LIGHTWEIGHT TEST-TIME GEOMETRY UNIFICATION FOR MULTI-VIEW RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Feed-forward 3D reconstruction models can predict several offset sheets for one physical surface. Reconciling these predictions requires both cross-view agreement and preservation of local surface structure: independent voxel shifts can narrow layers while introducing cracks between neighboring pixels. We introduce a self-supervised test-time refiner that addresses these coupled requirements without target-scene depth or 3D reference supervision. Closed-form per-view similarity alignment first corrects rigid-plus-scale drift. A 0.9 M-parameter residual network then corrects the remaining non-rigid component of multi-view geometric inconsistency, using multi-view image features and geometric self-supervision with displacement-field regularization to preserve local structure. Wall-corner and curved-edge comparisons show that this second stage brings residual sheets closer together while retaining surface continuity. It reduces cross-view centroid gaps in all 53 evaluated NRGBD/DTU scene–model cases. On $6{,}658$ patches from six NRGBD scenes, its adjacent-pixel GT edge-vector error is lower than full-strength direct ($\alpha = 1$) and smoothed voxel shifts in 94.7% and 87.5% of cases, respectively. The same refiner applies to Pi3 and VGGT-Ω without per-base retraining, improving Accuracy and Normal Consistency across the evaluated object-level, indoor, and outdoor benchmarks. The complete pipeline takes about 3.7 minutes for a representative 15-view scene on one A800 GPU. A fixed-count analysis shows that redundant layers can lower Completeness error while raising Accuracy error, motivating evaluation of surface structure alongside point distances.

1 INTRODUCTION

Dense 3D reconstruction from a set of uncalibrated images underpins spatial computing and embodied robotics, and has long been addressed by staged pipelines that cascade SfM and MVS (Schönberger & Frahm, 2016; Yao et al., 2018). Feed-forward 3D foundation models provide a direct-regression alternative to this staged pipeline: from pairwise point-map regression (Wang et al., 2024; Leroy et al., 2024) to single-pass models that predict the point maps and cameras of all views at once (Wang et al., 2025a;c), coupling views through cross attention.

Despite global attention across views (Wang et al., 2025a;c), the final geometry is decoded per view without an explicit constraint enforcing coincident predictions for overlapping observations. This residual disagreement can cause a physical surface to appear as several offset, parallel sheets—a "layering" artifact—and can introduce "flying points" near occlusion boundaries (Figure 1). These duplicated sheets can degrade estimated surface normals and propagate geometric artifacts to downstream meshing, mapping, and object tracking (Hu et al., 2026).

Classical SfM + MVS pipelines impose cross-view agreement through bundle adjustment (Schönberger & Frahm, 2016) and depth fusion. Feed-forward models can retain residual disagreement after a single pass. Correspondence-based registration depends on reliable matches and primarily corrects rigid drift; coordinate-only denoisers do not use view provenance to distinguish layering from local surface noise. Photometric optimization obtains additional constraints through rendering. Our approach uses geometric disagreement among overlapping predictions as the test-time supervisory signal, optimizing for a few hundred steps without rendering, feature matching, or target-scene depth and 3D ground truth.

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.

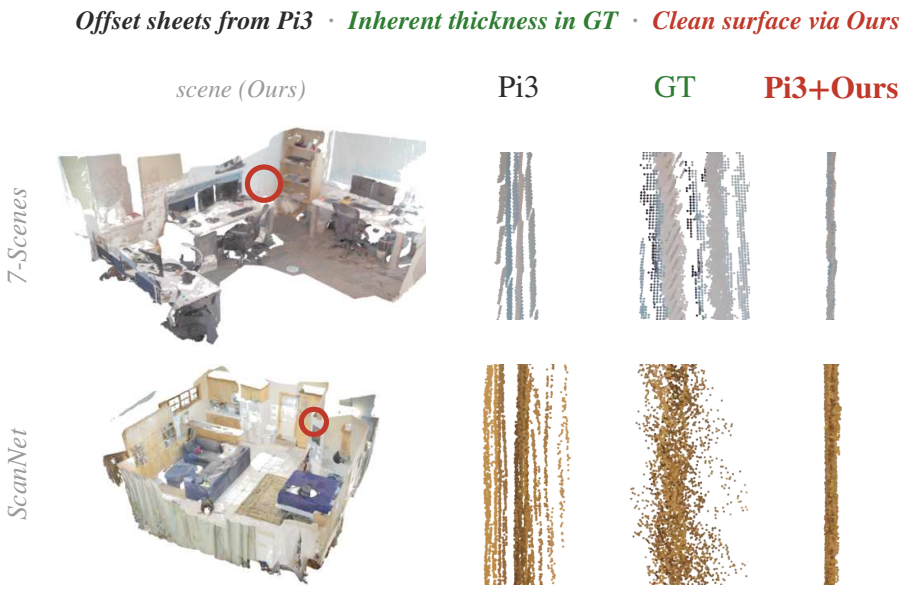

Figure 1: Feed-forward reconstruction creates offset surface "sheets." Along-normal sections in 7-Scenes (*top*) and ScanNet (*bottom*) show multiple Pi3 sheets; ours compresses them into a single thin band. The back-projected RGB-D references exhibit visible thickness and noise.

Stage 1 fits a closed-form per-view similarity in the predicted frame. Stage 2 corrects the remaining non-rigid component of multi-view geometric inconsistency with multi-view image features and regularized displacements (Section 3.4). The base model and tokenizer stay frozen, and each point retains its source pixel.

Our central objective is to bring overlapping predictions into agreement while preserving the local surface structure traced by neighboring pixels. One similarity moves a whole view together, leaving position-dependent offsets across a wall or corner. Figure 4 makes Stage 2's role explicit: Align retains separated sheets, Direct and Direct + smooth introduce surface breaks, and our residual narrows the separation while retaining continuous local surfaces. Scene-mean point distances alone do not describe this combination of cross-view agreement and within-view continuity. We therefore evaluate cross-view gaps, local GT edge vectors, and corresponding-pixel positions together.

The same refiner improves Accuracy and Normal Consistency for Pi3 and VGGT-Ω without per-base retraining. Beyond alignment, Stage 2 reduces cross-view gaps in all 53 assessed NRGBD/DTU scene–model cases. On 6,658 patches from six NRGBD scenes, its adjacent-pixel GT edge error is lower than full-strength Direct and Direct + smooth in 94.7% and 87.5% of patches (Section 4.3). A fixed-count test examines layering's effect on distance metrics (Section 4.4).

The primary contributions of this work are:

- A **test-time geometry unification approach** that reduces layering while preserving local surface structure, without target-scene depth or 3D reference supervision.
- A **two-stage formulation** combining closed-form alignment with image-conditioned correction of non-rigid multi-view differences and displacement-field regularization.
- A **joint evaluation of agreement, shape, and position**, with visual controls, complete-cohort diagnostics, and a fixed-count analysis of layering's effect on distance metrics.

2 Related Work

2.1 Feed-Forward 3D Reconstruction

Multi-view reconstruction has moved from SfM and MVS pipelines (Schönberger & Frahm, 2016; Yao et al., 2018; Gu et al., 2020) to foundation models that regress geometry from uncalibrated images. DUSt3R and MASt3R (Wang et al., 2024; Leroy et al., 2024) predict pairwise point maps; VGGT, VGGT-Ω, and Pi3 (Wang et al., 2025a; 2026; 2025c) jointly process multiple views, while DA3 (Lin et al., 2025) predicts a depth-ray field. Global attention, compact registers, spatial memory, and recurrent states exchange cross-view information (Wang et al., 2025a; 2026; Wang & Agapito, 2024; Wang et al., 2025b). Other work targets efficiency and scale (Shen et al., 2025; Elflein et al., 2026; Wang & Agapito, 2025). Such coupling and subsequent geometric alignment (Wang et al., 2024; Duisterhof et al., 2024) improve consistency, but spatially varying offsets can remain between decoded surface predictions. We target this residual disagreement in the predicted coordinate frame.

2.2 Refining and Reconciling Predictions

Coordinate-only denoisers, including PointCleanNet (Rakotosaona et al., 2020), score-based estimators (Luo & Hu, 2021), diffusion bridges (Vogel et al., 2024), and iterative filters (de Silva Edirimuni et al., 2023), regress points toward a clean manifold. Without access to the images, cameras, or view provenance associated with each point, these methods do not explicitly distinguish perspective-dependent inter-view offsets from local surface noise. DepthRefine (Fink et al., 2024) instead initializes scale from an SfM cloud and optimizes a meshed depth map with photometric and geometric losses through differentiable rendering; per-scene NeRF (Mildenhall et al., 2020) and 3D Gaussian Splatting (Kerbl et al., 2023) likewise obtain cross-view supervision through rendering. In their published settings, the NeRF-family variants evaluated by mip-NeRF 360 take 3.2–22.8 hours per scene, and NeuS takes 14–16 hours (Barron et al., 2022; Wang et al., 2021). IVGT (Wu et al., 2026) avoids explicit point-map redundancy by training a new continuous implicit SDF representation, whereas we retain an existing point-map backbone and correct its decoded geometry directly.

Test-time geometric adaptation is directly related. Align3R (Lu et al., 2024) combines monocular depth with depth/pose optimization for dynamic video. LoRA3D (Lu et al., 2025) refines point maps and confidence to construct pseudo labels for low-rank model adaptation. Test3R (Yuan et al., 2025) optimizes visual prompts so that image pairs sharing a reference view predict consistent point maps, while TTT3R (Chen et al., 2025) studies recurrent-state adaptation for long sequences. More recently, TCO (Zhou et al., 2026a) adapts VGGT with rendering-based cross-view compatibility and optional camera or depth priors; Online3R (Zhou et al., 2026b) uses prompt adaptation for sequential reconstruction; and Free Geometry (Dai & Yang, 2026) distills predictions from longer view sequences into a short-view model. Our external residual network instead corrects decoded multi-view point maps after one frozen backbone pass and per-view alignment, without updating prompts or backbone adapters. Training-time reprojection and consistency losses (Godard et al., 2019; Zhao et al., 2020; Bian et al., 2019) also impose geometric constraints; our scene-level correction uses cached point maps and image features at test time.

3 Method

3.1 Overview and Preliminaries

A frozen feed-forward model (Pi3 (Wang et al., 2025c) or VGGT-Ω (Wang et al., 2026)) maps K uncalibrated images $\{I_k\}_{k=1}^{K}$ to per-view point maps and confidence scores. After filtering low-confidence and depth-discontinuity pixels, we form $\mathcal{P} = \{\mathbf{p}_i^{(0)}\}$ with view assignment $v(i)$. Refinement uses cross-view agreement directly in the model's predicted frame, without target-scene depth or 3D ground truth; a similarity to the metric frame is fitted only for evaluation.

Algorithm pseudocode is in the supplement.

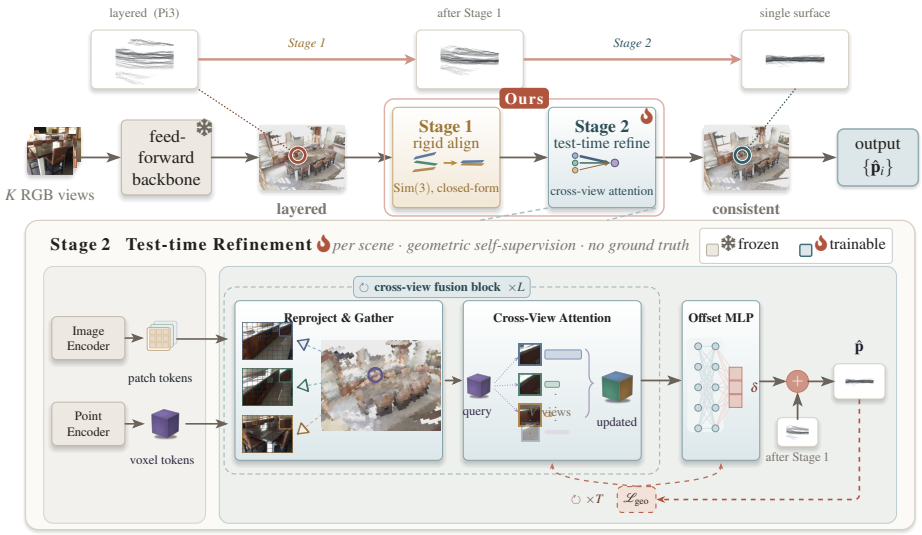

Figure 2: Overview of our method. From one base-model pass we build a shared voxel-token field with multi-view reprojection; Stage 1 aligns each view to the cross-view consensus in closed form; Stage 2 optimizes a residual network per scene by multi-view voxel attention under geometric self-supervision. All stages operate in the base model's predicted frame without target-scene depth or 3D reference geometry.

