# W3D: WHERE, WHO, AND WHY? DIAGNOSING MULTI-VIEW 3D FUSION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multi-view fusion assessment should explain local geometric errors rather than merely assign reliability scores. We introduce W3D, a query-centric framework that formulates fusion assessment as three coupled diagnostic objectives: Where localizes unreliable geometry, Who traces view-level influence through leave-one-view-out effects, and Why predicts multi-label observation factors. W3D reconnects each fused-surface query to its original multi-view observations using a point-normal-residual (PNR) representation that captures relative depth, viewing geometry, and surface orientation. A shared Set Transformer supports gated Where $\rightarrow$ Who $\rightarrow$ Why message passing, while stop-gradient operations decouple task-head updates through these messages without preventing joint learning of shared features. Supervision combines reference geometry with leave-one-view-out re-fusion, whereas inference requires only the fused surface, original observations, and camera calibration. On Skoltech3D, W3D consistently improves all six diagnostic metrics over the matched parallel Base with identical PNR inputs, remains effective across varying numbers of input views, and supports multiple fusion algorithms under the same diagnostic formulation through separate training. These results show that local fusion quality can be expressed as a traceable diagnosis linking geometric errors to their supporting observations.

## 1 INTRODUCTION

Multi-view fusion integrates depth observations across viewpoints into a unified geometric surface for downstream tasks such as navigation, grasping, and measurement. While global metrics summarize overall reconstruction accuracy, local quality estimates reveal where the fused geometry is unreliable. Recent work models such local reliability through uncertainty propagation, Bayesian fusion, evidential uncertainty, cross-view consistency, or visibility reasoning (Liao & Waslander, 2024; Yang et al., 2025; Zhu et al., 2026; Vats et al., 2024; Su et al., 2026). However, a local reliability score does not explain which observations are associated with the degradation or what abnormal observation patterns characterize it.

We formulate multi-view fusion quality as a structured diagnosis rather than an isolated reliability prediction. For each fused query point, diagnosis should answer three linked questions: Where is the local geometry unreliable, Who among the input views is most relevant to the degradation, and Why do the observations appear abnormal? These outputs are coupled by construction. Local quality provides context for view tracing, while view-level evidence supports factor diagnosis. A useful formulation must therefore reconnect fused geometry to its original observations while preserving source identity throughout prediction. This organization is also consistent with broader principles of evidence-based inference. In multisensory perception, biological systems distinguish whether heterogeneous signals should be integrated or attributed to different sources, while in physics, inverse problems infer latent causes from observed consequences under a forward model (Rohe et al., 2019; Calvetti & Somersalo, 2018).

We introduce W3D, a query-centric framework for structured diagnosis of multi-view 3D fusion. Each fused query is associated with its multi-view observation set and encoded using a point-normal-residual (PNR) representation that captures relative depth, viewing geometry, and surface orientation. A shared set encoder models cross-view relations, and gated message passing follows a Where $\rightarrow$ Who $\rightarrow$ Why hierarchy. Stop-gradient operations preserve forward information exchange while



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.

