.87% to 79.66%; debiasing further raises it to 79.90%.

Test-time pose optimization. Figure 5(a,b) shows Top-1 gains of 7.62–10.69 percentage points on Pascal3D and 3.93–5.91 percentage points on ImageNet3D, before deformation.

Deformation. With the Top-1 template and optimized pose fixed, deformation raises Pascal3D RS-IoU@Opt against the dataset-provided object mask by 0.76–2.93 percentage points. These moderate gains are consistent with the intended conservative adaptation regime, which corrects local image-

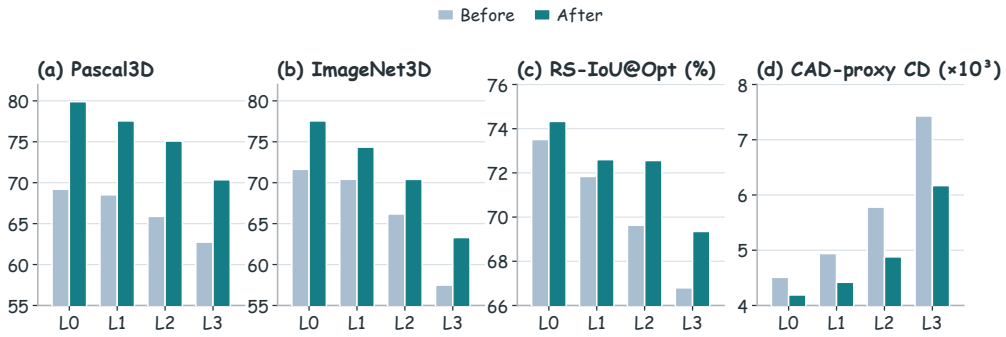

Figure 5: Top-1 accuracy (%) before and after pose optimization on Pascal3D (a) and ImageNet3D (b). Panels (c) and (d) show paired deformation results on Pascal3D with the Top-1 template and optimized pose fixed. Panel (c) reports RS-IoU@Opt (%) against the dataset-provided object mask. Panel (d) reports CAD-proxy CD as a complementary geometry check.

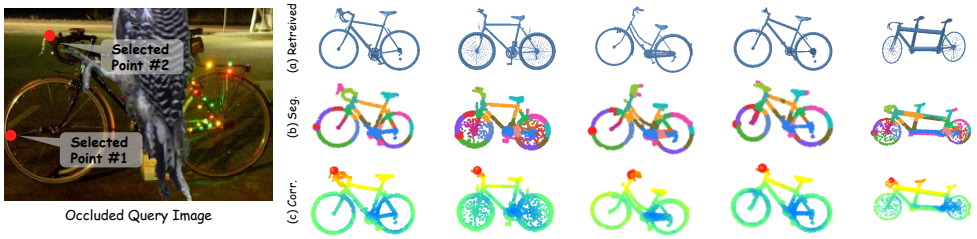

Figure 6: Qualitative semantic and correspondence visualizations. (a) Retrieval from an occluded image. (b) Zero-shot part clustering on the retrieved shape. (c) Similarity maps between selected image points and corresponding 3D points.

Table 3: Component ablation of DeformR on Pascal3D. Pose encodings are E for Euler angles, Q for the half-space quaternion, and R for 6D rotation.

	Components					Metrics			
#	DINOv3	DIY	Pose	CL	Debias	Top-1 ↑	Acc $_{\frac{\pi}{6}}$ ↑	Acc $_{\frac{\pi}{18}}$ ↑	MedErr ↓
1	✓	×	E	×	×	76.43	80.20	40.30	12.10
2	✓	×	E	✓	×	76.92	81.24	48.08	10.47
3	✓	×	Q	✓	×	77.24	83.37	56.50	8.25
4	×	✓	Q	✓	×	77.36	**88.50**	64.48	6.70
5	✓	✓	Q	✓	×	78.87	87.71	64.67	6.62
6	✓	✓	R	✓	×	79.66	87.84	65.70	6.41
7	✓	✓	R	✓	✓	**79.90**	88.31	**66.14**	**6.37**

supported residuals while preserving template structure. Deformation lowers $\mathrm{CD}_{\mathrm{CAD}}$ at every tested occlusion level on both datasets, with ImageNet3D reductions of 0.56, 0.75, 0.95, and 1.99 from L0 to L3. As a point-level diagnostic, one-way nearest-neighbor error on the 20% of source points with the largest pre-deformation errors decreases from 0.1267 to 0.1193 (5.8%), measured on the same point set before and after deformation. Retrieval ranking and pose accuracy remain unchanged.