3.2 Scene Tokenization and Multi-View Reprojection

We compute per-view normals on each point-map grid, use predicted poses, and recover pinhole intrinsics from predicted camera-frame geometry. A frozen point-transformer tokenizer (Utonia (Zhang et al., 2026)) voxelizes $\mathcal{P}$ into M tokens $\{(\mathbf{x}_m, \mathbf{f}_m)\}$ and point assignments $m(i)$. We reproject each voxel into the predicted cameras and retain visible image patches and viewing directions. Each point keeps its source pixel and image feature. Tokens, assignments, cameras, and reprojection associations are cached before Stage 1 and reused in Stage 2; voxel sizes and visibility tests are detailed in the supplement.

3.3 Stage 1: Closed-Form Per-View Alignment

Stage 1 fits one similarity per view to correct rigid-plus-scale drift. At iteration $t = 0, \ldots, T-1$ (T=4), we voxelize the current cloud $\{\mathbf{p}_i^{(t)}\}$ on a coarse grid, with assignment $m_t(i)$ distinct from the cached Stage-2 assignment $m(i)$. For each voxel and view we compute a centroid $\boldsymbol{\mu}_{m,v}^{(t)}$ and the cross-view consensus

$$\mathbf{c}_m^{(t)} = \frac{1}{|V_m^{(t)}|} \sum_{v \in V_m^{(t)}} \boldsymbol{\mu}_{m,v}^{(t)}, \tag{1}$$

where $V_m^{(t)}$ is the set of views contributing at least three points in iteration t. For each view v we then solve a single 7-DOF similarity ($s > 0$, $\mathbf{R} \in \mathrm{SO}(3)$, $\mathbf{t} \in \mathbb{R}^3$) by Umeyama's closed form (Umeyama, 1991),

$$(s_v^{(t)}, \mathbf{R}_v^{(t)}, \mathbf{t}_v^{(t)}) = \arg\min_{s,\mathbf{R},\mathbf{t}} \sum_{\substack{i:\, v(i)=v \\ |V_{m_t(i)}^{(t)}| \geq 2}} \left\| s\mathbf{R}\mathbf{p}_i^{(t)} + \mathbf{t} - \mathbf{c}_{m_t(i)}^{(t)} \right\|^2, \tag{2}$$

using points in voxels seen by ≥ 2 views. A well-constrained fit gives $\mathbf{p}_i^{(t+1)} = s_v^{(t)} \mathbf{R}_v^{(t)} \mathbf{p}_i^{(t)} + \mathbf{t}_v^{(t)}$ for all points of view v, preserving their relative geometry; otherwise those points remain unchanged. Normals are rotated by $\mathbf{R}_v^{(t)}$ when the fit is applied. Gating rules are in the supplement.

Because the same object's predicted local shape can differ across views, a per-view similarity cannot remove all layering. Stage 2 couples local corrections through an image-conditioned residual field and displacement regularization, bringing the remaining sheets closer while limiting disruption of neighboring pixels (Figure 4).

3.4 Stage 2: Ground-Truth-Free Test-Time Refinement

Residual parameterization. Write $\mathbf{p}_i := \mathbf{p}_i^{(T)}$ for the aligned point. A multi-view voxel-attention network f_θ predicts a per-point residual, scaled by $s = 0.05$:

$$\hat{\mathbf{p}}_i = \mathbf{p}_i + s\,\delta_\theta(i), \qquad s = 0.05. \tag{3}$$

We warm-start θ from a ScanNet-pretrained checkpoint; random initialization is compared in Table 4. In *multi-view voxel attention*, each voxel token attends to image patches from its visible views, conditioned on voxel-to-camera directions. Each point is decoded from nearby voxel tokens, its image feature, and raw attributes. Only θ is optimized per scene; point maps, base model, and tokenizer remain fixed.

Self-supervised objective. We minimize

$$\mathcal{L} = \lambda_c \mathcal{L}_{\text{cons}} + \lambda_p \mathcal{L}_{\text{col}} + \lambda_s \mathcal{L}_{\text{smooth}}, \tag{4}$$

with $(\lambda_c, \lambda_p, \lambda_s) = (1, 1, 2)$. Let ρ_β denote the Huber penalty and $\langle\cdot\rangle_w$ a w-weighted mean.

(i) Between-view-gated centroid consensus. For each voxel we pull the refined per-view centroids $\bar{\boldsymbol{\mu}}_{m,v}$ toward their current consensus $\bar{\mathbf{c}}_m(\theta)$ along a normal $\mathbf{n}_m^{\text{cons}}$ estimated by PCA from the aligned baseline points:

$$\mathcal{L}_{\text{cons}} = \Big\langle \frac{\sum_v a_{m,v}\, \rho_\beta\big((\bar{\boldsymbol{\mu}}_{m,v} - \bar{\mathbf{c}}_m(\theta))\cdot \mathbf{n}_m^{\text{cons}}\big)}{\sum_v a_{m,v}} \Big\rangle_{w_m^{\text{cons}}}. \tag{5}$$

The consensus $\bar{\mathbf{c}}_m(\theta) = \sum_v b_{m,v}\bar{\boldsymbol{\mu}}_{m,v} / \sum_v b_{m,v}$ is recomputed from refined centroids, with fronto-parallel views weighted more. Gradients pass through centroids, so the loss penalizes relative disagreement without anchoring the common surface to Stage 1. A between-view spread gate, planar shape weight, and grazing-view weights restrict where it acts; definitions are in the supplement.

(ii) Soft per-view collapse. Let $\mathbf{n}_m^{\text{col}}$ aggregate the rotated per-view normals, $\mathbf{c}_m^0$ be the aligned baseline centroid, and $o_{m,v} = (\boldsymbol{\mu}_{m,v}^{\text{base}} - \mathbf{c}_m^0)\cdot \mathbf{n}_m^{\text{col}}$. Robust IRLS on these offsets gives a common target t_m. The loss pulls each point's along-normal *displacement* toward its view's required shift:

$$\mathcal{L}_{\text{col}} = \Big\langle \rho_\beta\big(\big(s_i - s_i^{\text{base}}\big) - \big(t_{m(i)} - o_{m(i),v(i)}\big)\big)\Big\rangle_{w_{m(i)}^{\text{col}}}, \tag{6}$$

Here $s_i = (\hat{\mathbf{p}}_i - \mathbf{c}_{m(i)}^0)\cdot \mathbf{n}_{m(i)}^{\text{col}}$ and s_i^{base} uses $\mathbf{p}_i$. A view shares one displacement target within each voxel, preserving its within-view offsets rather than forcing every point to one location. Collapse uses an offset and support gate distinct from the centroid gate; full definitions are in the supplement.

(iii) Displacement-field Laplacian. On each view's point-map grid we penalize the curvature of the *displacement* field rather than of the surface:

$$\mathcal{L}_{\text{smooth}} = \big\langle \rho_{\beta'}\big(\|\Delta\hat{\mathbf{p}}_i - \Delta\mathbf{p}_i\|\big)\big\rangle_{e_i}, \tag{7}$$

Here Δ denotes grid second differences, β' a narrower Huber width, and e_i attenuates spatial and color discontinuities. The loss penalizes displacement curvature, though locally affine tangential shifts remain possible. Gates and geometric references come from the frozen aligned baseline. We optimize for 300 steps; implementation details are in the supplement.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on six datasets spanning object-level, indoor, and outdoor scenes. DTU (Jensen et al., 2014) provides object scans; NRGBD (Azinović et al., 2022) synthetic indoor

depth and meshes; ScanNet++ (Yeshwanth et al., 2023) laser-scan indoor geometry; 7-Scenes (Shotton et al., 2013) and ScanNet (Dai et al., 2017) real RGB-D captures; and ETH3D (Schöps et al., 2017) high-precision indoor and outdoor reference geometry. We use DTU, ScanNet++, and ETH3D for the main Completeness comparison, NRGBD for the controlled analysis in Section 4.4, and 7-Scenes as a separate supplementary benchmark. ScanNet is qualitative only. The supplementary material details each reference.

Protocol. For each scene, we sample K frames, run the base model, and apply Align + Refine. Reference point maps are back-projected from the provided reference depths and cameras at the prediction grid. The DTU object-level protocol applies the dataset-provided foreground silhouettes to predicted points before voxelization and refinement; the mask files are external inputs, while target-scene reference depth values and camera poses are used only for evaluation. Each output is independently registered to its corresponding valid reference pixels by dense Umeyama alignment followed by rigid ICP; nearest-neighbor distances use all valid points. Results are aggregated per dataset (mean over scenes).

Metrics. We report point-to-point Accuracy (Acc, prediction-to-GT distance), Completeness (Comp, GT-to-prediction distance), and Normal Consistency (NC). Acc and Comp use mean and median distances (lower is better); NC measures orientation agreement (higher is better). Units are millimetres for DTU and metres otherwise. Evaluation normals are recomputed by PCA on each output and reference cloud using 30 nearest neighbors, independently of the refinement normals. At fixed registration, adding predictions cannot increase one-sided Completeness error; our evaluator re-registers each output independently. We therefore interpret the two directions together with NC and analyze the coverage effect in Section 4.4.

Implementation. We keep the base model and tokenizer frozen and optimize a 0.9 M-parameter residual network. The original benchmark runs use a single NVIDIA A800 GPU (80 GB). For a representative Pi3 indoor scene (K=15 views, 3.5M points), the base forward pass, tokenization, and reprojection take 3.4 s, and Stage 1 takes 0.1 s. Stage 2 takes 220 s over 300 Adam steps at learning rate 2×10^{-4}, followed by 0.2 s for full-density inference; peak memory is 8.8 GB, for about 3.7 minutes in total.

4.2 Point Map Estimation

Table 1 compares methods under shared keyframes, registration, and metrics. Applying the same refiner to Pi3 and VGGT-Ω improves mean Accuracy and Normal Consistency on DTU, ETH3D, and ScanNet++. Pi3 + Ours attains the lowest DTU mean Accuracy in the table (0.806 versus 1.217 mm for Pi3); VGGT-Ω + Ours attains the lowest ETH3D and ScanNet++ mean Accuracy (0.037 and 0.0168 m), with ETH3D Normal Consistency increasing from 0.940 to 0.962. On ETH3D, refinement lowers Accuracy error in 11 of 13 scenes for each base (supplement). These gains use one external refiner without per-base retraining. Figure 3 shows the full pipeline, and Section 4.3 isolates Stage 2 on identical retained pixels. Completeness and comparator setup details are in the supplement.

4.3 Cross-View Agreement and Local Surface Structure

Stage 2 must reduce the residual layering after alignment while preserving local surface structure. Figure 4 shows both requirements: Align retains separated sheets in the multiview slices, while Direct and Direct + smooth introduce blockwise breaks in the single-view surfaces. Ours brings the sheets closer while retaining a continuous wall corner and curved edge. We compare PCA projection, direct collapse shifts, and unrestricted per-point offsets from the same aligned points and retained pixels. The neural refiner additionally uses cached multi-view image features and a shared residual field with grid regularization.