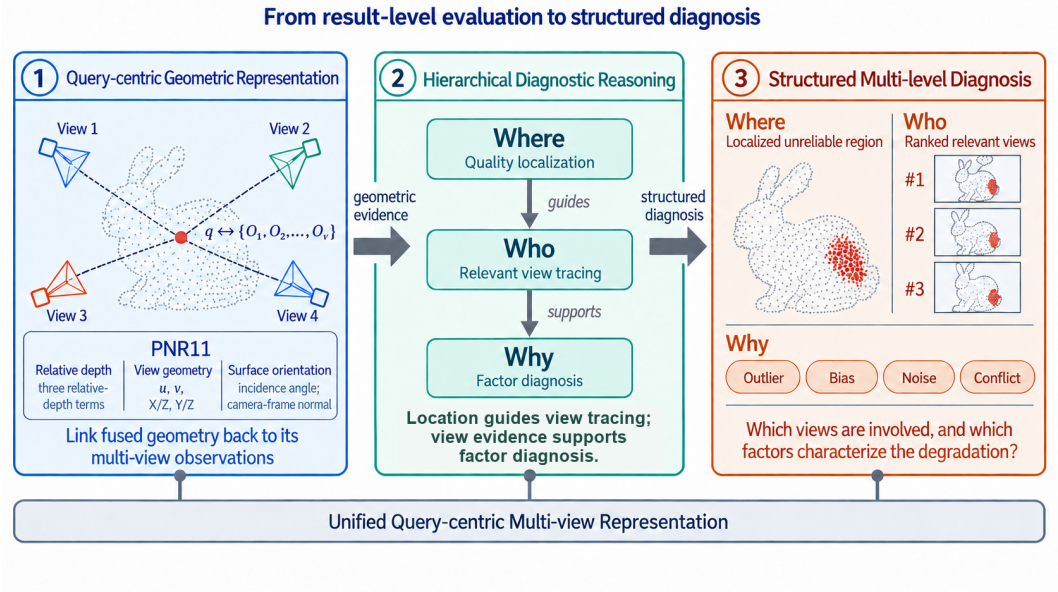


Figure 1: Overview of W3D. PNR links fused queries to multi-view observations, and Where → Who → Why organizes structured diagnosis. Factor labels are conceptual; PNR denotes the 11-dimensional PNR representation.

preventing downstream losses from directly updating upstream task heads. Offline supervision combines reference geometry with leave-one-view-out re-fusion, whereas inference uses only the fused surface, original observations, and camera calibration. W3D improves all six diagnostic metrics over the matched parallel Base with the same PNR input, remains effective across varying numbers of input views, and supports multiple fusion algorithms under the same diagnostic formulation through separate training.

Our contributions are threefold:

- **Query-centric geometric representation.** We associate each fused query with its original multi-view observations and encode their relative geometry using PNR.
- **Hierarchical diagnostic reasoning.** We organize diagnosis as Where → Who → Why with gated message passing and stop-gradient task decoupling.
- **Structured multi-level diagnosis**. We unify local quality localization, relevant-view tracing, and multi-label factor diagnosis into a traceable description of fusion anomalies.

## 2 Related Work

**Multi-view reconstruction and fusion.** Classical fusion methods aggregate geometric observations into explicit or implicit surfaces. TSDF integrates range measurements into a volumetric signed-distance representation (Curless & Levoy, 1996), while Screened Poisson reconstructs implicit surfaces from oriented points (Kazhdan & Hoppe, 2013). Recent learned approaches predict scene geometry directly: DUSt3R formulates geometric estimation through pointmaps (Wang et al., 2024), MASt3R combines image matching with 3D prediction (Leroy et al., 2024), MUSt3R extends joint prediction to multiple views (Cabon et al., 2025), and VGGT jointly estimates cameras, depth, and other geometric attributes (Wang et al., 2025). W3D is complementary to these reconstruction methods: it operates after fusion and diagnoses the quality and observational origins of the resulting geometry rather than redesigning reconstruction.

**Geometric reliability and multi-view evidence.** Recent work extends global reconstruction metrics with local reliability estimation. Liao & Waslander (2024) propagate uncertainty from shape

representations to signed-distance predictions and perform Bayesian fusion; GURecon models continuous geometric uncertainty fields (Yang et al., 2025); and Trust3R introduces evidential uncertainty for feed-forward pointmap reconstruction (Zhu et al., 2026). Other methods exploit relations among observations to improve reconstruction itself: GC-MVSNet enforces multi-view and multi-scale geometric consistency (Vats et al., 2024), while GVGS uses cross-view visibility for geometric supervision (Su et al., 2026). These methods primarily estimate or exploit reliability within the reconstruction pipeline. W3D instead treats local quality, view relevance, and diagnostic factors as linked outputs at a shared fused-surface query, enabling traceable post-fusion diagnosis.

**Set representations and diagnostic relations.** Set Transformer models interactions among unordered elements using attention, inducing points, and attention-based pooling (Lee et al., 2019). Such architectures suit multi-view observations with variable cardinality and ordering. W3D adopts a Set Transformer backbone for observation encoding, while our contribution lies in how diagnostic evidence is represented and exchanged: PNR encodes query-relative geometry, while detached gated messages organize information flow across Where, Who, and Why. Thus, the set encoder models view interactions, whereas PNR and message passing structure the diagnostic task.

# 3 W3D Preliminaries

Given a fused surface, calibrated multi-view depth maps, and a query point $q$, projection associates $q$ with valid observations from the input views. W3D predicts a local bad-quality probability (Where), view-level relevance weights (Who), and four multi-label diagnostic factor probabilities (Why). Who represents leave-one-view-out re-fusion effects, whereas Why predicts Bias, Residual Outlier, Noise, and Multi-view Conflict. All outputs are aligned to the same query and observation indices, preserving source traceability without using view identity as a PNR feature. Reference geometry is used only for offline supervision and is not required at inference.

# 4 Method

W3D combines a query-relative geometric representation (Section 4.1), a shared encoder with detached gated prediction messages (Section 4.2), and task-specific supervision and losses (Sections 4.3–4.4).