5 Conclusion

DeformR combines hybrid 2D–3D distillation, dense spatial contrastive learning, and 6D pose optimization for single-view retrieval under occlusion. Given the selected CAD template and pose, its lightweight adaptation conditions pointwise updates on pose-aligned image features, while ARAP and Laplacian regularizers encourage retention of local structure. Experiments on Pascal3D and ImageNet3D show improved retrieval over evaluated baselines and competitive pose estimation. At fixed retrieval and pose, adaptation improves Pascal3D overlap with the dataset-provided object mask; CAD-proxy CD remains a complementary benchmark-shape measure.

AI USE STATEMENT

We used generative AI tools for language polishing and code writing. All AI-assisted text and code were manually reviewed, tested, and verified by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human subjects or personally identifiable information. We are not aware of any significant ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

Details of the proposed method, experimental setup, implementation, and hyperparameters are provided in the main paper and supplementary materials to support reproducibility.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3d model retrieval. *Computer Graphics Forum*, 22(3):223–232, 2003.

Fupeng Chu, Yang Cong, and Ronghan Chen. Open: Occlusion-invariant perception network for single image-based 3d shape retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9):7998–8012, 2024.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. INSID3: Training-free in-context segmentation with DINOv3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yan Di, Chenyangguang Zhang, Ruida Zhang, Fabian Manhardt, Yongzhi Su, Jason Rambach, Didier Stricker, Xiangyang Ji, and Federico Tombari. U-red: Unsupervised 3d shape retrieval and deformation for partial point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8884–8895, 2023.

Yan Di, Chenyangguang Zhang, Chaowei Wang, Ruida Zhang, Guangyao Zhai, Yanyan Li, Bowen Fu, Xiangyang Ji, and Shan Gao. Shapematcher: Self-supervised joint shape canonicalization segmentation retrieval and deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21017–21028, 2024.

Olaf Dünkel, Thomas Wimmer, Christian Theobalt, Christian Rupprecht, and Adam Kortylewski. Do it yourself: Learning semantic correspondence from pseudo-labels. *arXiv preprint arXiv:2506.05312*, 2025.

Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. 3d-coded: 3d correspondences by deep deformation. In *Proceedings of the european conference on computer vision (ECCV)*, pp. 230–246, 2018.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

Nian Hu, Xiangdong Huang, Wenhui Li, Xuanya Li, and An-An Liu. Cross-domain image-object retrieval based on weighted optimal transport. *IEEE Transactions on Multimedia*, 25:9557–9571, 2023.

Ming-Xian Lin, Jie Yang, He Wang, Yu-Kun Lai, Rongfei Jia, Binqiang Zhao, and Lin Gao. Single image 3d shape retrieval via cross-modal instance and category contrastive learning. In *International Conference on Computer Vision (ICCV)*, 2021a.

Ming-Xian Lin, Jie Yang, He Wang, Yu-Kun Lai, Rongfei Jia, Binqiang Zhao, and Lin Gao. Single image 3d shape retrieval via cross-modal instance and category contrastive learning. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 11405–11415, 2021b.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding, 2023.

Wufei Ma, Angtian Wang, Alan Yuille, and Adam Kortylewski. Robust category-level 6d pose estimation with coarse-to-fine rendering of neural features. In *European Conference on Computer Vision*, pp. 492–508. Springer, 2022.

Wufei Ma, Guofeng Zhang, Qihao Liu, Guanning Zeng, Adam Kortylewski, Yaoyao Liu, and Alan Yuille. Imagenet3d: Towards general-purpose object-level 3d understanding. *Advances in Neural Information Processing Systems*, 38, 2024.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. *arXiv preprint arXiv:2504.20024*, 2025.

Van Nguyen Nguyen, Yinlin Hu, Yang Xiao, Mathieu Salzmann, and Vincent Lepetit. Templates for 3d object pose estimation revisited: Generalization to new objects and robustness to occlusions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6771–6780, 2022.

Harsh Pal, Ritwik Khandelwal, Shivam Pande, Biplab Banerjee, and Srikrishna Karanam. Domain adaptive 3d shape retrieval from monocular images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3192–3201, 2024.

SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. Sam 3d: 3dfy anything in images, 2025. URL https://arxiv.org/abs/2511.16624.

Jiaxin Shi, Mingyue Xiang, Hao Sun, Yixuan Huang, and Zhi Weng. Chain of semantics programming in 3d gaussian splatting representation for 3d vision grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24560–24569, 2025.

Jiaxin Shi, Guofeng Zhang, Wufei Ma, Naifu Liang, Adam Kortylewski, and Alan Vuile. Pasr: Pose-aware 3d shape retrieval from occluded single views, 2026. URL https://arxiv.org/abs/2604.22658.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Dan Song, Xue-Jing Jiang, Yue Zhang, Fang-Lue Zhang, Yao Jin, and Yun Zhang. Domain-specific modeling and semantic alignment for image-based 3d model retrieval. *Computers & Graphics*, 115:25–34, 2023.

Dan Song, Shumeng Huo, Xinwei Fu, Chumeng Zhang, Wenhui Li, and An-An Liu. Cross-modal contrastive learning with a style-mixed bridge for single image 3d shape retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2024.

Dan Song, Zekai Qiang, Chumeng Zhang, Lanjun Wang, Qiong Liu, You Yang, and An-An Liu. Adaptive clip for open-domain 3d model retrieval. *Information Processing & Management*, 62 (2):103989, 2025.

Johan WH Tangelder and Remco C Veltkamp. A survey of content based 3d shape retrieval methods. *Proceedings Shape Modeling Applications, 2004.*, pp. 145–156, 2004.

Mikaela Angelina Uy, Vladimir G. Kim, Minhyuk Sung, Noam Aigerman, Siddhartha Chaudhuri, and Leonidas J. Guibas. Joint learning of 3d shape retrieval and deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11713–11722, 2021. doi: 10.1109/CVPR46437.2021.01154.

Angtian Wang, Adam Kortylewski, and Alan Yuille. Nemo: Neural mesh models of contrastive features for robust 3d pose estimation. *arXiv preprint arXiv:2101.12378*, 2021.

Zhenwei Wang, Tengfei Wang, Zexin He, Gerhard Hancke, Ziwei Liu, and Rynson W. H. Lau. Phidias: A generative model for creating 3d content from text, image, and 3d conditions with reference-augmented diffusion. In *International Conference on Learning Representations*, 2025.

Xin Wen, Junsheng Zhou, Yu-Shen Liu, Hua Su, Zhen Dong, and Zhizhong Han. 3d shape reconstruction from 2d images with disentangled attribute flow. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3803–3813, 2022.

Qirui Wu, Daniel Ritchie, Manolis Savva, and Angel X. Chang. Generalizing single-view 3d shape retrieval to occlusions and unseen objects. In *Proceedings of the International Conference on 3D Vision (3DV)*, pp. 893–902, 2024. doi: 10.1109/3DV62453.2024.00060.

Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pp. 75–82. IEEE, 2014.

Yun-Peng Xiao, Yu-Kun Lai, Fang-Lue Zhang, Chunpeng Li, and Lin Gao. A survey on deep geometry learning: From a representation perspective. *Computational Visual Media*, 6(2):113–133, 2020.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1179–1189, 2023.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27091–27101, 2024.

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *International Conference on Learning Representations (ICLR)*, 2024.

A Training Details

The retrieval model is trained for 100 epochs on eight NVIDIA A5000 GPUs, each with 24 GB of VRAM. We employ the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay coefficient of 1×10^{-2}. Each mesh is downsampled to 4,096 points. The initial search uses 192 poses from 4 elevations, 12 azimuths, and 4 in-plane rotations.

The deformation module is trained separately for 300 epochs (batch size 32, AdamW with learning rate and weight decay 1×10^{-4}). Sources are sampled uniformly from same-category training records, with the annotated CAD proxy as target.

The module uses query-image features but no query masks. At test time, the selected Top-1 template is the source shape.

For the architectural backbone, we use PointNeXt as the 3D encoder and PyTorch3D for differentiable rendering. Camera intrinsics and extrinsics follow the dataset specifications. The 2D encoders are DINOv3 ViT-B/16 and DIY-SC `agg_dino_in3d`. We align 2D and projected 3D features at a spatial resolution of 37×37. During test-time pose estimation, we refine the 6D rotation parameters with AdamW for 50 iterations at a learning rate of 0.2.