As a separate position check, we freeze each scene's Align-only registration and compare every retained predicted point with the reference 3D point at the same input pixel. This paired error uses no nearest-neighbor search or per-output registration. Table 2 covers both bases on all three main benchmarks, with Ours attaining the lowest cohort-mean paired error in each row. The same test improves all nine Pi3/NRGBD scenes (supplement). Within the fixed Direct grid for Pi3, $\alpha = 0.75$

	DTU				ETH3D				ScanNet++			
	Acc↓		NC↑		Acc↓		NC↑		Acc↓		NC↑	
Method	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
CUT3R	4.343	2.378	0.683	0.769	0.615	0.512	0.768	0.871	0.1138	0.0741	0.720	0.850
Fast3R	2.798	1.697	0.678	0.764	0.578	0.446	0.760	0.847	0.0695	0.0485	0.815	0.922
FLARE	2.500	1.452	0.687	0.777	0.386	0.253	0.763	0.867	0.0500	0.0315	0.849	0.961
VGGT	1.488	0.825	0.681	0.771	0.261	0.167	0.862	0.956	0.0400	0.0257	0.824	0.952
MapAnything	2.160	1.185	0.692	0.783	0.400	0.310	0.826	0.913	0.0445	0.0298	0.836	0.968
DA3	1.311	0.795	0.695	0.790	0.241	0.197	0.852	0.913	0.0199	0.0129	**0.919**	**0.995**
AMB3R	1.107	0.661	**0.707**	**0.792**	0.244	0.152	0.867	0.952	0.0322	0.0205	0.867	0.982
Test3R	2.240	1.280	0.689	0.778	0.903	0.836	0.687	0.753	0.1569	0.1267	0.578	0.617
TCO	2.320	1.328	0.684	0.772	0.230	0.139	0.881	0.974	0.0365	0.0237	0.840	0.981
Free Geometry	1.528	0.857	0.686	0.779	0.304	0.209	0.857	0.952	0.0305	0.0194	0.852	0.979
Pi3 (base)	1.217	0.647	0.669	0.755	0.192	0.125	0.876	0.967	0.0349	0.0255	0.862	0.965
VGGT-Ω (base)	1.157	0.592	0.668	0.753	0.049	0.027	0.940	0.992	0.0278	0.0171	0.846	0.974
Pi3 + Ours	**0.806**	**0.464**	0.697	0.781	0.162	0.102	0.906	0.982	0.0306	0.0235	0.904	0.973
VGGT-Ω + Ours	0.868	0.482	0.683	0.766	**0.037**	**0.023**	**0.962**	**0.995**	**0.0168**	**0.0109**	0.889	0.982

Table 1: Point-map estimation on DTU, ETH3D, and ScanNet++: Accuracy and Normal Consistency. All methods share keyframes, dense Umeyama + ICP registration, and metrics. Test3R, TCO, and Free Geometry use their final outputs; TCO receives no target-scene GT depth, pose, or intrinsics. Test3R uses $K{=}10$ on DTU and $K{=}12$ on ScanNet++. The same refiner is applied to Pi3 and VGGT-Ω. Completeness and comparator setup details are in the supplement. **Best** per column in bold, second-best underlined.

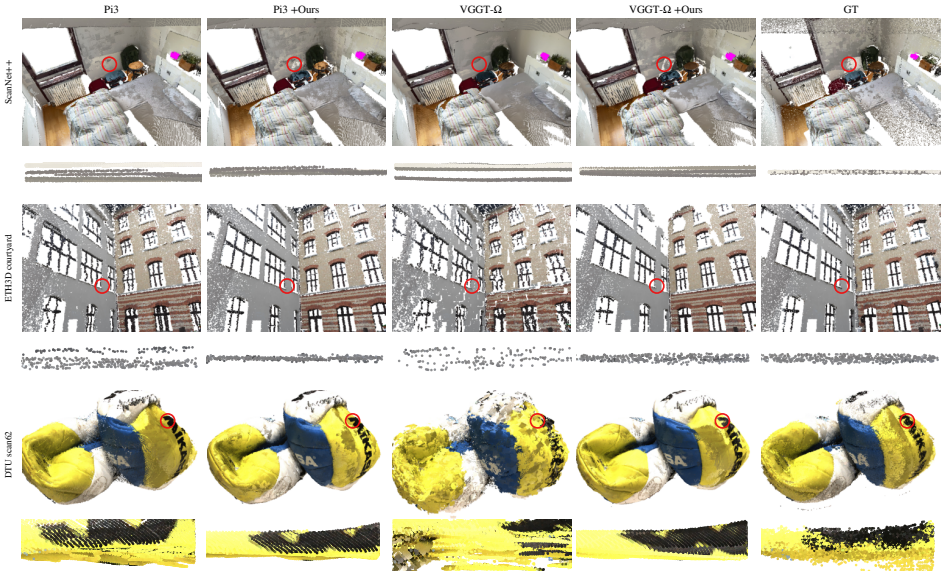

Figure 3: **Qualitative de-layering on ScanNet++, ETH3D, and DTU.** Each example shows a point-cloud rendering and a cross-section of the circled region. The complete pipeline brings offset predictions from multiple views closer to a common surface. These examples illustrate the visual effect; the incremental contribution of Stage 2 is evaluated separately against Align only.

gives the lowest paired errors, 21.022 and 6.121 mm on NRGBD and DTU, respectively. Figure 4 illustrates the gap–shape tradeoff in two regions from different scenes.

On fixed input voxel/view groups, Stage 2 reduces the cross-view centroid gap after Align in all 9 NRGBD/Pi3, 22 DTU/Pi3, and 22 DTU/VGGT-Ω scene–model cases, by 25%, 29%, and 18% on average, respectively. Gap reduction alone is insufficient: a full Direct shift yields smaller gaps but higher GT local-edge error in each cohort.

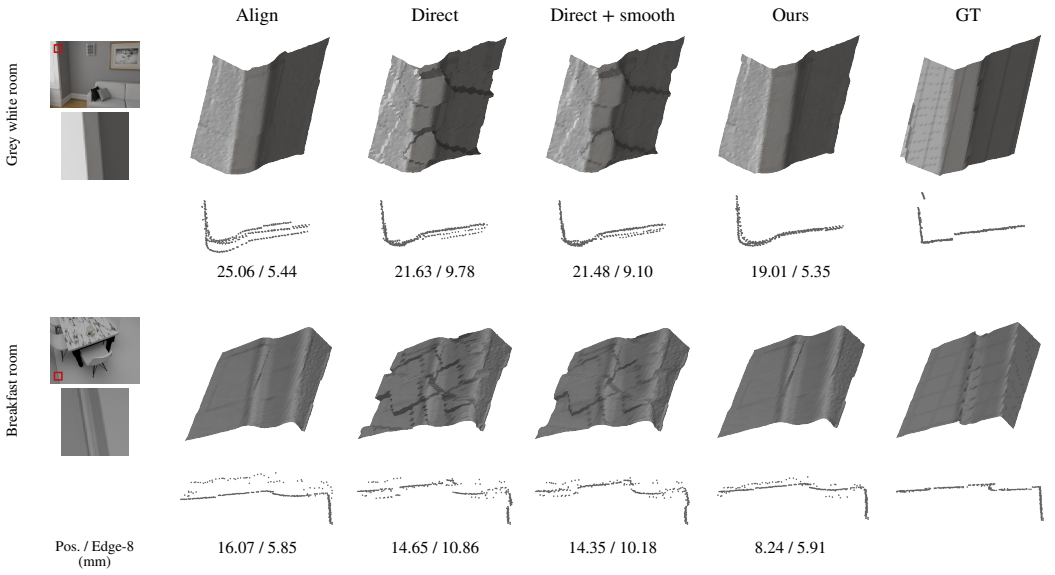

Figure 4: **Stage 2 reduces residual layering while retaining local surface structure.** In these two selected examples, Align retains separated sheets in the lower multiview slices. Direct and Direct + smooth introduce breaks in the upper single-view surfaces. Ours narrows the separation while retaining surface continuity. Pos./Edge-8 are upper-patch paired position / eight-pixel edge-vector errors (mm) under fixed Align registration. Direct uses $\alpha = 1$; point identities and display scale are shared within each row.

Base / dataset	Align	PCA	Direct	Direct+smooth	Free offsets	Ours
Pi3 / DTU	6.154	6.133	6.122	6.117	6.166	**6.069**
VGGT-Ω / DTU	7.154	7.133	7.095	7.090	7.133	**7.041**
Pi3 / ETH3D	498.512	498.349	498.298	498.258	498.330	**498.239**
VGGT-Ω / ETH3D	82.346	82.029	81.935	81.890	82.076	**81.299**
Pi3 / ScanNet++	153.711	153.628	153.629	153.602	153.654	**153.460**
VGGT-Ω / ScanNet++	57.695	57.612	57.565	57.530	57.698	**57.063**

Table 2: Position check for Stage 2 on the main benchmarks. Scene-mean corresponding-pixel error (mm, lower is better) uses the same retained pixels and one Align-only registration per scene; it differs from the nearest-neighbor Accuracy in Table 1. Complete DTU mean and median results are in the supplement.

Across all 6,658 valid 48×48 patches in six audited NRGBD scenes, Ours has lower one-pixel GT edge-vector error than full-strength Direct ($\alpha = 1$) in 94.7% of patches and its smoothed variant in 87.5%. Mean edge error remains close to, but above, Align (3.017 versus 2.993 mm). Figure 4 shows two examples selected for improvement over both Direct controls. Free offsets also has higher cohort-mean Edge-8 error than Ours. Weak Direct ($\alpha = 0.25$) has lower Edge-8 error in all three cohorts, while Ours has lower paired position error. The supplement gives the full strength sweep and selection protocol.

4.4 Layering Can Artificially Lower Completeness Error

Completeness measures the distance from each reference point to its nearest prediction. Extra off-surface sheets can shorten this distance even as prediction-to-GT Accuracy worsens (proof in the supplement). To separate layering from a larger point count, we split the *same* N points into K

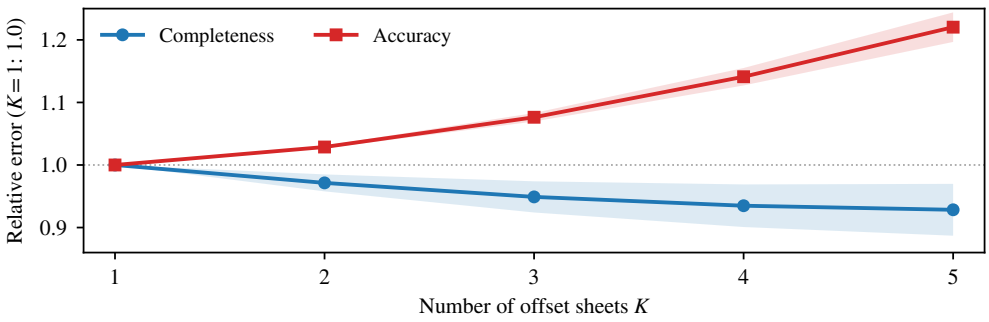

Figure 5: Fixed-count layering test on three NRGBD scenes. Curves show mean one-sided errors relative to $K=1$ with the point count held constant.

	7-Scenes (real)			NRGBD (synthetic)		
Method	Acc↓	Comp↓	NC↑	Acc↓	Comp↓	NC↑
Base model (raw VGGT-Ω)	15.47	16.49	0.683	10.58	5.89	0.904
+ Align only	9.06	15.38	0.706	6.55	5.29	0.948
+ Refine only	9.27	14.86	0.704	6.29	5.13	0.954
+ Align + Refine (Ours)	8.78	16.06	0.707	6.29	5.23	0.956
w/o consensus	8.91	15.74	0.706	6.32	5.21	0.955
w/o collapse	8.69	16.07	0.708	6.26	5.18	0.955
w/o Laplacian	8.76	16.27	0.703	6.36	5.32	0.955

Table 3: Ablation on VGGT-Ω for 7-Scenes and NRGBD. Acc./Comp. are scene-mean distances in millimetres; NC is unitless. Each output is independently registered and scored at full density. Small differences among loss variants are descriptive.