## 4.1 PNR Geometric Representation

The raw7 baseline contains normalized image coordinates, absolute observed depth, incidence cosine, and three view-slot codes: linear, squared, and sinusoidal functions of the view index. PNR replaces absolute depth with relative depth, removes slot codes, and adds normalized camera coordinates and camera-frame normals. Section 5.3 evaluates these cumulative changes.

PNR encodes each observation in 11 dimensions. For observed and query depths $d_s, d_q$, let $r = d_s - d_q$. Relative-depth terms comprise $\frac{r}{\ell_0}$, $r/(d_q + \varepsilon)$, and a log-depth ratio, with fixed scale $\ell_0 = 10\,\mathrm{mm}$. Normalized image and camera coordinates describe viewpoint geometry; incidence cosine $\kappa$ and camera-frame normals describe orientation:

$$x_{q,v} = [u_n, v_n, \kappa, r/\ell_0, r/(d_q + \varepsilon), \log(d_s/d_q), x_c/z_c, y_c/z_c, n_x^c, n_y^c, n_z^c].$$

PNR uses neither reference geometry nor fixed view identifiers. Normalization is fitted on valid training observations; invalid observations are zeroed and masked. Appendix B specifies numerical safeguards and feature dimensions.

## 4.2 Detached Gated Relational Network

A shared set encoder feeds the three task heads (Figure 2). Gated residuals incorporate upstream predictions into downstream heads. Stop-gradient (SG) preserves forward messages while blocking downstream losses from updating upstream heads through those messages. Shared features remain differentiable.

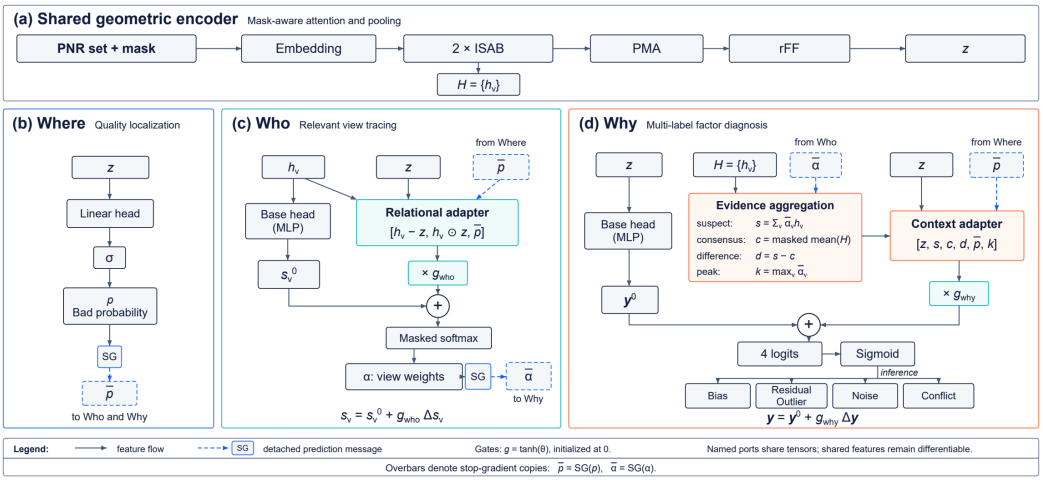


Figure 2: Detached gated message passing. Shared features H and z remain differentiable; SG copies of Where and Who predictions enter downstream adapters. Why outputs Bias, Residual Outlier, Noise, and Conflict probabilities in this order.

#### 4.2.1 Set Encoding and Where Prediction

The backbone uses Set Transformer ISAB and PMA modules (Lee et al., 2019).

Linear embedding and two ISAB layers produce $H = \{\mathbf{h}_v\}$; PMA and rFF produce $\mathbf{z}$. Attention and pooling use validity masks. $H$ is permutation-equivariant, and $\mathbf{z}$ is invariant. A linear head $f_W$ predicts bad-quality probability:

$$p = \mathrm{sigmoid}(f_W(\mathbf{z})), \overline{p} = \mathrm{SG}(p).$$

#### 4.2.2 Where-guided Who Residual

The base head produces view scores $s_v^0$. Adapter $f_{\mathrm{rel}}$ combines view–global differences, element-wise interactions, and detached quality prompt $\overline{p}$:

$$\Delta s_v = f_{\mathrm{rel}}([\mathbf{h}_v - \mathbf{z}, \mathbf{h}_v \odot \mathbf{z}, \overline{p}]), s_v = s_v^0 + g_{\mathrm{who}} \Delta s_v.$$

Masked softmax maps $s_v$ to weights $\alpha_v$, with zero weight for invalid views. Outputs follow the input permutation; $\overline{\alpha} = \mathrm{SG}(\alpha)$ supplies evidence to Why.