B Initial Search Setup

To initialize pose estimation and accelerate inference, we search candidate shapes and poses in parallel over a predefined pose set. Figure 7 shows the angle ranges for ImageNet3D and Pascal3D, and Table 4 lists the corresponding search intervals.

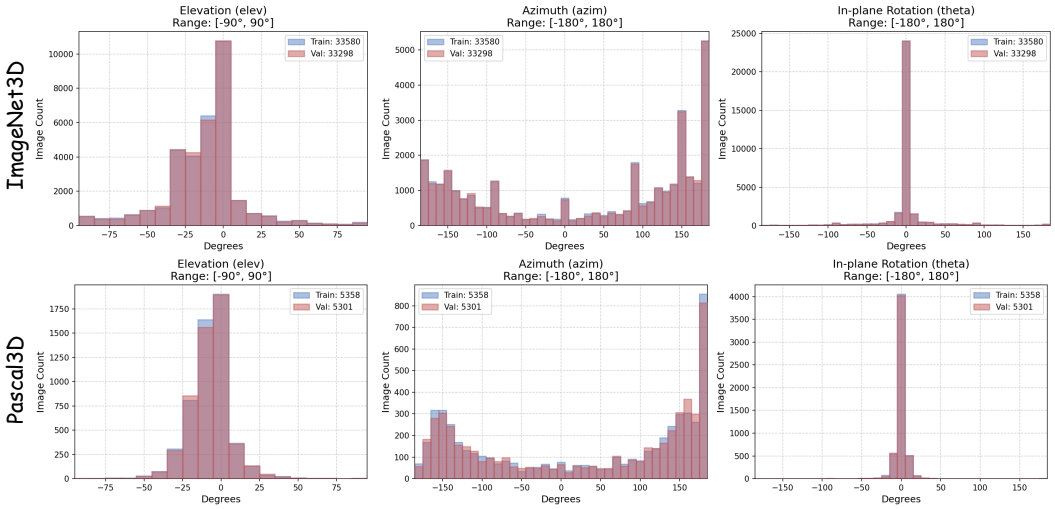

Figure 7: Angle ranges for the ImageNet3D and Pascal3D datasets.

Table 4: Initial search angle ranges in radians for Pascal3D and ImageNet3D.

Parameter (rad)	Pascal3D	ImageNet3D
Elevation	$[-1.57, 1.57]$	$[-1.57, 1.57]$
Azimuth	$[-3.14, 3.14]$	$[-3.14, 3.14]$
In-plane rotation	$[-3.14, 3.14]$	$[-3.14, 3.14]$

C Implementation Details on Pose Estimation

We use the top-5 shapes from the initial search, along with their coarse poses, to initialize pose optimization. To reduce sensitivity to local optima, we augment each pose as follows: $[\text{azim}, \text{elev}, \theta]$, $[\text{azim} + \pi, \text{elev}, \theta]$, $[\text{azim}, \text{elev}, \theta + \pi]$, $[\text{azim} - 1, \text{elev}, \theta]$, $[\text{azim} + 1, \text{elev}, \theta]$, and $[\text{azim} + \frac{\pi}{2}, \text{elev}, \theta]$.

We optimize these six initializations in parallel for each shape. Only the pose is learnable during this step, so point-cloud and image features are reused.

D Additional Correspondence Visualization

Figure 8 shows examples of correspondences between the left and right airplane wings, the front and rear bicycle wheels, the four car tires, and the two sides of a chair.

