3R-Adapter: Retrieval, Rewiring, and Refinement for Efficient Adaptation of 3D Reconstruction Model

Zixiang Ai[1] **Zhenyu Cui**[2] **Xu Zou**[3] **Jiahuan Zhou**[1,†]
[1]Peking University [2]Tsinghua University [3]Huazhong University of Science and Technology

Abstract

Feed-forward 3D reconstruction models have demonstrated strong generalization under large-scale pre-training, yet they struggle in long-tail scenarios such as low-light reconstruction, joint human-scene reconstruction, and test-time adaptation. We observe relational attention structure collapse, manifested as the failure of cross-view correspondences, as a primary cause of these limitations. While adaptation offers a practical way to bridge this gap, existing parameter-efficient fine-tuning (PEFT) approaches mainly perform channel-wise feature adjustment and lack mechanisms to explicitly adapt relational structures, leaving correspondence failures unresolved. In this work, we formulate adaptation of 3D reconstruction models as a problem of correspondence recovery under domain shift and propose 3R-Adapter, a PEFT paradigm for long-tail 3D reconstruction scenarios. Our method consists of three components: Deformable Retrieval for task-adaptive relation candidate retrieval, Relational Rewiring for reconstructing cross-view relational structures, and Consistency Refinement for injecting the rewired relations into stable predictions. Extensive experiments on low-light reconstruction, joint human-scene reconstruction, and test-time adaptation demonstrate that our approach significantly improves robustness over existing PEFT methods while maintaining efficient lightweight adaptation. Our code is released at https://github.com/PKU-OV3-LAB/3R-Adapter.

1 Introduction

Recent feed-forward 3D reconstruction models [1, 2, 3, 4, 5] have demonstrated strong generalization ability under large-scale pre-training. These models leverage large-scale data [6, 7, 8, 9, 10, 11, 12] to implicitly learn geometric structures through attention-based representations, enabling the estimation of camera poses, intrinsics, and dense geometry from uncalibrated image collections. Despite these advances, current models exhibit clear limitations under domain shift. In scenarios such as low-light environments [13], joint human-scene reconstruction [14, 15, 16], and test-time adaptation [17, 18], data distributions deviate significantly from pre-training conditions and reliable 3D supervision is often unavailable, as depicted in Fig. 1. As a result, pre-trained models frequently fail to generalize, leading to unstable geometric predictions.

We observe that this failure is closely related to the instability of cross-view relational modeling, as shown in Fig. 2. Specifically, attention patterns that are well-structured in in-domain settings become diffuse and less informative under domain shift, making it difficult to establish reliable cross-view correspondences. This suggests that the core challenge lies not in insufficient feature representation, but in the inability to preserve and adapt relational structures across views.

Addressing this issue requires effective adaptation mechanisms that can restore stable cross-view relations. Parameter-efficient fine-tuning (PEFT) [19, 20, 21, 22, 23] provides a practical solution

[†]Corresponding author: jiahuanzhou@pku.edu.cn.
First author homepage: https://azx030512.github.io/.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

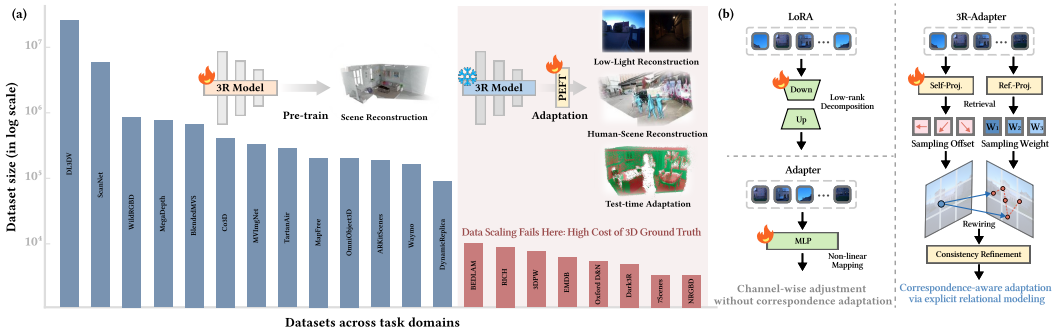


Figure 1: **The data scaling bottleneck in 3D pre-training and our relational adaptation paradigm.** (a) While foundation models thrive on massive data, they struggle to generalize to long-tail 3D domains due to the high cost of dense 3D ground truth. (b) Conventional PEFT methods (e.g., LoRA, Adapter) rely on implicit, channel-wise feature adjustment, which struggle to model the correspondence relationship in data-scarce scenarios. In contrast, our 3R-Adapter shifts the paradigm to explicit relational adaptation via Deformable Retrieval, Relational Rewiring, and Consistency Refinement, effectively establishing robust cross-view correspondences without destroying pre-trained priors.

by enabling lightweight adaptation of large pre-trained models. However, existing PEFT methods, such as LoRA [24] and Adapter [25], primarily perform channel-wise feature modulation and lack mechanisms to explicitly adapt relational structures. Consequently, they rely on implicit adjustments of attention to capture cross-view dependencies, offering limited control over relational routing and often failing to recover stable correspondences under domain shift, demonstrated in Fig. 1.

In this work, we argue that effective adaptation for feed-forward 3D reconstruction requires relational structure recovery. We therefore reformulate model adaptation as a problem of relational adaptation rather than feature reparameterization. Under this perspective, PEFT is not merely a tool for adjusting feature distributions, but a mechanism for introducing task-level structural constraints that explicitly guide cross-view correspondence modeling.

Based on this formulation, we propose 3R-Adapter, a relation-centric PEFT framework for long-tail adaptation. The proposed method consists of three components. Deformable Retrieval, which performs task-adaptive relational querying by selecting appropriate anchor tokens and retrieving candidate cross-view features via deformable sampling. Relational Rewiring, which reconstructs cross-view relational structures by reorganizing retrieved candidates into consistent associations. Consistency Refinement, which injects reconstructed relations into the prediction space to enforce stable and consistent geometric outputs. Together, these modules enable efficient and explicit adaptation of relational structures while preserving the generalization capability of the pre-trained backbone.

In summary, our contributions are threefold:

- We identify relational structure collapse as a fundamental failure mode of feed-forward 3D reconstruction models under domain shift, and show that effective adaptation requires restoring cross-view relational structures rather than merely adjusting features.
- We propose 3R-Adapter, a relation-centric PEFT framework that enables explicit adaptation of cross-view associations through retrieval, rewiring, and refinement. To our knowledge, this is the first work that explores parameter-efficient adaptation specifically designed for feed-forward 3D reconstruction models.
- Extensive experiments demonstrate that, with minimal parameter overhead, 3R-Adapter significantly improves robustness in long-tail scenarios, including joint human-scene reconstruction, low-light conditions, and test-time adaptation.

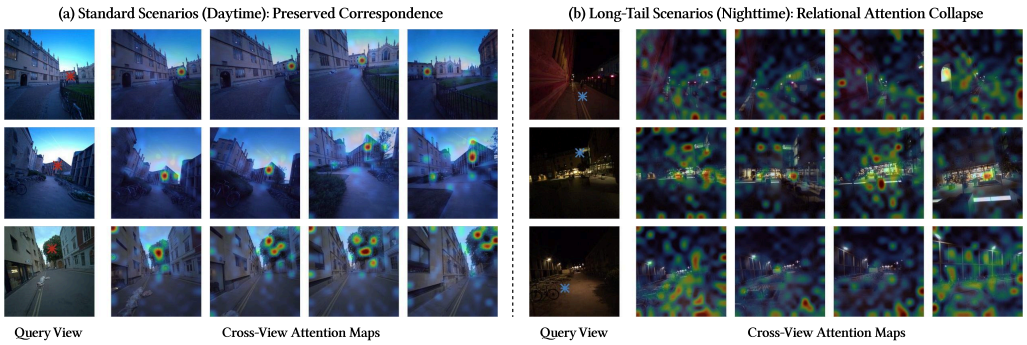


Figure 2: **Relational attention structure collapse in long-tail scenarios.** (a) In data-rich daytime scenes, the pre-trained 3D reconstruction model produces concentrated cross-view attention, allowing a selected query token to attend to consistent physical counterparts across reference views. (b) For selected low-light queries, the attention distribution becomes diffuse and fragmented, indicating degraded relational routing and unreliable correspondence modeling. This observation suggests that long-tail adaptation requires explicitly restoring cross-view relational structures, motivating the design of 3R-Adapter.