#### 4.2.3 Who-guided Why Residual

Detached Who weights aggregate suspect evidence; the unweighted valid-view mean supplies consensus. For valid-view set $\mathcal{V}_q$, define:

$$\mathbf{s} = \sum_{v \in \mathcal{V}_q} \overline{\alpha}_v \mathbf{h}_v, \mathbf{c} = \frac{1}{|\mathcal{V}_q|} \sum_{v \in \mathcal{V}_q} \mathbf{h}_v, \mathbf{d} = \mathbf{s} - \mathbf{c}, k = \max_v \overline{\alpha}_v.$$

Context adapter $f_{\mathrm{ctx}}$ combines global, suspect, consensus, difference, quality, and peak-weight information to correct base logits $\mathbf{y}^0$:

$$\Delta \mathbf{y} = f_{\mathrm{ctx}}([\mathbf{z}, \mathbf{s}, \mathbf{c}, \mathbf{d}, \overline{p}, k]), \mathbf{y} = \mathbf{y}^0 + g_{\mathrm{why}} \Delta \mathbf{y}, \hat{\mathbf{y}} = \mathrm{sigmoid}(\mathbf{y}).$$

Both scalar gates use $g = \tanh(\theta)$, with $\theta$ initialized to zero. Four sigmoids predict factors independently. Overbars denote SG copies, not means. Where→Who, Where→Why, and Who→Why messages are detached; shared $H, \mathbf{z}$ is jointly optimized.

### 4.3 Diagnostic Supervision

Reference geometry defines offline supervision. Where labels queries at least 5 mm from the reference surface as Bad. Why uses local residual shift, outliers, dispersion, and cross-view normal inconsistency. Who compares local distance-field errors $E_{\text{all}}$ and $E_{-v}$ before and after view removal, retaining positive reductions above 0.1 mm:

$$\Delta_v = E_{\text{all}} - E_{-v}, w_v = \Delta_v \mathbb{1}[\Delta_v > 0.1\,\text{mm}], t_{q,v} = \frac{w_v}{\sum_j w_j}.$$

Who requires Bad queries, complete counterfactual evidence, and positive effects representable by the input mask. Query-level Why labels are also conditioned on Bad. Missing evidence remains unknown and is masked from the losses. These targets represent protocol-specific removal effects and observation factors rather than established physical failure causes (Appendix A).

### 4.4 Joint Objective

Where and Why use binary and multi-label binary cross-entropy; Who uses cross-entropy between target $\mathbf{t}_q$ and prediction $\alpha_q$. Each loss is normalized using task-validity and confidence weights, and is zero without valid supervision:

$$\mathcal{L} = \mathcal{L}_{\text{Where}} + \mathcal{L}_{\text{Who}} + \mathcal{L}_{\text{Why}}.$$

Parallel Base and W3D train from scratch under the same backbone configuration, split, and training protocol, without pretrained Base weights or warm-up (Appendix C).

## 5 Experiments

### 5.1 Setup and Metrics

We use 101 Skoltech3D scenes (Voynov et al., 2023), split into 71/15/15 training/validation/test scenes. Controlled ablations use seven-view TSDF with fixed queries, labels, and training protocol. Section 5.5 trains separately for each fusion method. Models train for 30 epochs; the earliest maximum-validation-Where-AP checkpoint is selected. We report means $\pm$ sample standard deviations over seeds 42, 2023, and 2024. Section 5.7 evaluates native-score transfer; Appendix E reports additional controls.

We report six metrics spanning Where, Who, and Why; higher is better throughout.

- Where AP and AUROC measure ranking quality with Bad queries as positives; random-ranking AUROC is 0.5.
- Who Top-1 measures agreement between the highest-weight predicted and target views. Who Mass is the target probability assigned to the predicted top view, not prediction confidence.
- Why AP macro-averages AP over evaluable factors; Outlier AP reports Residual Outlier separately.

Metrics are weighted over valid queries within each scene, then averaged equally over valid scenes. AUROC uses area weights and requires both classes; AP terms without positives are excluded. Tables 1–3 use fixed-test re-inference to add AUROC; scene-level AP matches the original reports at all 30 training endpoints. External comparisons use common valid queries; fixed pretrained models have single-run results.

### 5.2 Representation $\times$ Architecture Ablation

Our $2 \times 2$ design crosses raw7/PNR inputs with parallel Base/detached-gated architectures. Backbone, labels, splits, and training remain fixed. The gated model adds adapters, so this is not a parameter-matched comparison.