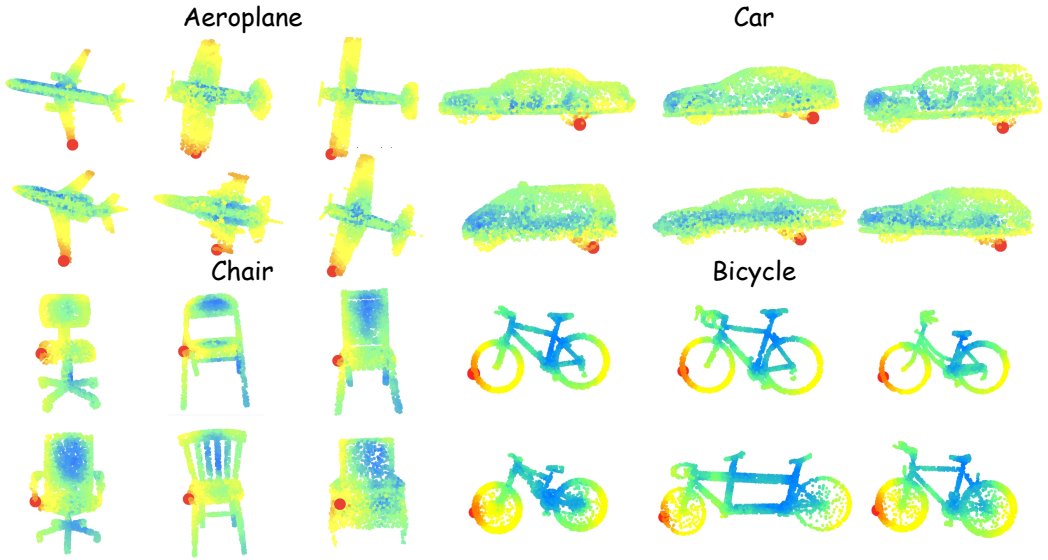

Figure 8: Additional correspondence visualization.

E Ablation on Steps of Test-Time Pose Estimation

Figure 9 shows that additional pose-optimization steps can improve performance, including under severe occlusion.

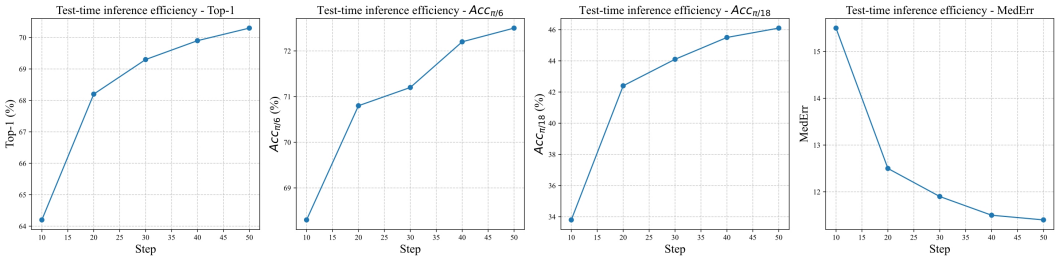

Figure 9: Effect of the number of test-time pose-optimization steps on Pascal3D at L3.

F Failure Case Analysis

Figure 10 shows eight selected L3 residual failures with the annotated CAD excluded from retrieval; row labels identify the query categories. Projected IoU is binary overlap with the dataset-provided object mask at the retrieval pose, distinct from the differentiable CAD-proxy projection loss. Point clouds are rendered at the GT pose.

G Inference Time

On the Pascal3D dataset, the initial search takes approximately 2 minutes, while test-time optimization takes 38 minutes. On ImageNet3D, the initial search takes roughly 30 minutes and test-time

Table 5: Per-query runtime and parameter count.

Method	Stage	Time (s/query)	Parameters
DeformR	Query feature extraction	0.019	–
	Initial search	0.023	–
	Pose optimization	0.430	–
	Deformation	0.055	–
	Total	0.527	186.0M
PASR	Query feature extraction	0.011	–
	Initial search	0.022	–
	Pose optimization	0.679	–
	Total	0.712	87.3M
SAM3D	Total	4.211	3,379M

Table 6: Baseline evaluation protocols.

Method	Predicted mask	Test-time optimization	Large-scale pretraining	Category-agnostic
OpenShape	×	×	✓	✓
Uni3D	×	×	✓	✓
SC-IBSR	×	×	×	✓
CMIC	×	×	×	✓
NeMo	×	✓	×	×
SAM3D	✓	×	✓	✓
PASR	×	✓	×	✓
DeformR	×	✓	×	✓

optimization takes 2.5 hours. These dataset-level durations include metric computation; per-query timings are reported separately below.

H Per-query Runtime and Model Size

Table 5 reports per-query latency and parameter counts where available. All reported runtime measurements, including the dataset-level durations above, were obtained on NVIDIA A5000 GPUs, the same GPU model used for training. Gallery features are precomputed. Table 6 describes the corresponding input and inference protocols.

I Baseline Evaluation Protocols

Table 6 summarizes input and inference protocols for the retrieval and pose-estimation baselines. None receives a ground-truth query mask; only SAM3D uses a SAM3-predicted mask. Dashes denote metrics unavailable for a method.