Metric	Pretrained	Random (no GT)
Steps	300	1500
Acc (mm)↓	6.3	6.1
Comp (mm)↓	5.2	5.5
NC↑	0.956	0.966

Table 4: Initialization on NRGBD (9 scenes): random weights reach comparable final quality after 1500 versus 300 steps. Unequal budgets do not establish equal-step performance.

offset sheets in three NRGBD scenes. Figure 5 shows that Completeness error falls while Accuracy error rises as K increases. The two one-sided metrics therefore describe different effects of layering.

Consolidation can raise Completeness error when the consensus surface lies farther from the reference. VGGT-Ω improves both distance directions on ETH3D and ScanNet++; on DTU, Accuracy and Normal Consistency improve while Completeness error rises (supplement). The corresponding-pixel test measures Stage 2's position correction directly.

4.5 Ablation Study

With fixed frames and seed, Align + Refine has the best stage-variant 7-Scenes Accuracy (8.78 vs. 9.06 mm Align, 9.27 mm Refine) and NRGBD Normal Consistency (0.956; Tables 3 and 4). Stage 1 removes similarity drift while preserving each view's shape; Stage 2 corrects residual non-rigid differences. Random initialization reaches comparable quality in 1,500 steps versus 300 with warm start.

On three prespecified Pi3/NRGBD scenes, optimization lowers fixed-pixel error from 36.816 mm (step 0) to 21.287 mm (step 300), versus 22.178 mm for Align. Across nine NRGBD scenes, removing image patches or collapse raises paired error from 20.500 to 21.163 or 21.779 mm, respectively (supplement).

5 Conclusion

Stage 1 corrects rigid-plus-scale drift; Stage 2 corrects the remaining non-rigid multi-view geometric differences. Across two backbones, this improves Accuracy and Normal Consistency without target-scene ground truth. Visual and cohort controls motivate evaluating local surface structure alongside point distances.

AI USE STATEMENT

We used OpenAI Codex to assist with manuscript editing, code inspection, and supplementary experiments. The authors are responsible for the experimental protocols, reported results, citations, and final claims.

ETHICS STATEMENT

This work studies geometric refinement for 3D reconstruction using established research benchmarks and does not collect new human-subject data. Because 3D reconstruction can be applied to privacy-sensitive scenes, downstream use should respect consent, privacy, dataset licenses, and applicable laws.

REPRODUCIBILITY STATEMENT

The main paper specifies the method, evaluation protocol, and runtime. The supplement provides the full algorithm, hyperparameters, software versions, per-scene results, and qualitative comparisons to support reproduction.

REFERENCES

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 (CVPR)*, 2022.

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025.

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 (CVPR)*, 2017.

Yuhang Dai and Xingyi Yang. Free Geometry: Refining 3D reconstruction from longer versions of itself. *arXiv preprint arXiv:2604.14048*, 2026. URL `https://arxiv.org/abs/2604.14048`.

Dasith de Silva Edirimuni, Xuequan Lu, Zhiwen Shao, Gang Li, Antonio Robles-Kelly, and Ying He. IterativePFN: True iterative point cloud filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. MASt3R-SfM: a fully-integrated solution for unconstrained structure-from-motion. *arXiv preprint arXiv:2409.19152*, 2024.

Sven Elflein, Ruilong Li, Sérgio Agostinho, Zan Gojcic, Laura Leal-Taixé, Qunjie Zhou, and Aljosa Osep. VGG-T^3: Offline feed-forward 3D reconstruction at scale. *arXiv preprint arXiv:2602.23361*, 2026.

Laura Fink, Linus Franke, Bernhard Egger, Joachim Keinert, and Marc Stamminger. Refinement of monocular depth maps via multi-view differentiable rendering. *arXiv preprint arXiv:2410.03861*, 2024.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9777–9786, 2020.

W. Hu, Z. Zhang, B. Li, et al. Ocean: Object-aware anchor-free tracking with matching-relation learning. *Machine Intelligence Research*, 23:565–592, 2026. doi: 10.1007/s11633-026-1634-0.

Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4), 2023.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision (ECCV)*, 2024.

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.

Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Taku Komura, Wenping Wang, Lingjie Liu, and Xiaoxiao Wang. Align3R: Aligned monocular depth estimation for dynamic videos. *arXiv preprint arXiv:2412.03079*, 2024.

Ziqi Lu, Heng Yang, Danfei Xu, Boyi Li, Boris Ivanovic, Marco Pavone, and Yue Wang. LoRA3D: Low-rank self-calibration of 3d geometric foundation models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6db7c49b14da8006892fda7350d76b6a-Abstract-Conference.html.

Shitong Luo and Wei Hu. Score-based point cloud denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

Marie-Julie Rakotosaona, Vittorio La Barbera, Paul Guerrero, Niloy J. Mitra, and Maks Ovsjanikov. PointCleanNet: Learning to denoise and remove outliers from dense point clouds. *Computer Graphics Forum*, 2020.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. FastVGGT: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025.

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 (CVPR)*, 2013.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

Mathias Vogel, Keisuke Tateno, Marc Pollefeys, Federico Tombari, Marie-Julie Rakotosaona, and Francis Engelmann. P2P-Bridge: Diffusion bridges for 3D point cloud denoising. In *European Conference on Computer Vision (ECCV)*, 2024.

Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024.

Hengyi Wang and Lourdes Agapito. AMB3R: Accurate feed-forward metric-scale 3D reconstruction with backend. *arXiv preprint arXiv:2511.20343*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-ω. *arXiv preprint arXiv:2605.15195*, 2026.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025b.

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 (CVPR)*, pp. 20697–20709, 2024.

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*, 2025c.

Yuqi Wu, Tianyu Hu, Wenzhao Zheng, Yuanhui Huang, Haowen Sun, Jie Zhou, and Jiwen Lu. IVGT: Implicit visual geometry transformer for neural scene representation. *arXiv preprint arXiv:2605.16258*, 2026.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2018.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yuheng Yuan, Qiuhong Shen, Shizun Wang, Xingyi Yang, and Xinchao Wang. Test3R: Learning to reconstruct 3d at test time. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/911dd89c81efc624c4e1c39381179505-Abstract-Conference.html.

Yujia Zhang, Xiaoyang Wu, Yunhan Yang, Xianzhe Fan, Han Li, Yuechen Zhang, Zehao Huang, Naiyan Wang, and Hengshuang Zhao. Utonia: Toward one encoder for all point clouds. In *International Conference on Machine Learning (ICML)*, 2026.

Wang Zhao, Shaohui Liu, Chang Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without PoseNet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Lei Zhou, Haoyu Wu, Akshat Dave, and Dimitris Samaras. Learning 3D reconstruction with priors in test time. *arXiv preprint arXiv:2604.03878*, 2026a. URL https://arxiv.org/abs/2604.03878. Accepted to CVPR 2026.

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*, 2026b. URL https://arxiv.org/abs/2604.09480.

LIGHTWEIGHT TEST-TIME GEOMETRY UNIFICATION FOR MULTI-VIEW RECONSTRUCTION

SUPPLEMENTARY MATERIAL

Anonymous authors
Paper under double-blind review

This supplement gives the full algorithm, implementation settings, Stage-2 controls, and additional benchmark results. The Stage-2 controls follow Figure 4 of the main paper: they measure position correction and local shape change after the same alignment. Unless stated otherwise, benchmark results use shared keyframes, dense Umeyama + ICP registration, and the same scoring code. Normal Consistency uses PCA normals from 30 nearest neighbors and averages the absolute cosine of matched normals in both directions.

Comparison with recent methods. We run Test3R, TCO, and Free Geometry with their released models and preprocessing, then score their final reconstructions with the same code. We found no publicly released reconstruction code for LoRA3D or Online3R, so we discuss them in Related Work without numerical comparisons.

1 IMPLEMENTATION AND EVALUATION PROTOCOL

Algorithm 1 summarizes the two-stage, per-scene pipeline (notation as in the main paper's Method section).

Algorithm 1 Self-supervised test-time de-layering by cross-view reconciliation.

Require: Images $\{I_k\}$; optional dataset foreground masks (DTU); frozen base model and tokenizer.

1: **Build:** run the base model once; retain point-map pixels passing confidence and depth-edge filters and, on DTU, the foreground mask; form cloud $\mathcal{P}$, per-view normals, and cameras. Tokenize $\mathcal{P}$ and reproject each voxel into every camera, keeping views where its depth test finds it unoccluded.
2: **Stage 1 (Align):** for T iterations, fit a per-view similarity to the cross-view voxel consensus and apply it to the dense points; rotate their normals. Keep the initial token field, assignments, cameras, and reprojection cache.
3: **Stage 2 (Refine):** initialize f_θ from the warm-start checkpoint.
4: **for** $t = 1$ to N_{steps} **do**
5: $\quad \hat{\mathbf{p}}_i \leftarrow \mathbf{p}_i + s\,\delta_\theta(i)$
6: $\quad$ compute baseline-derived loss statistics on the sampled points, the current refined-centroid consensus, and the displacement Laplacian on frame windows; update θ on $\lambda_c\mathcal{L}_{\text{cons}} + \lambda_p\mathcal{L}_{\text{col}} + \lambda_s\mathcal{L}_{\text{smooth}}$ (no ground truth).
7: **end for**
8: **return** refined points $\{\hat{\mathbf{p}}_i\}$ for every retained pixel.

1.1 IMPLEMENTATION DETAILS

The settings below extend the main-paper method.

Evaluation references. DTU uses scan-derived depth maps, NRGBD clean synthetic depths, ScanNet++ laser-mesh-rendered depths, and ETH3D laser-derived depths. The 7-Scenes and ScanNet references back-project Kinect and RGB-D depths, respectively, and therefore can contain accumulated surface thickness; 7-Scenes is evaluated separately and ScanNet is qualitative only.

Refiner architecture. The 0.89 M-parameter residual network uses token dimension 128 and two fusion blocks (`3d_static2d`). Each voxel attends to the fixed 2048-dimensional image-patch tokens of its observing views. Points are decoded in chunks from nearby voxel tokens, their patch feature, and raw attributes. The output is $\hat{\mathbf{p}}_i = \mathbf{p}_i + 0.05\,\delta_\theta(i)$.

Stage-1 alignment. The closed-form per-view Umeyama alignment iterates $T = 4$ times on the coarser 0.03 grid of the main paper (predicted units, not metres); the cross-view consensus $\mathbf{c}_m$ averages per-view centroids over views contributing ≥ 3 points, and each view's 7-DOF similarity is fitted on points in voxels seen by ≥ 2 views, then applied to all of that view's points. A view is left unchanged when the fit is under-constrained or degenerate: fewer than 50 of its points fall in ≥ 2-view voxels, those points span less than 0.3 of the view's extent, or the fitted scale falls outside $[0.5, 2]$. The frozen Utonia tokenizer (a pretrained Point Transformer V3) voxelizes on a 0.01 grid after rescaling the input: the object preset uses scale 1.0 and the indoor preset 0.5. The reported setup uses the resulting 0.01 grid on DTU and 0.02 on ETH3D, NRGBD, ScanNet++, and 7-Scenes, in predicted units rather than metres. These fixed scales are implementation presets, not a data-derived similarity normalization; arbitrary rescaling of an input can therefore change the voxel groups and active thresholds. Tokenization and reprojection precede alignment and are not rebuilt afterward. The camera centers used for Stage-2 viewing weights also remain the initial predictions.