Table 1: Representation × architecture ablation

| Architecture | Input | Where AP ↑ | Where AUROC ↑ | Who Top-1 ↑ | Who Mass ↑ | Why AP ↑ | Outlier AP ↑ |
|---|---|---|---|---|---|---|---|
| Parallel Base | raw7 | 0.89697 ±0.00811 | 0.70097 ±0.03308 | 0.39994 ±0.01857 | 0.36972 ±0.01249 | 0.86222 ±0.01801 | 0.55403 ±0.06478 |
| Parallel Base | PNR | 0.91825 ±0.00447 | 0.80348 ±0.00947 | 0.54580 ±0.02037 | 0.48895 ±0.02033 | 0.89102 ±0.00606 | 0.63369 ±0.02987 |
| Gated + SG | raw7 | 0.89510 ±0.01979 | 0.69311 ±0.05925 | 0.38161 ±0.01126 | 0.35582 ±0.00629 | 0.88783 ±0.01267 | 0.63293 ±0.04519 |
| Gated + SG | PNR | **0.92710 ±0.00169** | **0.80978 ±0.00696** | **0.55646 ±0.00323** | **0.49900 ±0.00294** | **0.90683 ±0.00562** | **0.67461 ±0.02272** |

Table 2: Progressive representation ablation

| Stage | Dim. | Where AP ↑ | Where AUROC ↑ | Who Top-1 ↑ | Who Mass ↑ | Why AP ↑ | Outlier AP ↑ |
|---|---|---|---|---|---|---|---|
| A0 raw7 | 7 | 0.89510 ±0.01979 | 0.69311 ±0.05925 | 0.38161 ±0.01126 | 0.35582 ±0.00629 | 0.88783 ±0.01267 | 0.63293 ±0.04519 |
| A1 Relative depth | 9 | 0.92071 ±0.00277 | 0.80630 ±0.00260 | 0.56627 ±0.01149 | 0.50658 ±0.00918 | 0.90059 ±0.00642 | 0.65338 ±0.01747 |
| A2 No slots | 6 | 0.91365 ±0.00516 | 0.78154 ±0.00832 | 0.54894 ±0.00606 | 0.49287 ±0.00439 | 0.90752 ±0.00609 | 0.67930 ±0.02201 |
| A3 + camXY | 8 | 0.91726 ±0.00554 | 0.79333 ±0.01005 | 0.54044 ±0.01419 | 0.48419 ±0.01362 | 0.90186 ±0.00232 | 0.65516 ±0.00780 |
| A4 Full PNR | 11 | 0.92710 ±0.00169 | 0.80978 ±0.00696 | 0.55646 ±0.00323 | 0.49900 ±0.00294 | 0.90683 ±0.00562 | 0.67461 ±0.02272 |

PNR improves all six metrics in both architectures. For parallel Base, Who Top-1 and Mass rise by 14.59 and 11.92 percentage points (pp), and Outlier AP by 7.97 pp. For the detached gated model, the two Who gains are 17.49 and 14.32 pp.

With identical PNR inputs, detached gated reasoning improves Where AP/AUROC by 0.89/0.63 pp, Who Top-1/Mass by 1.07/1.01 pp, and Why/Outlier AP by 1.58/4.09 pp. W3D has the highest means on all six metrics. With raw7, architectural gains concentrate on Why and Outlier, linking their benefits to geometric input design.

Together, these results support combining explicit relative geometry with hierarchical prediction messages. The architectural factor evaluates gating and gradient decoupling jointly.

### 5.3 Progressive Representation Ablation

We fix the detached gated architecture and training protocol and train each cumulative representation from scratch with three seeds. Only the input embedding changes with dimensionality. A0 and A4 reuse the matching endpoints of Section 5.2.

A0 uses raw7. A1 replaces absolute depth with three relative-depth terms; A2 removes slot codes; A3 adds camera x/z and y/z; A4 adds camera-frame normals to obtain PNR. A1–A4 use the same 10 mm physical scale.

Relative depth yields the largest view-tracing gain: A0→A1 improves Who Top-1/Mass by 18.47/15.08 pp and Where AP/AUROC by 2.56/11.32 pp. All six metrics improve.

Feature groups contribute differently across objectives. A1 has the strongest Who means; A2 has the highest Why and Outlier AP (0.90752/0.67930). Adding camera coordinates improves Where AP over A2; adding normals then improves every metric over A3.

Full PNR outperforms raw7 on all metrics and achieves the best Where AP/AUROC (0.92710/0.80978). These task-specific peaks indicate complementary feature contributions rather than uniform gains from every added component.