J Contextual Comparison with Reconstruction Methods

On Pascal3D L0, we compare DeformR with single-view reconstruction methods as shown in 7. OpenLRM receives the ground-truth foreground mask, Phidias uses rembg, and DeformR uses RGB without an additional mask. We compute R-SIoU@GT under the annotated camera pose. Reconstruction outputs are rigidly registered to their annotated CAD proxies using the evaluation's 15-iteration ICP protocol to account for unknown canonical axes. DeformR adapts a compatible CAD asset, whereas the baselines generate shapes from one image.

Figure 11 shows qualitative geometry comparisons between retrieval-preserving local adaptation and image-conditioned reconstruction baselines.

Table 7: Contextual comparison with single-view reconstruction methods on Pascal3D L0.

Method	$\mathrm{CD}_{\mathrm{CAD}} \times 10^3 \downarrow$	R-SIoU@GT (%) ↑	Time (s/query)
DeformR	4.19	88.6	0.53
Phidias Wang et al. (2025)	62.73	43.7	10.23
OpenLRM-S Hong et al. (2023)	131.50	12.1	0.87
OpenLRM-B Hong et al. (2023)	122.45	13.2	0.78
OpenLRM-L Hong et al. (2023)	113.08	14.8	0.91

K List of Selected Categories on ImageNet3D

We select 80 ImageNet3D categories with annotations suitable for training and evaluation. The selected categories are listed below.

hot_air_balloon	sax	ferris_wheel	cooling_tower	wind_turbine
french_horn	go_kart	snowmobile	millwheel	train
beaker	trolleybus	bookshelf	erlenmeyer_flask	billiard_table
interphone	cooker	boat	yurt	dune_buggy
segway	motorbike	aeroplane	riding_mower	tvmonitor
bowling_pin	garbage_truck	glider	light_bulb	lipstick
can	airship	grinder	sewing_machine	skateboard
jinrikisha	power_drill	pillow	violin	projector
chair	clock	racket	piano	lighter
bottle	radiator	sofa	stove	hand_barrow
sink	pay_phone	skate	padlock	stapler
bucket	printer	punching_bag	harvester	unicycle
fan	syringe	telephone	cone	washing_machine
mailbox	table_tennis_table	slipper	tank	blackboard
microwave	teapot	trophy	pan	bow
ax	bench	funnel	cap	camera

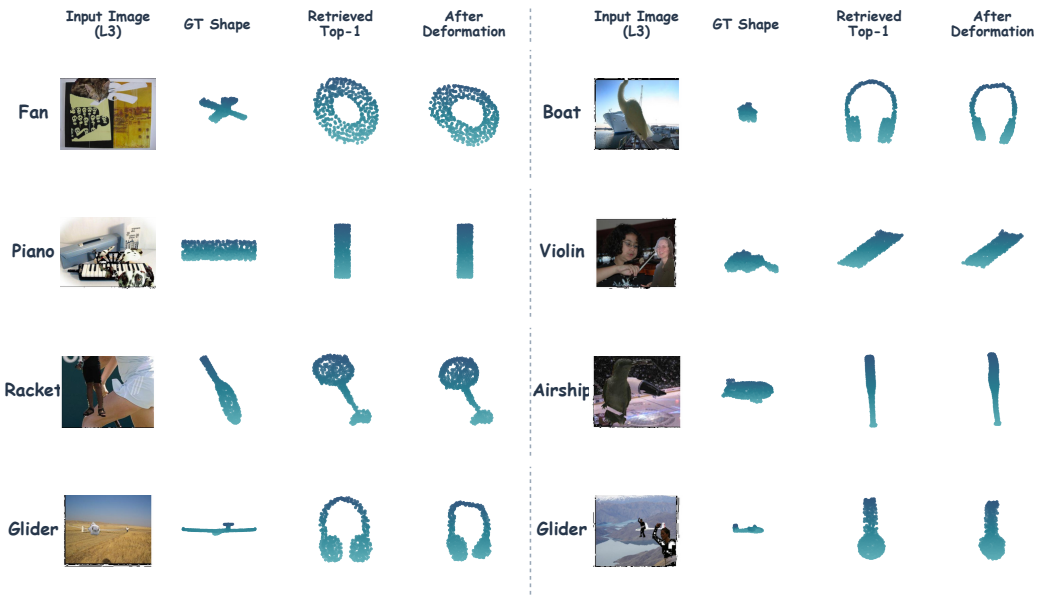

Figure 10: Eight selected L3 failure cases with the annotated CAD excluded from retrieval. Each row places two cases side by side, separated by a dashed line. For each case, the panels show the input image, annotated GT shape, retrieved Top-1 template, and deformed template. Point-cloud color varies with vertical position in the rendered view; all clouds use the GT camera pose. The examples include cross-category retrievals and cases with near-zero projected overlap.