2 Observation and Insight

Feed-forward 3D reconstruction models rely on attention-based relational structures to establish cross-view correspondences across input images [26]. While such structures are effectively learned from data-rich pre-training scenarios, their behavior in data-sparse long-tail scenarios remains less reliable. To better understand this failure mode, we visualize cross-view attention patterns of pre-trained reconstruction models in standard daytime scenes and challenging low-light scenes.

As shown in Fig. 2, in standard data-rich scenarios, the attention distribution of a selected query token is concentrated around consistent physical counterparts across reference views. This indicates that the pre-trained model can form stable relational routes when the input distribution is close to the pre-training regime. In contrast, for selected queries under extreme low-light conditions, the attention distribution becomes diffuse and fragmented, making it difficult to maintain reliable cross-view correspondences. We refer to the degradation of geometrically consistent cross-view associations as *relational attention structure collapse*. The quantitative correspondence analysis in Sec. 5.6 measures this geometric degradation and its recovery after adaptation.

This observation suggests that the degradation of feed-forward 3D reconstruction models in long-tail scenarios is not merely caused by weakened local feature representations. A key bottleneck lies in the instability of cross-view relational structures that support correspondence modeling. Therefore, effective adaptation should not only adjust channel-wise feature distributions, but also explicitly restore and reorganize cross-view relations. This motivates our 3R-Adapter, which retrieves task-relevant relational candidates, rewires them into structured cross-view associations, and refines predictions through consistency-preserving residual updates.

3 Related Work

3.1 3D Reconstruction Model

Traditional 3D reconstruction methods like structure-from-motion (SfM) [27] and multi-view stereo (MVS) [28] rely on establishing cross-view correspondences through handcrafted or learned local features, followed by geometrically grounded optimization with strong 3D inductive priors. These pipelines remain sensitive to weak texture, limited overlap, and viewpoint changes, and are computationally costly. Neural surface reconstruction also models geometric uncertainty, as in GURecon [29].

Feed-forward 3D reconstruction models [1, 30, 31, 2, 3, 5, 4] provide a promising alternative by learning geometric reasoning directly from data. Pioneering work DUSt3R [1] reformulates reconstruction as dense pointmap prediction from uncalibrated image pairs. CUT3R [2] introduces a persistent-state recurrent formulation, allowing the model to continuously update a shared 3D

state from streaming observations and produce metric-scale pointmaps online. VGGT [3] further scales feed-forward reconstruction to hundreds of views, directly predicting camera parameters, depth maps, point maps, and point tracks with a unified transformer architecture. More recent approaches such as π^3 [4] remove the reliance on a fixed reference view by adopting permutation-equivariant visual geometry learning, while Depth Anything 3 [5] investigates a minimal formulation based on a single transformer and depth-ray representation for joint any-view depth and pose estimation. These approaches demonstrate that correspondence and geometry can be implicitly learned at scale, marking a shift from explicit optimization to data-driven geometric reasoning.

Despite this progress, existing feed-forward models are primarily optimized for in-distribution performance and exhibit limited robustness under domain shift. In scenarios such as low-light conditions, the learned cross-view correspondences often become unreliable. This highlights the need for effective adaptation mechanisms that can recover stable cross-view relations under distribution shifts.

3.2 Parameter-efficient Fine-tuning

With the increasing scale of models, full fine-tuning has become computationally prohibitive for downstream adaptation. To address this challenge, parameter-efficient fine-tuning (PEFT) methods [19, 24, 25] have been widely explored in both vision and language domains.

Existing PEFT approaches can be broadly categorized into three families. Prompt-based methods, such as visual prompt tuning [19], introduce a small set of learnable tokens to steer pretrained models without modifying backbone parameters. Adapter-based methods [25, 32] insert lightweight bottleneck modules into intermediate layers to refine latent feature representations. Reparameterization methods, including LoRA [24] and its variants such as DoRA [33], approximate weight updates via low-rank decompositions, enabling efficient adaptation with minimal trainable parameters. Recent works have extended these paradigms to feed-forward 3D reconstruction. For example, Test3R [17] and Online3R [18] employ prompt-based adaptation with self-supervised objectives for test-time adaptation on single scenes, while Gen3R [34] introduces lightweight adapters to bridge representation gaps across models. Meanwhile, methods such as VIST3A [35] and VGGRPO [36] adopt LoRA-style reparameterization to enable efficient alignment and adaptation of large geometry models. These works demonstrate that PEFT techniques provide a practical solution for adapting feed-forward reconstruction models under limited supervision.

However, despite their effectiveness, existing methods primarily operate through feature-level modulation and rely on implicit updates of attention to capture cross-view dependencies. In multi-view 3D reconstruction, where geometric reasoning critically depends on structured cross-view relations, such implicit adaptation provides limited control over relational routing and often fails to maintain stable correspondences under domain shift. This limitation motivates our work, where we explicitly model and adapt cross-view relational structures through a relation-centric PEFT framework.

4 Method

In this section, we detail our method for efficiently adapting the pre-trained 3D reconstruction models, improving the correspondence capturing, thus enhancing the reconstruction quality. To explicitly guide cross-view correspondence modeling beyond feature shift in original pre-trained models, we introduce 3R-Adapter with three modules: Deformable Retrieval, Relational Rewiring, and Consistency Refinement, demonstrated in Fig. 3. Given a pre-trained model, we freeze the parameters and only update our newly inserted modules.

4.1 Preliminary: Feed-forward 3D Reconstruction

Given a set of input images $\mathcal{I} = \{I_i\}_{i=1}^{N}$, feed-forward 3D reconstruction models first encode them into patch-level feature tokens. We denote the flattened token set across all views as $\mathbf{X} = \{\mathbf{x}_k\}_{k=1}^{M} \in \mathbb{R}^{M \times C}$, where $M = N \times H \times W$ is the total number of tokens. The core mechanism of these models is to establish cross-view geometric correspondences by learning a relational function $\mathcal{R}_\theta$, yielding contextualized representations:

$$\mathbf{Z} = \mathcal{R}_\theta(\mathbf{X}). \tag{1}$$

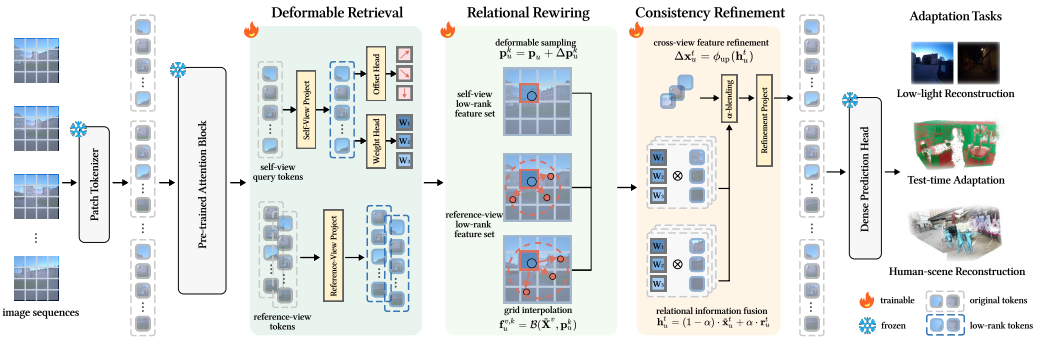


Figure 3: **The overall pipeline of the proposed 3R-Adapter.** To enable efficient adaptation for complex long-tail scenes, our 3R-Adapter modulates the network via three cohesive modules: Deformable Retrieval extracts view-specific anchors and estimates spatial offsets alongside routing weights; Relational Rewiring establishes explicit geometric correspondences across views through deformable spatial sampling; and Consistency Refinement integrates the relational features back into the main stream via a learned gating mechanism, effectively enforcing cross-view consistency.