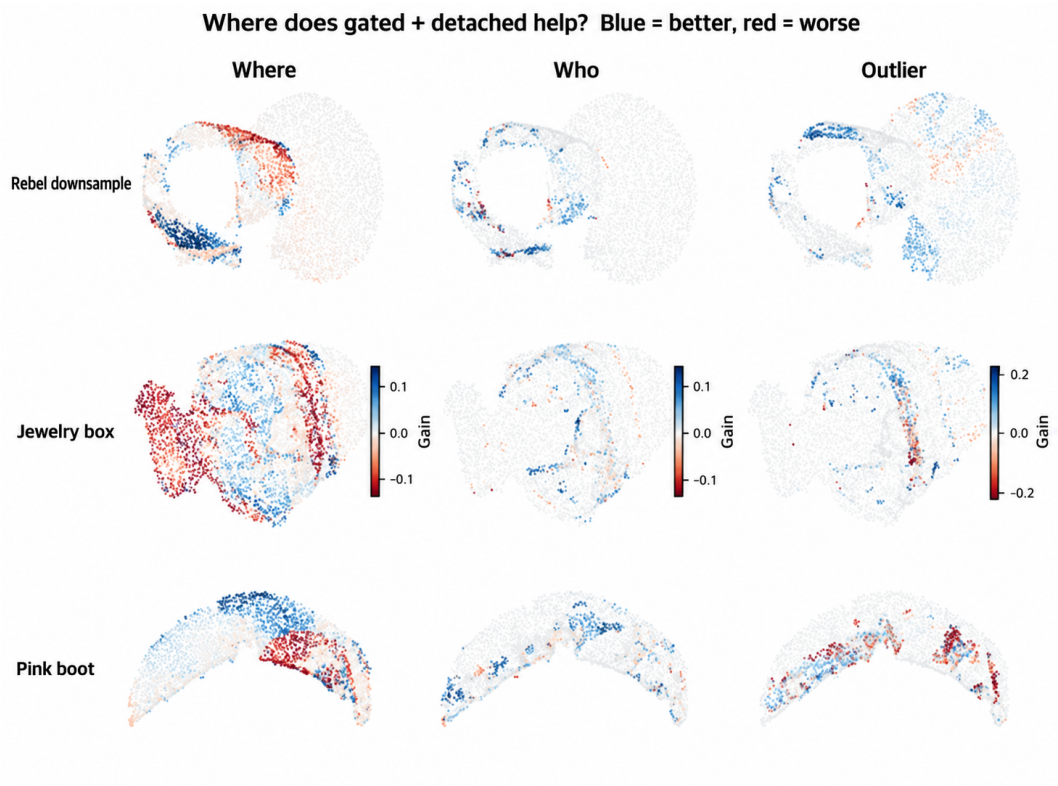


Figure 3: Pointwise gains over Base+PNR: blue indicates improvement, red degradation, and gray queries without valid labels. Rows are Rebel downsample, Jewelry box, and Pink boot, as labeled in the figure. Where and Outlier show the difference between Base and W3D squared errors; Who shows the target-weighted prediction difference.

### 5.4 Qualitative Comparison with Identical PNR Inputs

Figure 3 compares Base+PNR and W3D on paired test queries using identical coordinates, PCA projections, and validity masks. Predictions are averaged over three seeds. Scene selection uses paired Outlier AP differences, including two larger-gain cases and one declining case.

The gains are spatially heterogeneous. To test diagnostic utility, we also select one view per query and evaluate its signed local LOO effect. On Who-valid queries, W3D reduces local field error by $0.30832 \pm 0.00405$ mm, versus 0.22586 mm for residual-deviation ranking and 0.06613 mm for random selection (Appendix E.2). This evaluates local selection utility rather than whole-scene repair.

### 5.5 Adaptation across Fusion Methods

We construct data using TSDF, Voxel-centroid, Surfel, and Screened Poisson and train W3D independently for each. PNR, architecture, label rules, scene splits, and training protocol are shared; normalization is fitted separately, and weights are not transferred.

The same diagnostic design applies to all four methods after separate training (Table 3). Where AP spans 0.89000–0.97504, and Why AP spans 0.89284–0.90683.

Screened Poisson has the highest Where AP/AUROC and Outlier AP; TSDF has the highest Why AP and both Who metrics. Different rankings across objectives further motivate multi-level assessment.

This experiment evaluates reuse of the diagnostic interface rather than ranking reconstruction quality. Query geometry, label distributions, and supervision coverage differ across fusion methods (Appendix D).