Losses, normals, and reductions. The objective uses $(\lambda_c, \lambda_p, \lambda_s) = (1, 1, 2)$. Consensus and collapse use a Huber penalty of width $\beta = 0.01$, and smoothness uses $\beta' = 0.003$, in predicted units. Consensus takes a weighted mean over active voxels after averaging the losses of the participating views within each voxel. Collapse takes a weighted mean over active points, with a shared weight within each voxel. Thus these terms have different reductions. Consensus normals $\mathbf{n}_m^{\mathrm{cons}}$ are the minimum-variance axes of the aligned baseline voxel covariance. Collapse normals $\mathbf{n}_m^{\mathrm{col}}$ are the dominant axes of $\sum_{i:m(i)=m} \mathbf{n}_i \mathbf{n}_i^\top$, where $\mathbf{n}_i$ are the per-view finite-difference normals rotated by Stage 1. The latter aggregation is invariant to normal sign, but neither estimator is guaranteed accurate near depth boundaries or on thin structures.

Consensus gate and moving center. A voxel needs at least 8 baseline points and at least 2 views contributing 3 points each. Let σ_m be the square root of the smallest pooled covariance eigenvalue and σ_m^{btw} the standard deviation of the qualifying per-view centroids projected on $\mathbf{n}_m^{\mathrm{cons}}$, weighting those views equally. The implementation uses the within-view proxy

$$\tilde{\sigma}_m^{\mathrm{wth}} = \sqrt{\max\{0, \sigma_m^2 - (\sigma_m^{\mathrm{btw}})^2\}},$$

and requires $\sigma_m^{\mathrm{btw}}/(\tilde{\sigma}_m^{\mathrm{wth}} + 10^{-4}) > 1$ and $\sigma_m^{\mathrm{btw}} > 5 \times 10^{-4}$. Since pooled covariance weights points whereas between-view spread weights views, this subtraction is a heuristic proxy, not an exact variance decomposition when view counts differ. With sorted covariance eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$, define linearity $L = (\lambda_1 - \lambda_2)/\lambda_1$, planarity $P = (\lambda_2 - \lambda_3)/\lambda_1$, and scattering $S = \lambda_3/\lambda_1$. The soft shape weight is

$$g_m = \mathrm{sigmoid}((P - 0.35)/0.08)\ \mathrm{sigmoid}((0.65 - L)/0.08)\ \mathrm{sigmoid}((0.20 - S)/0.08).$$

Voxels with $g_m \leq 10^{-3}$ are excluded. With $r_m = \max(0, \sigma_m^{\mathrm{btw}} - 5 \times 10^{-4})$ and $\bar{r}$ its mean over active voxels, their weights are $w_m^{\mathrm{cons}} = \min(1 + 4r_m/\bar{r}, 17) g_m$, with numerical denominator clamps in the implementation. These are soft preferences rather than hard planarity guarantees. Viewing weights are $b_{m,v} = |\mathbf{r}_{m,v} \cdot \mathbf{n}_m^{\mathrm{cons}}|^2$ and $a_{m,v} = 0.1 + (1 - b_{m,v})$, where $\mathbf{r}_{m,v}$ points from the initial camera center to the current refined centroid and is detached from differentiation. The consensus $\bar{\mathbf{c}}_m(\theta) = \sum_v b_{m,v} \bar{\boldsymbol{\mu}}_{m,v} / \sum_v b_{m,v}$ is recomputed at every loss evaluation, retaining gradients through the centroids.

Collapse target and gate. Let $\mathbf{c}_m^0$ be the mean of the aligned baseline points and $o_{m,v} = (\boldsymbol{\mu}_{m,v}^{\mathrm{base}} - \mathbf{c}_m^0) \cdot \mathbf{n}_m^{\mathrm{col}}$. Starting at the mean of the qualifying view offsets, 3 IRLS iterations use weights $c/(c + |o_{m,v} - t_m|)$ with $c = 0.01$ to obtain t_m. The target displacement for each point of view v is $t_m - o_{m,v}$. The collapse risk is $q_m = \max_v |o_{m,v} - t_m|$ over qualifying views. Active voxels satisfy the same count thresholds as above and $q_m > 5 \times 10^{-4}$; their point weights are $w_m^{\mathrm{col}} = \min(1 + 4q_m/\bar{q}, 17)$. This term does not use the SNR or shape gate of the consensus term. Baseline-derived statistics are recomputed on the sampled baseline subset; the aggregated collapse normals and the

token assignments remain fixed. No gate is updated from refined geometry. A common along-normal offset can satisfy both objectives when their normal estimates agree because the centroid target moves with the refined centroids; their normal estimates can differ in general.

Displacement smoothness. Horizontal and vertical grid second differences are penalized separately. For each valid three-pixel stencil, the bilateral weight multiplies $\exp[-(\|\mathbf{p}_j - \mathbf{p}_i\|/0.03)^2]$ over its two neighbors j and the analogous RGB factors with scale 0.25; the loss is normalized by the sum of stencil weights. It attenuates across spatial and color discontinuities. The regularizer is zero for affine displacement on a local image grid and therefore does not bound tangential translation or affine distortion. Preserving the baseline Laplacian is an inductive bias, not an exact guarantee that surface detail or area is unchanged.

Optimization and keyframes. Per scene we run 300 Adam steps (β_1=0.9, β_2=0.999) at a learning rate of 2×10^{-4}, warm-started from a checkpoint pretrained with ScanNet RGB-D depth and pose-derived 3D targets. The pretraining scene is disjoint from every quantitative evaluation scene, and no depth values or 3D ground truth from a target evaluation scene enter test-time optimization; from a random initialization about 1500 steps are used to reach comparable final Accuracy (the main-paper initialization table). Keyframes follow the standard protocol: DTU/ETH3D every 5th frame (K=10 on DTU), ScanNet++ 12 frames linspace-sampled per scene, and NRGBD a dense sampling (4–17 frames/scene). The geometric losses sample up to 2 million points from all views per step; the Laplacian uses windows of up to 12 full point maps. The base model and tokenizer are frozen throughout.

DTU foreground inputs. The historical DTU object-level protocol supplies binary foreground silhouettes. We resize each silhouette to the reference depth-map dimensions (without reading depth values), apply the benchmark's 10×10 erosion, then resize it to the prediction grid. This mask selects predicted points before voxelization, Stage 1, and Stage 2. Reference depth values, reference camera poses, point maps, and surface normals enter evaluation only; they do not supervise the optimization.

Software and hardware. The benchmark uses one A800-SXM4-80 GB GPU, two Xeon Platinum 8350C CPUs, and 1 TB RAM under Ubuntu 22.04.3. Software: Python 3.10.19, PyTorch 2.5.1/CUDA 12.1, cuDNN 9.1, NumPy 2.2, SciPy 1.15, and Open3D 0.19. Pi3, VGGT-Ω, and Utonia use public weights; rescoring uses the original Open3D evaluator.

2 Stage-2 Controls: Position and Local Shape

Figure 4 in the main paper shows the coupled requirements of Stage 2: reducing the separated sheets that remain after alignment and preserving the local surface traced by neighboring pixels. Direct voxel shifts bring views closer but leave cracks where neighboring voxels receive different shifts; smoothing does not remove the breaks in those examples. Our network narrows the residual separation with a more continuous deformation. The controls below assess this behavior through cross-view agreement, local shape, and point position, which together describe the geometric correction.

2.1 Geometric alternatives

We compare four updates to the same Stage-1 points. PCA projects points onto local planes. Direct applies the voxel collapse shift, with a separate smoothed version where available. Free offsets optimize an independent displacement for every point under the same geometric losses, including the displacement smoothness term. Ours uses a pretrained residual network and 300 scene-specific steps. All variants retain the same predicted pixels; target-scene ground truth is used only for scoring.

Table 1 separates filtering, alignment, and refinement on all nine NRGBD scenes. Filtering accounts for 90.8% of the raw-to-full CD decrease ($13.643 \rightarrow 8.636 \rightarrow 8.127$ mm), while Stage 2 changes Align's 8.180 to 8.127 mm and improves seven scenes. The full pipeline's gain must therefore be distinguished from Stage 2's increment.

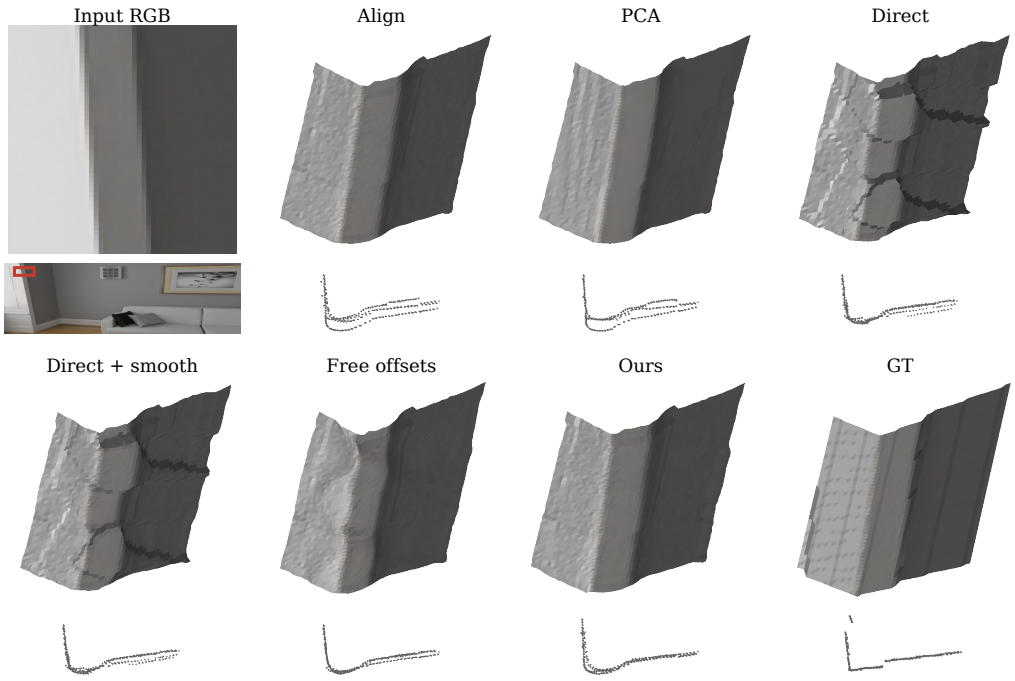

Figure 1: Post-processing controls on the wall corner selected for Figure 4 of the main paper. The Input RGB panel shows the image crop and its location in the full frame. The other upper panels use the same view, pixels, RGB, mesh connectivity, camera, and Align-derived registration; their lower panels show the same 470-point, four-view section with shared axes and scale. Free offsets leaves a visibly uneven wall despite improving Accuracy on average. This selected example illustrates local shape; the nine-scene scores are in Table 1.

Method	Acc. ↓	Comp. ↓	CD ↓	NC ↑
Raw base	14.091	13.195	13.643	0.8909
Filtered base	9.907	7.365	8.636	0.9055
Align only	8.963	**7.397**	8.180	0.9308
PCA	8.646	8.280	8.463	0.9401
Direct ($\alpha = 0.25$)	8.787	7.564	8.175	0.9383
Direct ($\alpha = 1$)	8.570	8.092	8.331	0.9462
Direct + smooth ($\alpha = 1$)	8.538	8.173	8.355	0.9480
Free offsets	**8.475**	8.316	8.395	**0.9502**
Ours	8.611	7.643	**8.127**	0.9439

Table 1: NRGBD controls on all nine original scenes (114 keyframes), scored with the original per-output Sim(3)+ICP and full-point evaluation. Rows below the divider share Stage-1 points and retained pixels; raw and filtered rows isolate preprocessing. Distances are mm. Bold marks the best post-alignment control.

Free offsets has the lowest mean Accuracy and highest NC, but worse Completeness and CD than Ours. Its selected wall-corner example has visible bumps and Edge-8 error 8.28 versus 5.35 mm for Ours. Differences in parameterization and warm-start prevent this example from isolating an architectural cause.