Recent vision foundation models typically instantiate $\mathcal{R}_\theta$ through two representative architectural paradigms:

Global Attention-based Models. Architectures such as VGGT [3] and π^3 [4] capture cross-view interactions via global attention. For a given token $\mathbf{x}_i$, its relational output $\mathbf{z}_i$ is computed by attending to all context tokens:

$$\mathbf{z}_i = \sum_j \text{Softmax}(\mathbf{Q}_i \mathbf{K}_j^\top)\mathbf{V}_j, \tag{2}$$

where $j \in \{1, \ldots, M\}$.

Recurrent State-based Models. Alternatively, models such as CUT3R [2] maintain a shared latent state to aggregate multi-view information sequentially:

$$\mathbf{S}_t = g_\theta(\mathbf{S}_{t-1}, \mathbf{X}_t), \tag{3}$$

where $\mathbf{S}_t$ encodes accumulated cross-view relations up to time step t. Despite their architectural differences, both paradigms rely on large-scale data to implicitly learn the relational structure $\mathcal{R}_\theta$. Under domain shift in long-tail scenarios, this learned structure often becomes unstable, leading to degraded correspondence estimation. This limitation motivates the need for explicit relational adaptation.

4.2 Deformable Retrieval

A straightforward way to adapt this relation function is to recompute dense interactions among all tokens across all views. However, for N views with feature resolution $H \times W$, dense all-to-all matching introduces quadratic interactions and can also amplify noisy correlations under domain shift. To this end, we design a lightweight Deformable Retrieval module to query task-relevant relational candidates.

Firstly, we restrict the relational search space by introducing a structured sliding window over the view dimension. For a target view t, tokens from $\mathbf{X}^t = \{\mathbf{x}_u^t\}_{u=1}^{HW} \in \mathbb{R}^{HW \times C}$ serve as self-view query tokens, while tokens from neighboring views $\mathbf{X}^v$ are used as reference tokens. Specifically, we restrict reference views by a sliding window over the view dimension:

$$\mathcal{N}(t) = \{v \mid |v - t| \leq w\}, \tag{4}$$

where w is the window radius. This reduces the number of view-level interactions from $O(N^2)$ to $O(Nw)$ and provides a coarse locality prior over the relational graph.

For each self-view query token $\mathbf{x}_u^t$ with spatial position $\mathbf{p}_u$, we first project it into a low-dimensional latent space:

$$\tilde{\mathbf{x}}_u^t = \phi_{self}(\mathbf{x}_u^t) \in \mathbb{R}^r, \tag{5}$$

where $r \ll C$. From this compact query representation, we predict K deformable offsets and their corresponding sampling weights:

$$\{\Delta\hat{\mathbf{p}}_u^k\}_{k=1}^K = \psi_\Delta(\tilde{\mathbf{x}}_u^t), \quad \{\mathbf{w}_u^k\}_{k=1}^K = \psi_w(\tilde{\mathbf{x}}_u^t). \tag{6}$$

In parallel, reference tokens are projected into the same low-rank space:

$$\tilde{\mathbf{X}}^v = \phi_{ref}(\mathbf{X}^v) \in \mathbb{R}^{HW\times r}, \quad v \in \mathcal{N}(t). \tag{7}$$

Thus, Deformable Retrieval defines a sparse set of candidate relation locations $\{\mathbf{p}_u^k\}$ and weights $\{\mathbf{w}_u^k\}$ for each self-view token, while keeping both query and reference features in a lightweight latent space.

Compared with dense attention, Deformable Retrieval introduces only a small number of trainable parameters and provides a structured search space for relational adaptation, making it well-suited for parameter-efficient fine-tuning without disrupting the pre-trained backbone. The sliding window constrains the view-level relational graph, while the deformable offsets provide fine-grained spatial flexibility within selected reference views. The resulting sparse candidates are then materialized and aggregated in the following Relational Rewiring stage.

4.3 Relational Rewiring

Given the offsets and weights predicted by Deformable Retrieval, Relational Rewiring explicitly constructs sparse cross-view relations through deformable sampling and weighted aggregation. For each self-view query token $\mathbf{x}_u^t$ with spatial position $\mathbf{p}_u$, we first obtain K candidate sampling locations:

$$\mathbf{p}_u^k = \mathbf{p}_u + \Delta\mathbf{p}_u^k, \tag{8}$$

where $\Delta\mathbf{p}_u^k$ is the k-th predicted offset.

With the predicted raw offsets $\{\Delta\hat{\mathbf{p}}_u^k\}_{k=1}^K$ in Equation 6, we constrain the search range by applying a bounded transformation:

$$\Delta\mathbf{p}_u^k = \rho \cdot \tanh(\Delta\hat{\mathbf{p}}_u^k), \tag{9}$$

where ρ is a radius hyper-parameter controlling the maximum deformable search range. The candidate sampling location is then computed as:

$$\mathbf{p}_u^k = \mathbf{p}_u + \Delta\mathbf{p}_u^k. \tag{10}$$

After that, we reshape the projected reference feature map $\tilde{\mathbf{X}}^v \in \mathbb{R}^{HW\times r}$ into a spatial feature grid $\tilde{\mathbf{X}}^v \in \mathbb{R}^{H\times W\times r}$ and view it as a continuous feature field via bilinear interpolation, where $v \in \mathcal{N}(t)$ The candidate relational feature is retrieved by:

$$\mathbf{f}_u^{v,k} = \mathcal{B}(\tilde{\mathbf{X}}^v, \mathbf{p}_u^k), \tag{11}$$

where $\mathcal{B}(\cdot,\cdot)$ denotes bilinear sampling on the reference feature grid. The sampled candidates are then aggregated into a rewired relational representation:

$$\mathbf{r}_u^t = \sum_{v\in\mathcal{N}(t)} \sum_{k=1}^{K} \mathbf{w}_u^k \, \mathbf{f}_u^{v,k}, \tag{12}$$

where a_u^k denotes the learned importance weight for the k-th candidate and $\mathbf{r}_u^t \in \mathbb{R}^r$ is the relation-enhanced representation for query token $\mathbf{x}_u^t$.

This process converts the offsets and weights predicted in Deformable Retrieval into explicit cross-view associations. Instead of relying on unconstrained all-to-all attention, Relational Rewiring constructs a sparse and task-adaptive relational graph, enabling more stable correspondence modeling under domain shift.

4.4 Consistency Refinement

After Relational Rewiring, each self-view query token $\mathbf{x}_u^t$ is associated with a rewired relational representation $\mathbf{r}_u^t \in \mathbb{R}^r$. Rather than directly modifying the pre-trained backbone features, we inject this relational information through a lightweight residual branch, which preserves the original representation while enabling task-specific adaptation.

Specifically, we fuse the self-view latent token $\tilde{\mathbf{x}}_u^t$ and the rewired relational representation $\mathbf{r}_u^t$ in the adapter space:

$$\mathbf{h}_u^t = (1-\alpha) \cdot \tilde{\mathbf{x}}_u^t + \alpha \cdot \mathbf{r}_u^t, \tag{13}$$

where α controls the strength of relational blending. The fused representation is then projected back to the original feature dimension:

$$\Delta \mathbf{x}_u^t = \phi_{\text{up}}(\mathbf{h}_u^t) \in \mathbb{R}^C. \tag{14}$$

Finally, the adapter output is added to the frozen backbone feature through a residual connection:

$$\hat{\mathbf{x}}_u^t = \mathbf{x}_u^t + \gamma \Delta \mathbf{x}_u^t, \tag{15}$$

where γ denotes the scaling factor. This residual formulation constrains the adaptation to be a lightweight correction to the pre-trained representation, reducing the risk of disrupting the original geometric priors while allowing the rewired cross-view relations to refine downstream predictions.