Table 3: Adaptation across fusion methods

| Fusion method | Where AP ↑ | Where AUROC ↑ | Who Top-1 ↑ | Who Mass ↑ | Why AP ↑ | Outlier AP ↑ |
|---|---|---|---|---|---|---|
| TSDF | 0.92710 ±0.00169 | 0.80978 ±0.00696 | 0.55646 ±0.00323 | 0.49900 ±0.00294 | 0.90683 ±0.00562 | 0.67461 ±0.02272 |
| Voxel-centroid | 0.90666 ±0.00237 | 0.79027 ±0.00703 | 0.52857 ±0.01599 | 0.48284 ±0.01308 | 0.89284 ±0.01118 | 0.62388 ±0.03964 |
| Surfel | 0.89000 ±0.00509 | 0.75729 ±0.01302 | 0.50373 ±0.01173 | 0.48020 ±0.01046 | 0.90012 ±0.00085 | 0.63731 ±0.00328 |
| Screened Poisson | 0.97504 ±0.00187 | 0.82084 ±0.00659 | 0.44656 ±0.03378 | 0.41573 ±0.02873 | 0.90043 ±0.00666 | 0.67828 ±0.02147 |

Table 4: Comparisons across input view counts

| Views | Model / input | Where AP ↑ | Where AUROC ↑ | Who Top-1 ↑ | Who Mass ↑ | Why AP ↑ | Outlier AP ↑ |
|---|---|---|---|---|---|---|---|
| 3 | Base / raw7 | 0.86239 ±0.00877 | 0.64123 ±0.02786 | 0.69534 ±0.04387 | 0.65289 ±0.03953 | 0.76774 ±0.00665 | 0.30177 ±0.01840 |
| 3 | W3D / PNR | 0.90708 ±0.00301 | 0.77185 ±0.00437 | 0.77355 ±0.00569 | 0.72254 ±0.00824 | 0.80940 ±0.00508 | 0.40863 ±0.01688 |
| 5 | Base / raw7 | 0.87272 ±0.01535 | 0.67261 ±0.05197 | 0.46774 ±0.02911 | 0.42720 ±0.02797 | 0.81449 ±0.00432 | 0.43761 ±0.01205 |
| 5 | W3D / PNR | 0.91045 ±0.00254 | 0.79051 ±0.01258 | 0.58147 ±0.00501 | 0.53980 ±0.00416 | 0.85453 ±0.01703 | 0.52215 ±0.06863 |
| 7 | Base / raw7 | 0.89697 ±0.00811 | 0.70097 ±0.03308 | 0.39994 ±0.01857 | 0.36972 ±0.01249 | 0.86222 ±0.01801 | 0.55403 ±0.06478 |
| 7 | W3D / PNR | 0.92710 ±0.00169 | 0.80978 ±0.00696 | 0.55646 ±0.00323 | 0.49900 ±0.00294 | 0.90683 ±0.00562 | 0.67461 ±0.02272 |
| 9 | Base / raw7 | 0.92044 ±0.00097 | 0.75555 ±0.00296 | 0.39205 ±0.04839 | 0.36678 ±0.04284 | 0.91805 ±0.00475 | 0.70435 ±0.01281 |
| 9 | W3D / PNR | 0.93520 ±0.00328 | 0.81040 ±0.00278 | 0.53112 ±0.00597 | 0.48984 ±0.00530 | 0.93632 ±0.00927 | 0.76247 ±0.03645 |
| 11 | Base / raw7 | 0.92013 ±0.00327 | 0.74098 ±0.02093 | 0.37460 ±0.05511 | 0.35058 ±0.04899 | 0.92769 ±0.00436 | 0.73799 ±0.01893 |
| 11 | W3D / PNR | 0.93529 ±0.00526 | 0.81711 ±0.00995 | 0.50859 ±0.00312 | 0.46387 ±0.00278 | 0.94617 ±0.00049 | 0.79616 ±0.00294 |

### 5.6 Effectiveness across View Counts

We compare Base+raw7 with W3D+PNR at 3, 5, 7, 9, and 11 views, training each configuration from scratch with three seeds. This evaluates the complete representation–architecture combination; Table 1 isolates architecture at fixed inputs.

W3D improves all six metrics across the tested view counts. Gains are 1.48–4.47 pp for Where AP, 5.49–13.06 for AUROC, 7.82–15.65 for Who Top-1, 6.97–12.93 for Who Mass, 1.83–4.46 for Why AP, and 5.81–12.06 for Outlier AP.