Scene	Align	PCA	Direct $\alpha = .25$	Direct $\alpha = 1$	Direct + smooth	Free offsets	Ours
thin geometry	4.177	4.236	**4.168**	4.182	4.183	4.207	4.172
breakfast room	7.598	7.926	7.577	7.665	7.684	7.720	**7.521**
morning apartment	**9.765**	10.020	9.778	9.958	10.020	10.164	9.838
complete kitchen	9.987	10.305	9.985	10.126	10.154	10.183	**9.893**
green room	**7.863**	8.246	7.904	8.126	8.155	8.191	7.900
grey white room	7.931	8.258	7.941	8.151	8.180	8.213	**7.831**
kitchen	8.916	9.172	8.897	9.041	9.052	9.052	**8.812**
staircase	11.131	11.592	11.144	11.379	11.403	11.432	**11.015**
whiteroom	6.253	6.416	6.184	6.351	6.367	6.397	**6.162**
Nine-scene mean	8.180	8.463	8.175	8.331	8.355	8.395	**8.127**

Table 2: Per-scene NRGBD symmetric distance (mm, lower is better), with each output independently registered. All controls share retained pixels and keyframes; Direct $\alpha = .25$ is the best cohort-mean fixed strength. Bold marks the lowest CD in each row.

2.2 Complete DTU controls

Formal DTU controls. Table 3 evaluates both base models on all 22 DTU scenes with the original full-point scoring protocol. For Pi3, compared with alignment alone, our Stage 2 improves Accuracy in 19/22 scenes and Normal Consistency in 22/22 scenes, while worsening Completeness and symmetric distance in all 22 scenes. For VGGT-Ω, Ours improves Accuracy in 20/22 scenes and NC in 22/22, but worsens Completeness in all 22 and CD in 19/22 (mean CD 1.505 versus 1.477 mm). Both bases use the same DTU scenes, keyframes, and evaluator, with each base model's native preprocessing.

	Pi3				VGGT-Ω			
Method	Acc	Comp	CD	NC	Acc	Comp	CD	NC
Raw base	1.217	1.729	1.473	0.669	1.157	**1.785**	**1.471**	0.668
Filtered base	1.063	**1.598**	1.331	0.673	1.040	1.969	1.504	0.672
Align only	0.816	1.773	**1.295**	0.685	0.883	2.071	1.477	0.677
PCA	0.802	1.957	1.379	0.693	0.868	2.234	1.551	0.686
Direct	0.795	1.975	1.385	0.703	0.844	2.374	1.609	0.702
Direct + smooth	0.793	2.006	1.400	0.706	–	–	–	–
Free offsets	**0.788**	2.055	1.421	**0.710**	**0.838**	2.462	1.650	**0.708**
Ours	0.806	1.885	1.346	0.697	0.868	2.143	1.505	0.683

Table 3: DTU controls on the same 22 scenes for both bases. Distances are mm (↓); NC is unitless (↑). Each output uses independent Sim(3)+ICP and full-point scoring. All post-alignment controls share retained pixels; Direct uses $\alpha = 1$. Bold marks best values within each base; – denotes an unreported score.

2.3 Corresponding-pixel position and direct shifts

Fixed-registration corresponding-pixel error. Table 4 uses the Align-only Sim(3)+ICP transform once per scene for all methods, matching each retained pixel to its reference 3D point without nearest-neighbor search. Ours lowers mean error in 9/9 NRGBD/Pi3, 21/22 DTU/Pi3, and 19/22 DTU/VGGT-Ω scenes. This tests point position on shared support, not coverage.

Direct-shift strength sweep. Table 5 reports independently registered full-point metrics for the shared grid $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. The fixed-registration position minimum among these Direct strengths is at $\alpha = 0.75$ for Pi3/NRGBD and Pi3/DTU (21.022 and 6.121 mm), versus the full model at 20.500 and 6.069 mm. DTU coverage and CD remain worse than Align for the full model.

Method	NRGBD / Pi3 (9)			DTU / Pi3 (22)			DTU / VGGT-Ω (22)		
	Mean	Med.	Wins	Mean	Med.	Wins	Mean	Med.	Wins
Align only	21.272	15.961	–	6.154	3.256	–	7.154	3.964	–
PCA	21.078	15.788	9/9	6.133	3.235	22/22	7.133	3.943	22/22
Direct	21.029	15.756	9/9	6.122	3.223	22/22	7.095	3.907	22/22
Direct + smooth	20.994	15.722	9/9	6.117	3.218	22/22	7.090	3.903	22/22
Free offsets	21.053	15.788	9/9	6.166	3.265	9/22	7.133	3.943	13/22
Ours	**20.500**	**15.170**	9/9	**6.069**	**3.157**	21/22	**7.041**	**3.860**	19/22

Table 4: Corresponding-pixel 3D error (mm, ↓) under one fixed Align-only registration and identical pixels. Mean and median are computed per scene, then averaged equally. Wins count scenes with lower mean error than Align. Direct uses $\alpha = 1$; Ours and Free offsets use 300 steps. Bold marks cohort minima.

Method	Acc ↓	Comp ↓	CD ↓	NC ↑
NRGBD / Pi3 (9 scenes)				
Direct $\alpha = 0$ (Align)	8.963	7.397	8.180	0.9308
Direct $\alpha = 0.25$	8.787	7.564	8.175	0.9383
Direct $\alpha = 0.5$	8.660	7.763	8.212	0.9436
Direct $\alpha = 0.75$	8.586	7.972	8.279	0.9464
Direct $\alpha = 1$	8.570	8.092	8.331	0.9462
Ours (300 steps)	8.611	7.643	**8.127**	0.9439
DTU / Pi3 (22 scenes)				
Direct $\alpha = 0$ (Align)	0.816	1.773	**1.295**	0.6849
Direct $\alpha = 0.25$	0.806	1.835	1.320	0.6921
Direct $\alpha = 0.5$	0.800	1.897	1.348	0.6988
Direct $\alpha = 0.75$	0.796	1.952	1.374	0.7029
Direct $\alpha = 1$	0.795	1.975	1.385	0.7029
Ours (300 steps)	0.806	1.885	1.346	0.6968

Table 5: Fixed-strength Direct-shift interpolation and Ours on the same cohorts. Distances are in mm; CD is the mean of Accuracy and Completeness, with the lowest CD in each cohort bold. Intermediate strengths use identical ordered Stage-1 points and are scored with the original full-point, per-output Sim(3)+ICP protocol. The Direct sweep is an internal geometric control.

2.4 Cross-view separation and local shape on complete cohorts

We evaluate the same aligned points and retained pixel identities used in the corresponding-pixel control. For each scene, voxel assignments, view memberships, normal axes, and the supported-voxel mask are fixed from Align only and reused for every output. A supported voxel contains at least eight points and at least two views with three or more points each. Within each voxel, we average the absolute along-normal separation over all distinct unordered pairs of view centroids; supported voxels and then scenes receive equal weight. This is a prediction-only measure of local cross-view compactness, not a verified correspondence or a measure of surface correctness. We report the result in millimetres using each scene's fixed Align-only similarity scale.

To test whether a smaller gap accompanies a more accurate surface, we also measure corresponding-pixel 3D position error and local 3D edge-vector error against GT. Both use the same Align-only Sim(3)+ICP registration for all outputs in a scene. Edge pairs lie eight pixels apart horizontally or vertically within one view; a fixed GT-continuity mask excludes pairs whose reference edge length exceeds five times the scene median. The retained pixels, edge pairs, and GT validity are shared across methods. GT selects evaluation pairs and supplies error targets only; it does not enter refinement or the cross-view grouping. Free offsets and Ours use 300 test-time steps. Direct interpolates the saved full collapse displacement with one grid $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ shared across scenes, without per-scene GT selection.

Ours reduces the gap relative to Align in all 53 scene–model cases, by 25%, 29%, and 18% in the three cohort means. Full Direct ($\alpha = 1$) and Free offsets compress the groups more strongly, but have higher mean position and edge errors than Ours. Weak Direct ($\alpha = 0.25$) has lower Edge-8

Base / dataset	Update after Align	Gap	Pos.	Edge-8
Pi3 / NRGBD (9)	Align ($\alpha = 0$)	5.792	21.272	6.800
	Direct ($\alpha = 0.25$)	4.388	21.146	**6.785**
	Direct ($\alpha = 0.5$)	2.983	21.062	6.861
	Direct ($\alpha = 0.75$)	1.579	21.022	7.004
	Direct ($\alpha = 1$)	**0.175**	21.029	7.198
	Free offsets	0.983	21.053	7.258
	Ours	4.339	**20.500**	6.903
Pi3 / DTU (22)	Align ($\alpha = 0$)	0.836	6.154	2.316
	Direct ($\alpha = 0.25$)	0.634	6.137	**2.313**
	Direct ($\alpha = 0.5$)	0.433	6.126	2.320
	Direct ($\alpha = 0.75$)	0.232	6.121	2.335
	Direct ($\alpha = 1$)	**0.030**	6.122	2.357
	Free offsets	0.159	6.166	2.399
	Ours	0.592	**6.069**	2.318
VGGT-Ω / DTU (22)	Align ($\alpha = 0$)	1.105	7.154	**2.250**
	Direct ($\alpha = 0.25$)	0.840	7.125	**2.250**
	Direct ($\alpha = 0.5$)	0.575	7.105	2.262
	Direct ($\alpha = 0.75$)	0.311	7.095	2.282
	Direct ($\alpha = 1$)	**0.046**	7.095	2.310
	Free offsets	0.159	7.133	2.357
	Ours	0.903	**7.041**	2.275

Table 6: Complete-cohort Direct-strength and free-offset controls (mm; lower is better). All methods share Align-derived voxel groups, registration, pixels, and GT edge pairs. Gap is along-normal view-centroid separation; Pos. is paired 3D error; Edge-8 uses eight-pixel GT edges. Entries are equal-scene means; bold marks minima at shown precision.

error than Ours in all three cohorts, while Ours has lower position error. On NRGBD, their gaps nearly match (4.388 versus 4.339 mm). Ours' Edge-8 error also rises slightly relative to Align in each cohort. These results establish a position–shape tradeoff, rather than uniform superiority over every strength.

For view-centroid coordinates $h_v = s + \epsilon_v$ along a fixed normal, the shared surface coordinate s cancels in $|h_v - h_w|$. Gap therefore measures relative offsets under the shared-surface assumption; it cannot establish correct correspondence, tangential agreement, or GT accuracy. Incorrect points can also collapse to zero gap.

2.5 Patch audit and selection of illustrations

The patch audit covers all 6,658 valid 48×48 grid patches in the six NRGBD scenes listed in Table 7. These were the scenes included in the saved audit; no independent scene-selection criterion was recorded, so this subset should not be treated as representative of all nine NRGBD scenes. Patch locations and validity use GT and input pixel identities, with at least 75% GT-valid retained pixels and sufficient connected edges. Every valid grid patch is included regardless of method outcome. All methods share pixels and the Align-only registration.

The two main-paper Figure 4 patches are selected illustrations: candidates must improve over both full-strength Direct controls by a factor exceeding 1.03 in paired 3D position, position error projected onto the GT normal, and adjacent-edge error, and are ranked by the product of these relative gains. The projected position error is measured in mm, not a normal-angle error. The audit uses one-pixel edges; the displayed examples report eight-pixel edges.

Ours has lower adjacent-edge error than Direct and Direct + smooth in 94.7% and 87.5% of patches, but lower error than Align in only 9.5%. Its pooled edge error is 3.017 versus 2.993 mm for Align. This supports retaining local structure close to Align while improving agreement and position, with better edge accuracy than the full-strength controls; it does not show improved local shape over Align.