5 Experiments

5.1 Adaptation Tasks

We evaluate 3R-Adapter on three representative adaptation settings: low-light reconstruction, joint human-scene reconstruction, and test-time adaptation. These tasks cover diverse forms of domain shift and long-tail scenarios where feed-forward 3D reconstruction models struggle to maintain stable cross-view relations.

Low-light reconstruction. Under low-light conditions, severe image degradation and distribution shifts from pre-training data lead to unreliable feature extraction. Consequently, cross-view correspondences become unstable, making accurate geometric reconstruction challenging.

Joint human-scene reconstruction. This task requires reconstructing both parametric human geometry (e.g., SMPL [37]) and background scene structure within a unified framework. The presence of non-rigid motion and cross-entity interactions introduces significant complexity, making it difficult for pre-trained models to recover consistent geometry and fine-grained human meshes.

Test-time adaptation. In this setting, models must adapt to a single scene without additional supervision, relying on self-supervised signals or internal consistency at test time. This scenario is particularly sensitive to errors in cross-view relations, as no external supervision is available to correct them.

Overall, these tasks share a common challenge: maintaining stable cross-view correspondences under domain shift, making them suitable benchmarks for evaluating relational adaptation.

5.2 Comparing Methods

Our 3R-Adapter is evaluated across multiple feed-forward 3D reconstruction backbones, including DUSt3R [1], CUT3R [2], and π^3 [4], to demonstrate its generality and backbone-agnostic design.

We compare against three categories of methods. First, we include recent feed-forward reconstruction models across all tasks. Second, we consider task-specific approaches tailored to individual settings. For low-light reconstruction, we compare with Dark3R [13], which leverages contrastive learning between low-light and normal images. For joint human-scene reconstruction, we include Josh3R [16], Human3R [14], and UniSH [15], which incorporate specialized designs for modeling human-scene interactions. For test-time adaptation, we compare with Test3R [17] and Online3R [18], which refine predictions via self-supervised objectives on single scenes. Finally, we compare with representative parameter-efficient fine-tuning (PEFT) methods from NLP [33, 38], 2D vision [32, 39], and 3D vision [21]. These baselines share the same training protocols and evaluation settings as ours, enabling a fair comparison of adaptation efficiency and effectiveness.

Table 1: Comparison on the *Oxford Day&Night* [40] test dataset for camera pose estimation and multi-view scene reconstruction.

Method	Publication	Params.	Camera Pose Estimation			Scene Reconstruction			
			ATE ↓	RPE-T ↓	RPE-R ↓	Accuracy ↓	Complete ↓	Chamfer ↓	Normal ↑
DUSt3R [1]	CVPR 24	0.57B	0.937	0.729	10.351	0.741	0.598	1.339	0.526
CUT3R [2]	CVPR 25	0.79B	0.784	0.542	9.924	0.562	0.639	1.201	0.534
VGGT [3]	CVPR 25	1.19B	0.401	0.324	4.236	0.356	0.585	0.941	0.522
Dark3R [13]	CVPR 26	0.69B	0.387	0.285	4.019	0.473	0.424	0.897	0.532
π^3 [4]	ICLR 26	0.96B	0.345	0.261	3.657	0.190	0.746	0.936	0.518
Parameter-efficient Fine-tuning on CUT3R [2]									
DoRA [33]	ICML 24	5.6M	0.565	0.342	6.784	0.439	0.504	0.943	0.541
HydraLoRA [38]	NeurIPS 24	10.0M	0.612	0.457	7.719	0.514	0.578	1.092	0.536
GAPrompt [21]	ICML 25	5.2M	0.577	0.410	6.694	0.467	0.484	0.951	0.532
NightAdapter [32]	ICCV 25	9.5M	0.489	0.310	5.331	0.375	0.416	0.791	0.544
T-LoRA [39]	AAAI 26	7.0M	0.516	0.368	5.362	0.350	0.489	0.839	0.538
3R-Adapter	This Paper	**2.8M**	**0.416**	**0.253**	**4.781**	**0.254**	**0.327**	**0.581**	**0.550**
Parameter-efficient Fine-tuning on π^3 [4]									
DoRA [33]	ICML 24	6.4M	0.376	0.245	3.459	0.198	0.543	0.741	0.524
HydraLoRA [38]	NeurIPS 24	12.0M	0.312	0.260	3.359	0.178	0.624	0.802	0.526
GAPrompt [21]	ICML 25	6.2M	0.396	0.291	3.852	0.215	0.589	0.804	0.518
NightAdapter [32]	ICCV 25	10.3M	0.297	0.225	3.341	0.182	0.560	0.742	0.531
T-LoRA [39]	AAAI 26	7.9M	0.326	0.269	3.340	0.192	0.613	0.805	0.528
3R-Adapter	This Paper	**3.2M**	**0.258**	**0.199**	**3.199**	**0.152**	**0.481**	**0.633**	**0.538**

Table 2: Scene reconstruction under test-time adaptation (TTA) on *7-Scenes* [41] and *NRGBD* [42] datasets. Mean metrics are reported. Additional PEFT baselines are included for comparison.

Method	Publication	Params.	*7-Scenes* [41]				*NRGBD* [42]			
			Accuracy ↓	Complete ↓	Chamfer ↓	Normal ↑	Accuracy ↓	Complete ↓	Chamfer ↓	Normal ↑
DUSt3R [1]	CVPR 24	0.57B	0.146	0.181	0.327	0.736	0.144	0.154	0.299	0.871
MASt3R [30]	CVPR 24	0.63B	0.189	0.211	0.400	0.687	0.085	0.063	0.148	0.794
CUT3R [2]	CVPR 25	0.79B	0.126	0.154	0.280	0.727	0.099	0.076	0.175	0.837
Test3R [17]	NeurIPS 25	0.8M	0.105	0.136	0.241	0.746	0.083	0.079	0.162	0.870
Online3R [18]	CVPR 26	1.2M	0.112	0.145	0.257	0.733	0.084	0.081	0.165	0.829
Parameter-Efficient Fine-Tuning on DUSt3R Backbone										
HydraLoRA [38]	NeurIPS 24	9.3M	0.114	0.148	0.262	0.730	0.095	0.183	0.318	0.869
GAPrompt [21]	ICML 25	4.6M	0.106	0.136	0.242	0.731	0.101	0.179	0.308	0.861
NightAdapter [32]	ICCV 25	8.1M	0.120	0.147	0.267	0.729	0.093	0.142	0.235	0.865
T-LoRA [39]	AAAI 26	6.4M	0.102	0.135	0.237	0.733	0.089	0.131	0.220	0.872
3R-Adapter	This Paper	**2.5M**	**0.094**	**0.122**	**0.216**	**0.738**	**0.082**	**0.123**	**0.205**	**0.879**

5.3 Analysis on Low-light Reconstruction

Table 1 reports results on the *Oxford Day&Night* [40] dataset. Low-light scenes severely degrade visual features and make cross-view matching unreliable, providing a challenging testbed for evaluating whether adaptation can recover stable relational structures.

3R-Adapter achieves the strongest PEFT performance on both CUT3R and π^3 backbones with the fewest trainable parameters. On CUT3R, it reduces Chamfer distance by about **27%** compared with the strongest PEFT baseline, while using only **2.8M** trainable parameters. On the stronger π^3 backbone, 3R-Adapter also consistently improves pose estimation and reconstruction quality, reducing Chamfer distance by about **32%** over the original backbone.