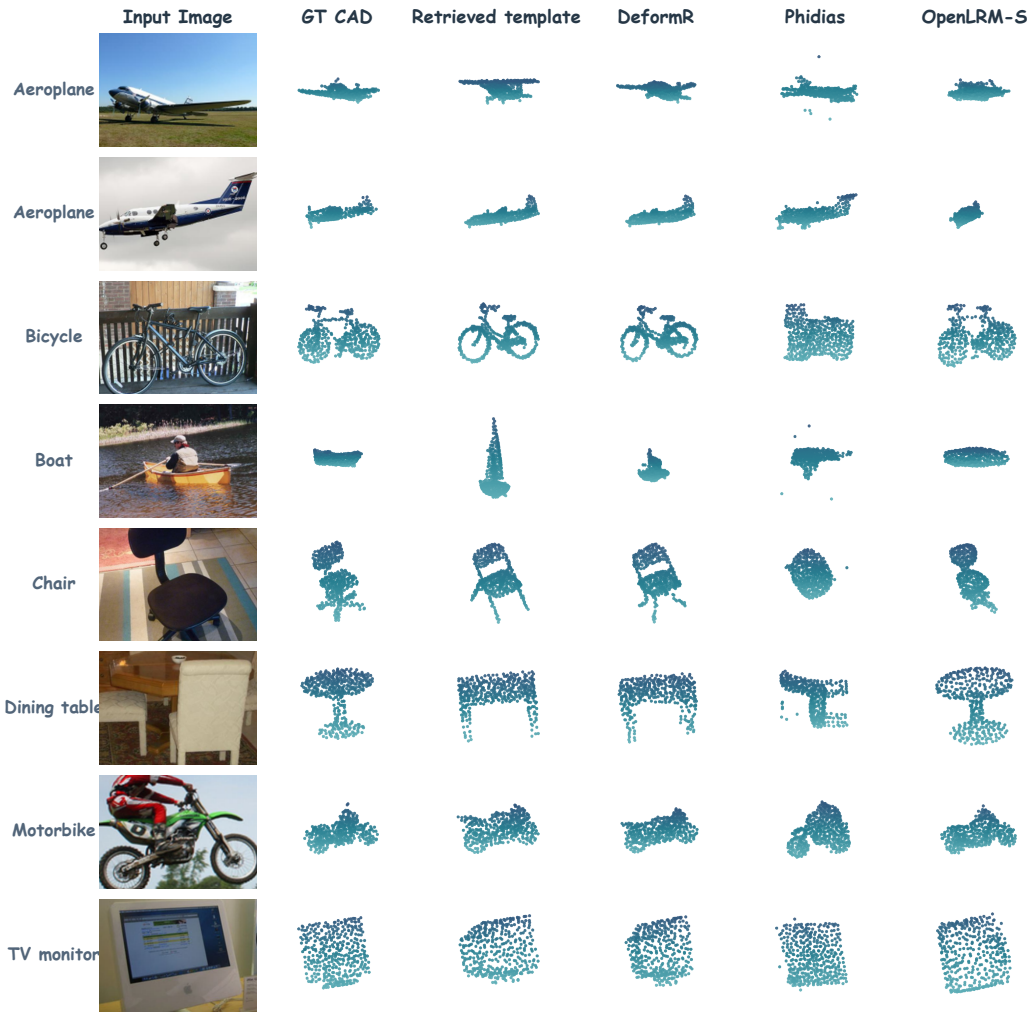

Figure 11: Eight selected Pascal3D L0 examples spanning seven categories. Columns show the input image and shapes from GT CAD, retrieval, DeformR, Phidias, and OpenLRM-S. All point clouds use the annotated GT pose. Reconstruction outputs are independently unit-sphere normalized and rigidly registered to GT CAD for visualization because their canonical axes are unknown. Quantitative R-SIoU follows the 15-iteration ICP evaluation protocol described above. GT, retrieval, and DeformR share a common normalization. OpenLRM-S receives a GT foreground mask, while Phidias uses rembg.