Scene	Patches	Align	Direct	Direct + smooth	Ours
Paired 3D position error (mm)					
Breakfast room	959	16.667	16.446	16.414	**15.830**
Complete kitchen	1,037	26.083	25.793	25.759	**24.886**
Grey white room	1,095	18.261	18.008	17.973	**17.464**
Kitchen	1,298	24.400	24.186	24.149	**23.367**
Morning apartment	856	26.985	26.816	26.792	**25.759**
Whiteroom	1,413	16.209	15.860	15.797	**15.207**
Pooled	6,658	21.133	20.876	20.837	**20.123**
Adjacent one-pixel GT edge-vector error (mm)					
Breakfast room	959	**2.111**	2.334	2.217	2.133
Complete kitchen	1,037	**3.159**	3.415	3.282	3.189
Grey white room	1,095	**2.947**	3.194	3.060	2.972
Kitchen	1,298	**3.215**	3.473	3.334	3.242
Morning apartment	856	**3.191**	3.453	3.322	3.207
Whiteroom	1,413	**3.180**	3.543	3.337	3.204
Pooled	6,658	**2.993**	3.266	3.119	3.017
Ours edge wins (pooled)		9.5%	94.7%	87.5%	–

Table 7: Six-scene NRGBD patch audit, with every valid 48×48 grid patch retained regardless of outcome. Direct and Direct + smooth use full displacement ($\alpha = 1$). Means weight patches equally; bold marks the lowest error in each row. Edge wins count patches where Ours has lower error than the named comparator. These correlated patches do not constitute 6,658 independent scenes.

3 Ablations and Optimization Diagnostics

Pi3 loss and stage controls on NRGBD. Table 8 separates the three prespecified scenes from the nine-scene collapse extension. In the three-scene subset, removing collapse improves mean CD from 7.967 to 7.843 mm, while removing Laplacian worsens it to 8.072 mm. Across all nine original NRGBD scenes, the full network lowers fixed-pixel position error from 21.779 to 20.500 mm relative to the no-collapse variant, with a gain in 8/9 scenes. It also slightly lowers cohort-mean CD (8.172 to 8.127 mm), although only 5/9 individual scenes favor it, and the no-collapse variant has slightly higher NC. The effect of the Laplacian loss is still supported only by the three-scene subset; no loss is claimed necessary for every metric.

Three-scene loss controls	Acc ↓	Comp ↓	CD ↓	NC ↑
Align only	8.588	**7.365**	7.977	0.927
Without consensus	8.312	7.592	7.952	0.938
Without collapse	**8.142**	7.545	**7.843**	**0.940**
Without Laplacian	8.349	7.795	8.072	0.938
Refine only	8.329	7.576	7.953	0.939
Ours (300)	8.269	7.666	7.967	0.939

Nine-scene collapse control	Fixed-pixel error (mm)↓	CD (mm)↓	NC↑
Align only	21.272	8.180	0.931
Without collapse loss	21.779	8.172	**0.945**
Ours (300)	**20.500**	**8.127**	0.944

Table 8: Pi3/NRGBD loss controls. Top: three prespecified scenes, original full-point `score1` protocol with per-output Sim(3)+ICP; distances are mm. Bottom: all nine original scenes; removing collapse changes only its loss weight, retaining the cached base predictions, initialization, seed, and 300 steps. Fixed-pixel error uses one Align-only registration and identical pixel IDs; CD and NC use original full-point scoring. Bold marks the best score within each panel; the panels cover different scene sets.

DTU collapse-loss control. The complete 22-scene collapse-loss ablation in Table 9 tests whether that term alone causes the DTU coverage cost. Removing it increases mean fixed-pixel error from 6.069 to 6.394 mm and CD from 1.346 to 1.359 mm; Completeness also worsens from 1.885 to

1.914 mm. Accuracy and NC improve slightly. Thus removing this term does not recover bidirectional distance or coverage in this cohort; the cause of the tradeoff remains unresolved.

22 DTU scenes	Fixed pixel↓	Acc↓	Comp↓	CD↓	NC↑
Align only	6.154	0.816	**1.773**	**1.295**	0.685
Without collapse loss	6.394	**0.804**	1.914	1.359	**0.700**
Ours (300)	**6.069**	0.806	1.885	1.346	0.697

Table 9: Pi3/DTU collapse-loss ablation over all 22 original scenes. Except for the collapse weight (1 versus 0), the full and ablated runs use identical cached predictions, pretrained initialization, seed, and 300-step optimization. Full refinement has lower fixed-pixel error in 20/22 scenes and lower CD in 19/22; the ablation slightly improves mean Accuracy and NC. Distances are in mm. Fixed-pixel error uses one Align-only registration and identical pixels; other metrics use the original full-point evaluator.

Pretrained refiner before and after scene adaptation. On three prespecified NRGBD scenes (Table 10), the actual pretrained output at step 0 is worse than Align on every scene. After 300 steps, paired position error is lower on every scene. This isolates the benefit of adaptation in this subset, without separating architecture from losses.

Three NRGBD scenes	Fixed-pixel error (mm)↓	CD (mm)↓
Align only	22.178	7.977
Pretrained refiner, 0 steps	36.816	12.793
Ours (300)	**21.287**	**7.967**

Table 10: Equal-scene means on three prespecified Pi3/NRGBD scenes. All three outputs use the same cached backbone prediction, alignment, pretrained refiner weights, and retained pixels. The 0-step row is an actual pretrained-network forward pass, not Align only. Fixed-pixel error uses the Align-only registration for every output; CD follows the original independently registered full-point protocol.

Input sensitivity. Table 11 holds all nine original NRGBD scenes and the optimization budget fixed while zeroing one input stream at a time. Removing image patch tokens raises corresponding-pixel error from 20.500 to 21.163 mm, close to Align's 21.272 mm; the full network has lower error in all nine scenes. Removing pretrained 3D descriptor features leaves the mean essentially unchanged at 20.485 mm; voxel positions and raw point attributes remain available. This supports the practical value of image conditioning in this cohort but does not establish that the pretrained 3D descriptor branch or attention architecture is necessary. Bidirectional CD changes much less than corresponding-pixel error.

On DTU, image conditioning has the same direction on both bases (Table 12). Full refinement has lower fixed-pixel error than the zero-image input control in 21/22 Pi3 scenes and 19/22 VGGT-Ω scenes. Its mean error is 6.069 versus 6.155 mm for Pi3 and 7.041 versus 7.172 mm for VGGT-Ω. The full network has worse bidirectional CD than the no-image control on both bases, so this result supports image-conditioned position correction, not an unqualified surface-distance gain.

Measured optimization budget. Table 13 reports saved iterates from one Pi3/NRGBD scene. The warm-started refiner has its lowest measured symmetric distance at 30 steps (9.655 mm versus 9.765 mm for Align only); Normal Consistency continues to rise through 300 steps. This single-scene, single-seed comparison does not establish a general stopping rule.

3.1 Input-normal reliability

We evaluate the normal estimators used by the geometric losses on two Pi3 test sequences, `heads/01` and `chess/03` from 7-Scenes, using 25 views per sequence. The comparison includes per-pixel finite-difference normals (FD), their sign-invariant voxel aggregation (the dominant eigenvector of $\sum_i \mathbf{n}_i \mathbf{n}_i^\top$), and the minimum-variance axis of the pooled aligned points in each original

NRGBD scene	Align	Full	Zero 3D feat.	Zero image
breakfast room	16.966	16.226	16.190	17.105
complete kitchen	27.261	26.107	26.043	26.868
green room	17.662	17.386	17.443	17.632
grey white room	18.683	17.883	18.028	18.588
kitchen	25.850	24.870	24.679	25.668
morning apartment	28.148	26.803	26.815	27.992
staircase	28.630	28.106	27.968	28.395
thin geometry	11.124	10.952	11.015	11.185
whiteroom	17.127	16.168	16.185	17.038
Mean fixed-pixel error ↓	21.272	20.500	**20.485**	21.163
Mean CD ↓	8.180	8.127	**8.122**	8.159

Table 11: Pi3/NRGBD input sensitivity over all nine original scenes. All errors are mm. Scene rows report fixed-registration corresponding-pixel 3D error, with one Align-only registration and identical valid pixels; the final row reports full-point CD with independent registration. The 300 steps, losses, seed, and base predictions are fixed. Zeroing pretrained 3D descriptors retains voxel positions and raw attributes; zeroing image patch tokens retains voxel descriptors. Both perturbations act during scene optimization without retraining. Full has lower fixed-pixel error than zero image in all nine scenes; the comparison with zero 3D features is mixed (5/9 favor Full).

Base	Method	Fixed-pixel error (mm)↓	CD (mm)↓
Pi3	Align only	6.154	**1.295**
	Zero image patch tokens	6.155	1.298
	Ours (300)	**6.069**	1.346
VGGT-Ω	Align only	7.154	**1.477**
	Zero image patch tokens	7.172	1.481
	Ours (300)	**7.041**	1.505

Table 12: Image-input sensitivity over all 22 original DTU scenes for each base. Each ablation keeps the cached prediction, network initialization, seed, losses, and 300-step budget, but zeros image patch tokens during scene optimization. Fixed-pixel error shares the Align-only registration and exact pixel identities; CD uses the original full-point scoring with independent registration. Position improves with image tokens in 21/22 Pi3 and 19/22 VGGT-Ω scenes, while the full network has higher CD than either control.

Method	Acc ↓	Comp ↓	CD ↓	NC ↑
Align only	10.421	**9.109**	9.765	0.887
Ours (30 steps)	10.120	9.189	**9.655**	0.893
Ours (50 steps)	10.081	9.320	9.700	0.899
Ours (100 steps)	10.008	9.462	9.735	0.905
Ours (300 steps)	**9.975**	9.700	9.838	0.909
Random init (30)	10.404	9.212	9.808	0.891
Random init (50)	10.248	9.334	9.791	0.897
Random init (100)	10.069	9.543	9.806	0.904
Random init (300)	10.014	9.873	9.944	**0.910**
Free offsets (30)	10.133	10.057	10.095	0.906
Free offsets (100)	10.083	10.204	10.144	0.909
Align + Free offsets (300)	10.060	10.268	10.164	**0.910**

Table 13: Pi3 optimization budgets: original NRGBD morning apartment scene. Original full-point, per-output Sim(3)+ICP `score1`; distances are mm. Rows report measured saved iterates at the indicated step counts. Lower is better for distances, higher for NC; best values at shown precision are bold.

voxel (PCA). FD normals are rotated by the per-view Stage-1 transformations. All three estimates are then rotated into the same reference frame.

Protocol. We uniformly sample 100,000 shared valid input pixels per scene without replacement (seed 0) and use common valid queries across the three estimators. Reference normals are estimated independently within each input view from the full-density back-projected Kinect cloud, using $k = 30$ or 100 nearest neighbors. The error is $\arccos|\mathbf{n}^\top \mathbf{n}_{\mathrm{ref}}|$ in degrees. A query is excluded for every estimator if any estimate is invalid or the pooled voxel lacks two-dimensional support; this excludes 55 queries in `heads` and 98 in `chess`. Reference-derived strata are fixed independently of the predicted corrections. A boundary is an image edge, a missing axial reference neighbor, or an axial 3D jump above 5 cm. Stable interior points have no such boundary, reference-normal agreement within 10° between $k = 30$ and 100, and a second reference covariance eigenvalue above 10^{-14}.

Scene	Reference stratum	k	Queries	FD	Aggregated FD	PCA
heads/01	All sampled	30	99,945	18.76	**17.12**	18.46
	Stable interior	30	66,920	15.03	**13.45**	14.77
	Stable interior	100	66,920	13.11	**11.21**	12.67
	Boundary	30	4,195	31.00	**27.83**	29.83
chess/03	All sampled	30	99,902	24.25	**22.39**	25.33
	Stable interior	30	49,270	17.27	**15.44**	18.15
	Stable interior	100	49,270	15.31	**13.09**	16.18
	Boundary	30	4,115	39.74	**36.81**	39.67

Table 14: Mean sign-invariant angular error in degrees (↓) on common valid queries; best estimator per row is bold. The reference comes from sensor depth; its normals contain measurement and registration noise.