These gains indicate that the improvement does not come from increasing adaptation capacity, but from the proposed relation-centric design. By retrieving sparse relational candidates and rewiring them into structured cross-view associations, 3R-Adapter more effectively restores correspondence under low-light degradation than channel-wise feature adjustment.

5.4 Analysis on Test-time Adaptation

Test-time adaptation evaluates whether a model can refine reconstruction for each test scene without relying on additional supervised training data. This setting is particularly sensitive to unstable cross-view relations, since correspondence errors must be corrected using only scene-level self-supervised signals.

As shown in Table 2, 3R-Adapter achieves the best PEFT performance on both *7-Scenes* and *NRGBD*. It reduces Chamfer distance by about **9%** over the strongest PEFT baseline on *7-Scenes* and by about **7%** on *NRGBD*, while using only **2.5M** trainable parameters. The consistent improvement across two datasets suggests that our method does not overfit to a single scene distribution.

Compared with specialized TTA methods such as Test3R [17] and Online3R [18], 3R-Adapter remains competitive while following a lightweight relation-centric adaptation design. Instead of relying solely on self-supervised refinement at test time, our method explicitly retrieves and rewires reliable cross-view relations, providing a more structured way to improve scene-level consistency.

These results further support our central claim that effective adaptation for feed-forward 3D reconstruction should target relational structures, rather than only modulating channel-wise features.

5.5 Analysis on Joint Human-Scene Reconstruction

Joint human-scene reconstruction requires the model to go beyond standard point-level correspondence and capture structured associations between human proxy tokens, body regions, and surrounding scene context. This makes it a particularly suitable benchmark for evaluating whether relational adaptation can transfer a generic feed-forward reconstruction backbone to more structured downstream tasks.

As shown in Table 3, 3R-Adapter achieves the strongest overall performance among PEFT baselines under the same CUT3R backbone, while using the fewest trainable parameters. This indicates that the improvement does not simply come from larger adaptation capacity, but from explicitly adapting relational structures through retrieval, rewiring, and refinement. Compared with channel-wise PEFT methods, our design is better aligned with the human-scene setting, where reconstruction depends on structured cross-entity relations rather than only local feature adjustment.

3R-Adapter also remains competitive with task-specific joint reconstruction methods, despite operating as a lightweight adaptation module for a generic reconstruction backbone. This suggests that relation-centric PEFT can effectively extend pre-trained 3D reconstruction models to structured downstream tasks without requiring a fully specialized architecture. The qualitative results in Fig. 4 further support this observation: our method produces more coherent human meshes and more consistent scene reconstruction under challenging human-scene interactions, indicating that the adapted relational structures benefit both human and scene modeling.

5.6 Quantitative Cross-View Correspondence Matching Accuracy

We further evaluate the cross-view correspondence matching accuracy of attended or retrieved reference locations with respect to geometry-consistent ground-truth correspondences. Using π^3, we analyze 204 four-view inputs from 51 *Oxford Day&Night* [40] sequences, with correspondences obtained by 3D projection using ground-truth depth and camera geometry. Each 518×518 image contains 1,369 patch tokens on a 37×37 grid, using 14×14-pixel patches.

Match@1 measures the fraction of valid query/reference pairs whose highest-attention token lies within a $w \times w$ patch window around the projected correspondence. Window size w denotes the geometric tolerance width in patch tokens; $w = 1$ requires an exact token match. The widest 7×7 window covers at most 49 tokens (3.6% of the grid) and has a grid radius of three tokens (42 pixels per axis); the default 3×3 window covers at most nine tokens (0.66%). These tolerances therefore retain local geometric constraints. All reported metrics are percentages; adapted Match@1 and deformable offset coverage are averaged over three training seeds.

Table 4(a) shows lower frozen Match@1 at night across all window sizes. At $w = 3$, it falls from 18.49% by day to 11.14% at night, confirming weaker cross-view matching.

Deformable offset coverage measures whether any reference location sampled by 3R-Adapter's learned offsets falls within the same window. Table 4(b) uses four candidates from one layer and 32 across eight layers per query/reference pair, then up to 64 across valid reference views per query. At $w = 3$, coverage is 28.72%, 82.00%, and 86.19%, respectively. Pooling candidates thus provides broader geometric support for rewiring; these coverage rates reflect different budgets and counting units and do not evaluate learned ranking accuracy.

Table 4(c) further shows higher nighttime Match@1 after adaptation in all four reported blocks, using $w = 3$. Block 21 improves from 14.38% to 19.21%, and block 25 from 36.02% to 44.41%. Block 31 changes from 52.94% to 52.16%, indicating that recovery is layer-dependent. These results provide direct evidence that 3R-Adapter recovers geometric associations at intermediate decoder layers, consistent with our relation-centric design.

Table 3: Comparison on the *BEDLAM* [43] validation dataset for joint human-scene reconstruction.

Method	Publication	Params.	Scene Reconstruction			Human Mesh Reconstruction			
			Accuracy ↓	Complete ↓	Chamfer ↓	PA-PVE ↓	Metric-PVE ↓	PA-MPJPE ↓	Metric-MPJPE ↓
DUST3R [1]	CVPR 24	0.57B	0.237	0.159	0.396	-	-	-	-
CUT3R [2]	CVPR 25	0.79B	0.185	0.096	0.281	-	-	-	-
VGGT [3]	CVPR 25	1.19B	0.092	0.044	0.136	-	-	-	-
Joint Human-Scene Reconstruction									
JOSH3R [16]	ICLR 26	0.66B	0.231	0.107	0.338	46.2	245.7	41.5	239.4
Human3R [14]	ICLR 26	1.12B	0.157	0.059	0.216	43.8	222.1	37.2	220.9
UniSH [15]	CVPR 26	0.85B	0.124	0.056	0.180	45.3	231.0	39.6	228.4
Parameter-efficient Fine-tuning on CUT3R [2]									
DoRA [33]	ICML 24	42.5M	0.184	0.086	0.270	46.1	245.1	39.1	236.9
HydraLoRA [38]	NeurIPS 24	48.0M	0.201	0.124	0.325	47.8	258.8	40.5	251.4
NightAdapter [32]	ICCV 25	46.7M	0.167	0.077	0.244	45.2	231.0	38.2	222.0
T-LoRA [39]	AAAI 26	40.3M	0.159	0.068	0.227	45.5	227.2	37.9	225.4
3R-Adapter	This Paper	**36.1M**	**0.135**	**0.052**	**0.187**	**44.4**	**218.5**	**37.6**	**217.2**

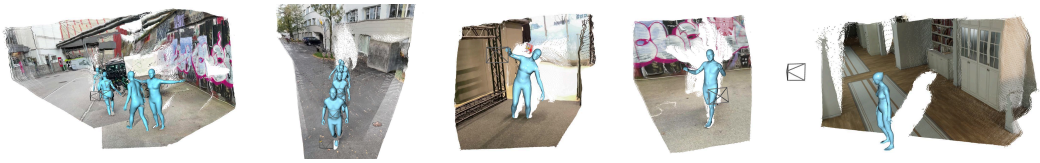


Figure 4: Qualitative results on joint human-scene reconstruction. 3R-Adapter produces more consistent human mesh recovery and cleaner scene geometry under complex human-scene interactions.

Table 4: Cross-view correspondence diagnostics on *Oxford Day&Night* with π^3. (a) Frozen Match@1 pooled over blocks 17/19/21/23. (b) Coverage of 3R-Adapter's deformable retrieval offsets, pooled over day/night and evaluated within one layer, across all eight adapter layers, or across valid reference views. (c) Paired nighttime Match@1 with a fixed 3×3 tolerance.

(a) Frozen Match@1

Window size	Day	Night
1	7.56	3.69
3	18.49	11.14
5	24.38	16.64
7	29.60	22.22

(b) Deformable Offset Coverage

Window size	Single layer	All layers	All references
1	3.71	24.03	31.27
3	28.72	82.00	86.19
5	62.44	91.55	93.50
7	83.26	94.74	96.05

(c) Night Match@1

Block	Frozen	Adapted
19	2.86	**3.85**
21	14.38	**19.21**
23	26.47	**33.19**
25	36.02	**44.41**

6 Conclusion

We present 3R-Adapter, a relation-centric parameter-efficient adaptation framework for feed-forward 3D reconstruction models. Motivated by the observation that pre-trained models exhibit relational attention structure collapse in data-sparse long-tail scenarios, we reformulate adaptation as the recovery of cross-view relational structures rather than channel-wise feature adjustment. Our method retrieves task-relevant relational candidates, rewires them into structured cross-view associations, and refines predictions through consistency-preserving residual updates. Experiments on low-light reconstruction, joint human-scene reconstruction, and test-time adaptation demonstrate that 3R-Adapter improves robustness over existing PEFT methods with minimal trainable parameters. These results suggest that explicitly adapting relational structures is a promising direction for efficient and reliable adaptation of 3D reconstruction foundation models.

Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China (62376011) and the National Key R&D Program of China (2024YFA1410000).

References

[1] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20697–20709, 2024.

[2] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025.

[3] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[4] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025.

[5] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[6] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024.

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[8] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22389, 2024.

[9] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018.

[10] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021.

[11] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, 2023.

[12] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020.

[13] Andrew Y Guo, Anagh Malik, SaiKiran Tedla, Yutong Dai, Yiqian Qin, Zach Salehe, Benjamin Attal, Sotiris Nousias, Kiriakos N Kutulakos, and David B Lindell. Dark3r: Learning structure from motion in the dark. *arXiv preprint arXiv:2603.05330*, 2026.

[14] Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3r: Everyone everywhere all at once. *arXiv preprint arXiv:2510.06219*, 2025.

[15] Mengfei Li, Peng Li, Zheng Zhang, Jiahao Lu, Chengfeng Zhao, Wei Xue, Qifeng Liu, Sida Peng, Wenxiao Zhang, Wenhan Luo, et al. Unish: Unifying scene and human reconstruction in a feed-forward pass. *arXiv preprint arXiv:2601.01222*, 2026.

[16] Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4d human-scene reconstruction in the wild. *arXiv preprint arXiv:2501.02158*, 2025.

[17] Yuheng Yuan, Qiuhong Shen, Shizun Wang, Xingyi Yang, and Xinchao Wang. Test3r: Learning to reconstruct 3d at test time, 2025.

[18] Shunkai Zhou, Zike Yan, Fei Xue, Dong Wu, Yuchen Deng, and Hongbin Zha. Online3r: Online learning for consistent sequential reconstruction based on geometry foundation model. *arXiv preprint arXiv:2604.09480*, 2026.

[19] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[20] Zhengxiang Shi and Aldo Lipani. Dept: Decomposed prompt tuning for parameter-efficient fine-tuning. *arXiv preprint arXiv:2309.05173*, 2023.

[21] Zixiang Ai, Zichen Liu, Yuanhang Lei, Zhenyu Cui, Xu Zou, and Jiahuan Zhou. GAPrompt: Geometry-aware point cloud prompt for 3D vision model. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 796–808. PMLR, 13–19 Jul 2025.

[22] Zixiang Ai, Zichen Liu, and Jiahuan Zhou. Vision graph prompting via semantic low-rank decomposition. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 809–821. PMLR, 13–19 Jul 2025.

[23] Zixiang Ai, Zhenyu Cui, Yufei Guo, Wenwen Qiang, Lei Chen, Jiwen Lu, and Jiahuan Zhou. GAPrompt++: Multi-granular geometry-aware point cloud prompt for 3D vision model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[24] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[25] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[26] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022.

[27] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[28] Michael Goesele, Brian Curless, and Steven M Seitz. Multi-view stereo revisited. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2402–2409. IEEE, 2006.

[29] Zesong Yang, Ru Zhang, Jiale Shi, Zixiang Ai, Boming Zhao, Hujun Bao, Luwei Yang, and Zhaopeng Cui. GURecon: Learning detailed 3D geometric uncertainties for neural surface reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9363–9372, 2025.

[30] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pages 71–91. Springer, 2024.

[31] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pages 78–89. IEEE, 2025.

[32] Qi Bi, Jingjun Yi, Huimin Huang, Hao Zheng, Haolan Zhan, Yawen Huang, Yuexiang Li, Xian Wu, and Yefeng Zheng. Nightadapter: Learning a frequency adapter for generalizable night-time scene segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23838–23849, 2025.

[33] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

[34] Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3r: 3d scene generation meets feed-forward reconstruction. *arXiv preprint arXiv:2601.04090*, 2026.

[35] Hyojun Go, Dominik Narnhofer, Goutam Bhat, Prune Truong, Federico Tombari, and Konrad Schindler. Vist3a: Text-to-3d by stitching a multi-view reconstruction network to a video generator. *arXiv preprint arXiv:2510.13454*, 2025.

[36] Zhaochong An, Orest Kupyn, Théo Uscidda, Andrea Colaco, Karan Ahuja, Serge Belongie, Mar Gonzalez-Franco, and Marta Tintore Gazulla. Vggrpo: Towards world-consistent video generation with 4d latent reward. *arXiv preprint arXiv:2603.26599*, 2026.

[37] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[38] Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Chengzhong Xu. Hydralora: An asymmetric lora architecture for efficient fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[39] Vera Soboleva, Aibek Alanov, Andrey Kuznetsov, and Konstantin Sobolev. T-lora: Single image diffusion model customization without overfitting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9051–9059, 2026.

[40] Zirui Wang, Wenjing Bian, Xinghui Li, Yifu Tao, Jianeng Wang, Maurice Fallon, and Victor Adrian Prisacariu. Seeing in the dark: Benchmarking egocentric 3d vision with the oxford day-and-night dataset. *arXiv preprint arXiv:2506.04224*, 2025.

[41] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013.

[42] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6290–6301, 2022.

[43] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023.

[44] Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024.

A Technical appendices and supplementary material

A.1 Analysis and Discussion

Dense attention offers a flexible way to learn cross-view relations during large-scale pre-training, but it also introduces an unconstrained all-to-all relational search space. In data-sparse long-tail scenarios, directly adapting such dense relations is inefficient and may amplify noisy correspondences. Our Deformable Retrieval can be interpreted as a structured sparse approximation to dense cross-view attention. The sliding window mask provides a weak view-locality prior, while learnable offsets retain spatial flexibility for task-adaptive correspondence search. This design reduces the difficulty of relational discovery without imposing hard geometric constraints.

This perspective also distinguishes 3R-Adapter from conventional PEFT methods. Rather than using trainable parameters only for channel-wise feature adjustment, our method uses them to adapt how cross-view relations are retrieved, rewired, and injected into predictions. Thus, the parameter efficiency of 3R-Adapter comes not from a larger adaptation capacity, but from placing lightweight trainable modules at the relational bottleneck of feed-forward 3D reconstruction.

A.2 Limitation

Although 3R-Adapter effectively alleviates the degradation of feed-forward reconstruction models in data-sparse long-tail scenarios, it still relies on the representational and geometric priors learned by the pre-trained backbone. Our method adapts and reorganizes existing cross-view relational structures rather than learning reconstruction ability from scratch. Thus, when the backbone fails to extract meaningful visual features or the input distribution lies far outside the pre-training regime, relational retrieval and rewiring may remain limited. In addition, the current sliding-window design improves efficiency but may restrict long-range relation modeling in extremely long sequences or scenes with large viewpoint changes. Future work may explore adaptive view selection and hierarchical relational memory for more scalable adaptation.