Results and scope. Aggregating the per-view FD normals reduces input angular error relative to raw FD and pooled-point PCA on these two sequences (Table 14); the stable-interior ordering holds for both reference neighborhood sizes. Boundary errors remain much higher for all estimators, including 27.83° and 36.81° for aggregated FD. Lower input angular error does not imply better final reconstruction: in a paired two-sequence fixed-registration pilot, replacing the collapse normal with pooled PCA slightly lowered CD on both sequences (5.241 to 5.226 mm and 9.332 to 9.250 mm). Neither pilot establishes correct cross-view correspondence or robustness to normal corruption. Spatially correlated pixel queries are not independent scene samples. Pooled PCA here uses all aligned points in each voxel, whereas training may estimate covariance from a random subset; this is an estimator diagnostic rather than a replay of the full alternative optimization.

3.2 Coverage effect and limitations

For fixed reference G and registration, define $\mathrm{Comp}(G, P) = |G|^{-1} \sum_{g \in G} \min_{p \in P} \|g - p\|$.

Proposition 1 (Completeness under set inclusion)**.** *If $P \subseteq P'$, then $\mathrm{Comp}(G, P') \leq \mathrm{Comp}(G, P)$, strictly when a newly added point is closer to at least one reference point than every point in P.*

Proof. For each g, minimizing over the superset P' cannot increase its distance. Averaging preserves the inequality and the stated strict case. □

Thus adding off-surface sheets can lower Completeness. This result assumes fixed registration and set inclusion: it does not establish monotonicity for moving a fixed number of points or independently re-registering outputs. The three-scene fixed-count experiment in main-paper Figure 5 demonstrates that effect for its chosen perturbations. It does not establish the cause of every benchmark regression.

Scope of the method. Proximity-based voxel groups can mix distinct surfaces or lack sufficient overlap. Displacement regularization does not guarantee detail preservation, and large updates can invalidate cached view associations. Gains are not universal across scenes or metrics.

4 Additional Benchmark Results

4.1 Completeness and symmetric distance

Method	DTU		ETH3D		ScanNet++	
	Mean	Med.	Mean	Med.	Mean	Med.
CUT3R	3.523	1.419	0.768	0.612	0.1013	0.0541
Fast3R	2.696	0.896	0.606	0.334	0.0658	0.0343
FLARE	3.247	1.404	0.580	0.352	0.0601	0.0370
VGGT	1.853	0.967	0.278	0.159	0.0312	0.0172
MapAnything	2.182	0.777	0.552	0.379	0.0358	0.0208
DA3	1.878	0.620	0.369	0.263	**0.0159**	<u>0.0093</u>
AMB3R	2.225	1.233	0.311	0.173	0.0258	0.0139
Test3R	3.550	0.961	1.742	1.291	1.4049	1.3481
TCO	3.097	2.276	0.249	0.131	0.0274	0.0156
Free Geometry	1.948	1.030	0.341	0.206	0.0232	0.0130
Pi3 (base)	**1.729**	<u>0.580</u>	0.207	0.120	0.0532	0.0397
VGGT-Ω (base)	<u>1.785</u>	**0.577**	<u>0.041</u>	<u>0.023</u>	0.0190	0.0108
Pi3 + Ours	1.885	0.684	0.207	0.108	0.0550	0.0411
VGGT-Ω + Ours	2.143	0.657	**0.035**	**0.021**	<u>0.0174</u>	**0.0090**

Table 15: Completeness (Comp↓; mean and median) under the main paper's shared-keyframe and registration protocol. DTU distances are mm; ETH3D and ScanNet++ distances are m. **Best** per column in bold, <u>second-best</u> underlined.

Symmetric distance. We additionally report the unsquared symmetric mean distance $\mathrm{CD} = \frac{1}{2}(\mathrm{Acc} + \mathrm{Comp})$. Table 16 is computed algebraically from full-benchmark directional means under the original evaluation protocol. The NRGBD/Pi3 row uses the current nine-scene full-point rerun with raw base output on every scene; other rows use archived aggregates. This criterion balances the two directions, although it remains sensitive to point density and the reference sampling. Improvement is not universal: VGGT-Ω on DTU and Pi3 on 7-Scenes have higher symmetric error after refinement. Consequently, improved Accuracy and Normal Consistency should not be interpreted as an unconditional improvement in overall reconstruction quality.

Dataset	Backbone	Raw base CD	Full pipeline CD	Reduction (%)
DTU	Pi3	1.473	1.346	+8.63
DTU	VGGT-Ω	1.471	1.505	-2.31
ETH3D	Pi3	199.312	184.586	+7.39
ETH3D	VGGT-Ω	45.199	35.936	+20.49
ScanNet++	Pi3	44.061	42.784	+2.90
ScanNet++	VGGT-Ω	23.400	17.077	+27.02
7-Scenes	Pi3	16.924	18.184	-7.45
7-Scenes	VGGT-Ω	15.979	12.420	+22.27
NRGBD	Pi3	13.643	8.127	+40.43
NRGBD	VGGT-Ω	8.234	5.761	+30.03

Table 16: Symmetric distance in mm, derived from full-point directional means. Reductions compare raw base outputs with the complete pipeline (input filtering, Align, and Refine), not Stage 2 alone; percentages use unrounded values. The NRGBD/Pi3 row uses the consistent raw-base rerun. Positive reduction is an improvement; negative values are regressions.

4.2 PER-SCENE ETH3D RESULTS

Table 17 gives the ETH3D Accuracy result for every scene under the main paper's alignment and scoring protocol. Our method improves 11 of 13 scenes for each base, and both indoor and outdoor means decrease. Pi3's large improvement on `meadow` accounts for 70% of its total reduction; excluding that scene, mean error still falls from 0.123 to 0.113 m. The four scene-level regressions are at most 7 mm.

	Pi3		VGGT-Ω	
Scene	base	+Ours	base	+Ours
Indoor (7 scenes)				
delivery_area	0.198	0.205	0.089	0.070
kicker	0.048	0.047	0.019	0.015
office	0.037	0.031	0.016	0.014
pipes	0.086	0.088	0.025	0.026
relief	0.067	0.045	0.031	0.018
relief_2	0.174	0.158	0.040	0.023
terrains	0.150	0.139	0.019	0.011
Indoor mean	0.109	0.102	0.034	0.025
Outdoor (6 scenes)				
courtyard	0.113	0.092	0.075	0.059
electro	0.106	0.088	0.045	0.029
facade	0.138	0.124	0.123	0.087
meadow	1.022	0.746	0.040	0.047
playground	0.266	0.256	0.067	0.055
terrace	0.091	0.085	0.055	0.029
Outdoor mean	0.289	0.232	0.067	0.051
All (13 scenes)	**0.192**	**0.162**	**0.049**	**0.037**

Table 17: Per-scene ETH3D Accuracy (metres, mean; lower is better) for both base models and after our test-time refinement (**+Ours**), under the same dense Umeyama + ICP alignment as the main paper. Aggregates match the ETH3D rows of the main paper's point-map table.

Outdoor examples. Figure 2 shows `electro` and `terrace` for both bases. Each whole-scene render is paired with a wall cross-section colored by input view. The complete Align + Refine pipeline narrows the view-offset bands in these examples, consistent with their Accuracy reductions; these are illustrations of the full pipeline, not isolated Stage 2 comparisons.

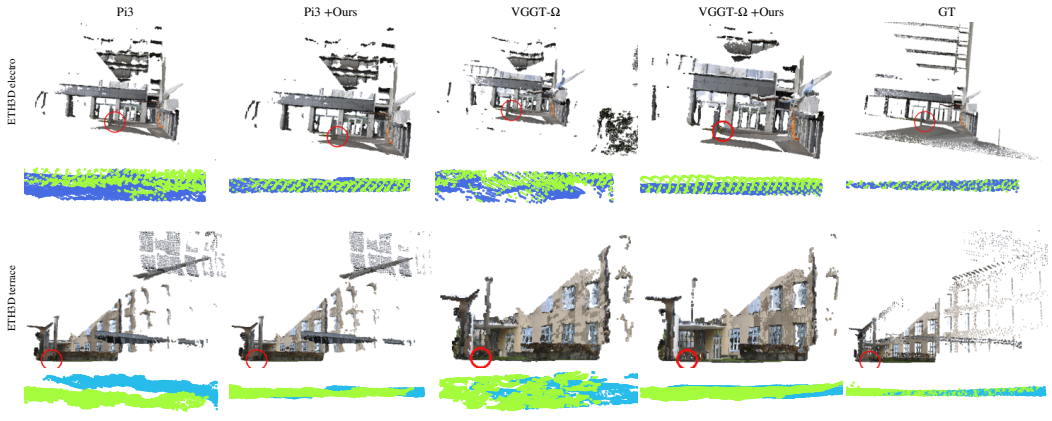

Figure 2: Outdoor ETH3D reconstructions before and after our complete two-stage pipeline. Columns: raw Pi3, Pi3 after Align + Refine (**+Ours**), raw VGGT-Ω, VGGT-Ω after Align + Refine (**+Ours**), and ground truth. Top of each cell: full-density RGB render. Bottom: along-normal cross-section of the flat wall patch (red circle). **Each color in a cross-section denotes one input view**, exposing view-offset sheets before correction and their narrower spread after Align + Refine.

4.3 7-SCENES AND NRGBD

The 7-Scenes reference is built by back-projecting Kinect depths from multiple frames and has visible surface thickness (Figure 3). We therefore report it separately from the main benchmarks. Our method lowers Accuracy error and improves Normal Consistency for both bases (Table 18); the mixed Completeness and symmetric-distance changes are retained. This single reference example does not establish the cause of those regressions.

	Acc↓		Comp↓		NC↑	
Method	Mean	Med.	Mean	Med.	Mean	Med.
7-Scenes (real captures, finite-thickness reference)						
Pi3 (base)	0.0136	0.0063	0.0203	0.0101	0.697	0.800
Pi3 + Ours	0.0131	0.0060	0.0233	0.0138	0.712	0.818
VGGT-Ω (base)	0.0155	0.0067	0.0165	0.0063	0.683	0.781
VGGT-Ω + Ours	0.0088	0.0048	0.0161	0.0084	0.707	0.809
NRGBD (single-surface synthetic reference)						
Pi3 (base)	0.0141	0.0078	0.0132	0.0048	0.891	0.986
Pi3 + Ours	0.0086	0.0056	0.0076	0.0045	0.944	0.993
VGGT-Ω (base)	0.0106	0.0062	0.0059	0.0039	0.904	0.990
VGGT-Ω + Ours	0.0063	0.0045	0.0052	0.0035	0.956	0.995

Table 18: Point-map estimation on 7-Scenes and NRGBD (metres; Acc/Comp lower is better, NC higher). All numbers use the main-paper full-point protocol. The NRGBD/Pi3 base row uses raw model output on all nine scenes; +Ours includes input filtering. On 7-Scenes the complete pipeline improves Accuracy and Normal Consistency on both bases; on NRGBD it improves all three mean metrics. Stage-2 effects are assessed against Align only on identical retained pixels in the main-paper controls.

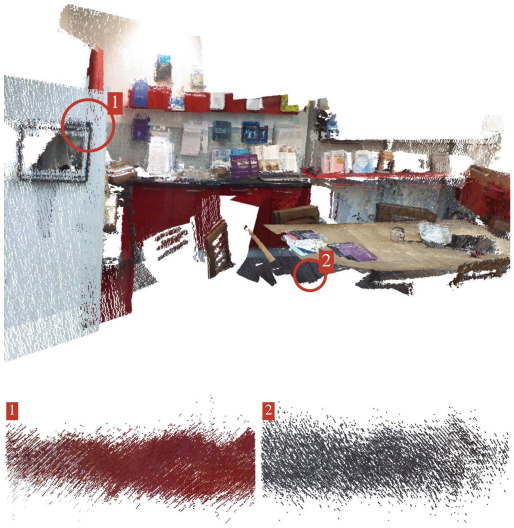

Figure 3: The 7-Scenes reference cloud (`redkitchen`, `seq-03`) at full density. *Below*: the two circled patches viewed edge-on, in the points' own colors with the along-normal axis stretched 6× (visible range ±2.3 cm; both panels share this scale). At this magnification, a single-surface reference would appear as a thin line.