A.3 Details of Datasets and Metrics

Due to the scarcity of high-quality annotations in long-tail scenarios, we evaluate our method on representative benchmarks that reflect diverse forms of domain shift. We report task-specific metrics for camera pose estimation, scene reconstruction, and human mesh reconstruction. Unless otherwise specified, camera trajectory and scene reconstruction errors are measured in meters, while human reconstruction errors are measured in millimeters.

Low-light reconstruction. We use the *Oxford Day&Night* [40] dataset, which provides calibrated camera poses and semi-dense point clouds captured under varying illumination conditions, including day, dusk, and night. This dataset enables systematic evaluation of reconstruction robustness under severe appearance degradation. For camera pose estimation, we report Absolute Trajectory Error (ATE), Relative Pose Error in translation (RPE-T), and Relative Pose Error in rotation (RPE-R). ATE measures the global trajectory discrepancy after alignment, RPE-T measures relative translation drift between frames, and RPE-R measures relative rotational error. ATE and RPE-T are reported in meters, while RPE-R is reported in degrees.

For scene reconstruction, we report Accuracy, Completeness, Chamfer Distance, and Normal Consistency. Accuracy measures the average distance from predicted points to the ground-truth point cloud, while Completeness measures the average distance from ground-truth points to the prediction. Chamfer Distance is computed as the sum of the two directional distances:

$$\mathrm{CD}(\mathcal{P},\mathcal{G}) = \frac{1}{|\mathcal{P}|}\sum_{\mathbf{p}\in\mathcal{P}}\min_{\mathbf{g}\in\mathcal{G}}\|\mathbf{p}-\mathbf{g}\|_2 + \frac{1}{|\mathcal{G}|}\sum_{\mathbf{g}\in\mathcal{G}}\min_{\mathbf{p}\in\mathcal{P}}\|\mathbf{g}-\mathbf{p}\|_2, \tag{16}$$

where $\mathcal{P}$ and $\mathcal{G}$ denote the predicted and ground-truth point sets. Accuracy, Completeness, and Chamfer Distance are reported in meters. Normal Consistency measures the cosine similarity between predicted and ground-truth surface normals, where higher values indicate better local geometric consistency.

Joint human-scene reconstruction. We adopt the *BEDLAM* [43] dataset, which provides both 3D scene geometry and human SMPL parameters as ground truth. The dataset is synthetically generated

across diverse indoor and outdoor environments, allowing controlled evaluation of human-scene interaction.

For scene reconstruction, we use the same Accuracy, Completeness, Chamfer Distance, and Normal Consistency metrics as described above. For human mesh reconstruction, we report PA-PVE, Metric-PVE, PA-MPJPE, and Metric-MPJPE. PVE measures the mean per-vertex error between predicted and ground-truth human meshes, while MPJPE measures the mean per-joint position error. The prefix PA indicates evaluation after Procrustes alignment, which removes global similarity transformation differences. In contrast, Metric denotes evaluation in the original metric space without alignment, and therefore reflects absolute human localization and the spatial consistency between the reconstructed human body and the scene. All human reconstruction metrics are reported in millimeters.

Test-time adaptation. Following the Test3R/CUT3R sparse-view protocol, we sample every 200 frames on *7-Scenes* [41] and every 500 frames on *NRGBD* [42] to simulate test-time adaptation scenarios. Each sequence therefore contains approximately 3–5 views on *7-Scenes* and 2–4 views on *NRGBD*, requiring the model to reconstruct geometry under self-supervised conditions without additional training data. We evaluate scene reconstruction using Accuracy, Completeness, Chamfer Distance, and Normal Consistency, with distance-based metrics reported in meters.

Cross-view attention diagnostics. GT-valid query/reference pairs are determined by projection and reprojection with dataset depth and camera geometry, and shared by all models. $H_{\mathrm{n}} = -\sum_j p_j \log p_j / \log N$ uses the conditional reference distribution over N patch tokens. The frozen backbone is evaluated once; adapted statistics use three released checkpoints (seeds 1111/1112/1113). All confidence intervals use 10,000 whole-sequence bootstrap replicates stratified by site and day/night, preserving input, layer, and model pairing. They describe sequence sampling within the five observed sites, conditional on the three checkpoints, rather than training-seed population uncertainty. Passive measurement leaves model outputs unchanged.

Table 5: Ablation of components in 3R-Adapter on the *Oxford Day&Night* [40] test dataset for low-light reconstruction.

Deformable Retrieval	Relational Rewiring	Consistency Refinement	Accuracy ↓	Complete ↓	Chamfer ↓	Normal ↑
-	-	-	0.562	0.639	1.201	0.534
✓	-	-	0.540	0.612	1.152	0.534
✓	✓	-	0.397	0.445	0.842	0.542
-	-	✓	0.483	0.529	1.012	0.536
✓	✓	✓	**0.254**	**0.327**	**0.581**	**0.550**

A.4 Detailed Ablation Study

Ablation on main components. Table 5 studies the contribution of each component in 3R-Adapter on the Oxford Day&Night dataset using CUT3R as the backbone. The vanilla CUT3R baseline obtains a Chamfer Distance of 1.201, indicating substantial degradation under low-light conditions.

Adding Deformable Retrieval alone slightly improves the reconstruction results, reducing Chamfer distance from 1.201 to 1.152. This suggests that sparse candidate querying can reduce noisy dense interactions, but retrieving candidate locations alone is insufficient to recover stable cross-view structure. When Relational Rewiring is further introduced, performance improves significantly, with Chamfer distance reduced to 0.842 and Accuracy improved from 0.540 to 0.397. This confirms that the key benefit comes from explicitly reorganizing retrieved candidates into structured cross-view relations.

Consistency Refinement also provides clear gains when used alone, improving Chamfer distance from 1.201 to 1.012. This shows that lightweight residual correction can stabilize predictions even without explicit retrieval and rewiring. However, the full 3R-Adapter achieves the best results across all metrics, reducing Chamfer distance to **0.581** and improving Normal consistency to **0.550**. These results demonstrate the complementary roles of the three components: Deformable Retrieval identifies reliable candidates, Relational Rewiring reconstructs cross-view associations, and Consistency Refinement safely injects the adapted relations into the prediction stream.

Ablation on Hyperparameters. We further analyze the sensitivity of 3R-Adapter to four key hyperparameters: the number of sampling locations K, the search radius ρ, the relational fusion factor α, and the residual scaling factor γ. We report Chamfer Distance on the *Oxford Day&Night* dataset, where lower values indicate better reconstruction quality.

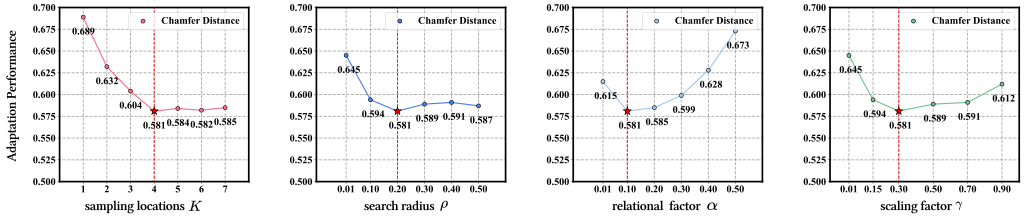


Figure 5: Ablation on the hyperparameters of 3R-Adapter on the *Oxford Day&Night* [40] test dataset for scene reconstruction.

As shown in Fig. 5, increasing the number of sampling locations from $K = 1$ to $K = 4$ consistently improves performance, suggesting that multiple candidates provide richer relational evidence for cross-view reconstruction. However, using more than four samples brings marginal gains and may introduce redundant or noisy candidates. The search radius ρ exhibits a similar trend. A moderate radius improves local relation discovery, with the best performance achieved at $\rho = 0.20$, while an overly large radius tends to include irrelevant distant features.

The relational fusion factor α controls the contribution of rewired cross-view relations. The model performs best at $\alpha = 0.10$, while larger values degrade performance, indicating that excessive relational injection can disturb the pre-trained feature distribution. The residual scaling factor γ shows a clear balance between adaptation strength and prior preservation, with the best result obtained at $\gamma = 0.30$.

Overall, 3R-Adapter achieves the best configuration with $K = 4$, $\rho = 0.20$, $\alpha = 0.10$, and $\gamma = 0.30$, which is used as the default setting in our experiments.

A.5 Implementation Details

We build 3R-Adapter on top of pretrained feed-forward 3D reconstruction backbones, including DUSt3R [1], CUT3R [2], and π^3 [4]. Across all experiments, the pretrained backbone is frozen, and only the lightweight adapter parameters are updated. This setting allows the model to preserve geometric priors learned from large-scale pre-training while adapting to data-sparse long-tail scenarios with minimal trainable parameters. We optimize the adapter modules using AdamW. The hyperparameters of 3R-Adapter are set according to the ablation study: the number of deformable sampling points is $K = 4$, the search radius is $\rho = 0.20$, the relational fusion factor is $\alpha = 0.10$, and the residual scaling factor is $\gamma = 0.30$. All experiments are conducted on 8 NVIDIA RTX 5090 GPUs. For low-light reconstruction and joint human-scene reconstruction, we train the adapters on the corresponding training splits and report results on the held-out evaluation sets. For test-time adaptation, the adapter is optimized independently on each test scene using self-supervised reconstruction loss [17], without accessing additional training data or ground-truth geometry.

A.6 Training Objectives

We optimize 3R-Adapter using the standard objective of each adaptation task, while keeping the pre-trained backbone frozen. This ensures that the comparison focuses on the adaptation mechanism rather than additional supervision.

For low-light reconstruction, we use geometric reconstruction losses on predicted point maps or depth representations, together with pose supervision when camera trajectories are available:

$$\mathcal{L}_{\text{low}} = \mathcal{L}_{\text{geo}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}}. \tag{17}$$

For joint human-scene reconstruction, we jointly supervise scene geometry and human body reconstruction:

$$\mathcal{L}_{\text{hs}} = \mathcal{L}_{\text{scene}} + \lambda_{\text{mesh}}\mathcal{L}_{\text{mesh}} + \lambda_{\text{joint}}\mathcal{L}_{\text{joint}} + \lambda_{\text{smpl}}\mathcal{L}_{\text{smpl}}. \tag{18}$$

Following Human3R [14], we incorporate human prior tokens from Multi-HMR [44] and fuse them with reconstruction tokens to enhance human-aware relational modeling. These tokens provide body-structure cues that help the model associate human proxy representations with body regions, while 3R-Adapter adapts the cross-view relational structures of the reconstruction backbone.

For test-time adaptation, we follow the self-supervised objective of Test3R [17]. Given sparse input views, the model predicts pointmaps under different reference-view choices. Since these predictions correspond to the same underlying scene geometry, we enforce consistency among pointmaps generated from different references:

$$\mathcal{L}_{\text{tta}} = \sum_{r_1 \neq r_2} \left\| \mathbf{P}^{(r_1)} - \mathbf{P}^{(r_2)} \right\|_1, \tag{19}$$

where $\mathbf{P}^{(r)}$ denotes the predicted pointmap when view r is selected as the reference view. This self-supervised loss enables scene-level adaptation without ground-truth geometry.

A.7 Broader Impact

3R-Adapter aims to improve the adaptability and robustness of feed-forward 3D reconstruction models in data-sparse scenarios, such as low-light environments, joint human-scene reconstruction, and test-time adaptation. By reducing the number of trainable parameters required for downstream adaptation, our method may lower the computational and data costs of deploying 3D reconstruction systems in practical settings, including robotics, augmented reality, digital content creation, and embodied intelligence.

At the same time, improved 3D reconstruction capabilities may raise privacy and safety concerns when applied to real-world environments or human-centered scenes. In particular, reconstructing human geometry or private spaces from visual inputs should be conducted with appropriate consent, data protection, and usage constraints. Our work focuses on technical improvements for efficient adaptation and does not introduce new data collection procedures. We encourage future applications of such systems to follow responsible data governance practices and to carefully consider potential misuse in surveillance or unauthorized 3D reconstruction.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes].

   Justification: The abstract and introduction summarize the proposed 3R-Adapter framework, its parameter-efficient adaptation setting, and its empirical scope on long-tail 3D reconstruction scenarios. The claims are supported by the method description and experimental results in Sections 4 and 5, and we avoid claiming generalization beyond the evaluated backbones, datasets, and adaptation settings.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes].

   Justification: We discuss limitations in Section A.2, including the dependence on pretrained feed-forward reconstruction backbones, and the scope of tested long-tail domains.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A].

Justification: The paper presents a model architecture, task objectives, and empirical evaluations, but does not claim formal theoretical results requiring proofs.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes].

Justification: Sections 4 and 5 describe the architecture, task objectives, benchmark protocols, and evaluation metrics. Appendix A.5 provides implementation details, and the released code includes training and evaluation entrypoints with release configurations for the main experiments.

Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes].

   Justification: Our code is released at https://github.com/PKU-OV3-LAB/3R-Adapter, with training and evaluation commands, release configurations, and checkpoint documentation. The datasets and pretrained backbones are available from their original sources, subject to their respective access conditions and licenses.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes].

   Justification: Section 5 and Appendix A.5 describe the benchmark splits, input sampling, optimizer, task objectives, frozen and trainable modules, evaluation protocols, and default adapter hyperparameters. Detailed training configurations are provided with the released code.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No].

Justification: Table 4 reports means over three seeds for adapted correspondence accuracy and deformable offset coverage. However, the paper does not report error bars or confidence intervals for all main reconstruction experiments; a full characterization of variability across every task is not provided.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [No].

   Justification: Appendix A.5 identifies the use of eight NVIDIA RTX 5090 GPUs, but does not provide per-experiment memory usage, execution times, or the total compute budget including preliminary experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes].

   Justification: The research uses publicly available datasets and pretrained models under their respective terms, and the work is conducted for scientific evaluation of 3D reconstruction and adaptation methods. We do not collect private user data, perform human-subject experiments, or deploy the system in a real-world decision-making setting.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes].

    Justification: Appendix A.7 discusses positive impacts such as improving efficient adaptation for 3D perception in robotics, AR/VR, and digital content creation, as well as potential risks such as misuse in surveillance, privacy-sensitive reconstruction, or unsafe deployment under severe domain shifts. We also recommend responsible use with appropriate data consent, privacy safeguards, and domain-specific validation before deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A].

    Justification: The paper does not introduce a high-risk pretrained generative model, language model, or scraped dataset. It presents an adaptation technique for existing 3D reconstruction backbones; use of the released code and checkpoints remains subject to the terms of the underlying models and datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [No].

    Justification: We cite the original datasets, pretrained backbones, and baseline methods, but the paper does not explicitly enumerate the license names and terms for every asset. Readers should consult the original distributions for their applicable terms.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes].

    Justification: The released implementation provides installation instructions, task entrypoints, release configurations, and checkpoint loading documentation. The paper describes the adapter design and its limitations in Sections 4 and A.2; no new dataset is introduced.

    Guidelines:

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

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

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

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted on existing public datasets and benchmarks.

    Guidelines:

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

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

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

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or human-subject studies, and therefore IRB approval or equivalent review is not applicable. The human-related datasets used in the experiments are existing datasets, and we follow their original terms of use.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A].

    Justification: The core method does not use LLMs as an important, original, or non-standard component. Any use of LLM-based tools, if any, was limited to writing, editing, or formatting assistance and did not affect the methodology, experiments, scientific claims, or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.