The gains persist across all tested view counts. In a separate transfer test, seven-view weights and normalization remain frozen while each target count uses its own fusion and labels. At 3/5/9/11 views, Who Top-1 is 0.75980/0.58826/0.52916/0.50036, within 1.38 pp of count-specific training; all six means exceed Base+raw7 at the corresponding count (Appendix E.3).

### 5.7 Task-Specific Model Comparisons

We compare Where with Trust3R uncertainty on ETH3D, Who with native Vis-MVSNet and Patch-MatchNet view-score proxies on DTU, and Why with a Bayesian-last-layer variant on Skoltech3D. The latter changes the final Why residual layer to diagonal-Gaussian variational weights while retaining the other components. Unevaluated tasks are marked "—".

W3D uses 7/2 DSLR training/validation scenes and five rig test scenes on ETH3D, and a 63/17/22 scene split on DTU. External models retain pretrained weights; Trust3R uncertainty is averaged

Table 5: Task-specific comparisons across datasets

| Task | Dataset | Method | Where AP ↑ | Where AUROC ↑ | Who Top-1 ↑ | Who Mass ↑ | Why AP ↑ | Outlier AP ↑ |
|---|---|---|---|---|---|---|---|---|
| Where | ETH3D | Trust3R | 0.76285 | 0.46907 | — | — | — | — |
| Where | ETH3D | W3D | **0.76831** ±**0.00415** | **0.48429** ±**0.00348** | — | — | — | — |
| Who | DTU | Vis-MVSNet | — | — | 0.26235 | 0.24919 | — | — |
| Who | DTU | PatchMatchNet | — | — | 0.19315 | 0.18671 | — | — |
| Who | DTU | W3D | — | — | **0.53338** ±**0.02721** | **0.52882** ±**0.02686** | — | — |
| Why | Sk3D | Bayesian Why / PNR | — | — | — | — | 0.89382 ±0.02092 | 0.64318 ±0.06936 |
| Why | Sk3D | W3D / PNR | — | — | — | — | **0.90683** ±**0.00562** | **0.67461** ±**0.02272** |

over valid views, and the MVS models receive no Who head. Scores use the same commonly valid labeled queries and scene-weighted macro-averaging. W3D receives target-domain diagnostic supervision, so these are comparisons of the stated usage protocols, not matched-training architecture rankings.

On DTU, W3D reaches 0.53338 Who Top-1 and 0.52882 Who Mass, exceeding Vis-MVSNet by 27.10/27.96 pp. This comparison covers 625 common queries, or 39.79% of Who-valid area.

On Skoltech3D, the deterministic Why layer improves Why/Outlier AP by 1.30/3.14 pp over the Bayesian variant, supporting the selected configuration.

On ETH3D, W3D exceeds Trust3R by 0.55 pp in AP and 1.52 pp in AUROC. However, AP is near the 76.70% macro-averaged Bad prevalence, and both AUROCs are below 0.5. These laser-local-plane proxy labels do not establish reliable cross-scene localization.

## 6 Discussion

**Limitations.** Frozen-weight transfer is demonstrated across the tested TSDF view counts, whereas fusion-method experiments require separate training. Individual message and gradient paths need targeted interventions. Bias and Noise labels are highly positive-skewed, so their near-unit AP must be read with class prevalence and two-class scene counts (Appendix E.4). Who effects and Why factors do not identify physical sensor faults.

**Broader impact.** Observation-linked diagnosis can support targeted inspection. The LOO analysis quantifies local selection utility, while automatic whole-scene repair and reduced reacquisition costs require downstream evaluation. Coverage must accompany predictions: missing evidence does not imply reliability, and view rankings alone should not dictate deletion.

**Conclusion.** W3D connects fused geometry to multi-view evidence through PNR and detached gated message passing. At fixed PNR inputs, it improves all six metrics over parallel Base. Frozen seven-view weights transfer across the tested TSDF counts, while the diagnostic design supports four fusion methods through separate training.

## Acknowledgements

The authors thank the anonymous reviewers for their valuable feedback.

**AI Use Statement.** We used LLMs for manuscript polishing, literature retrieval, narrative structuring, and section drafting. All core ideas, methods, experiments, and conclusions are the sole work of the authors. All AI-assisted text has been thoroughly reviewed and verified.

## REFERENCES


Henrik Aanæs, Rasmus R. Jensen, George Vogiatzis, Engin Tola, and Anders B. Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016.

Yohann Cabon, Lukas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jérôme Revaud, and Vincent Leroy. MUSt3R: Multi-view network for stereo 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Daniela Calvetti and Erkki Somersalo. Inverse problems: From regularization to bayesian inference. *WIREs Computational Statistics*, 10(3):e1427, 2018.

Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of SIGGRAPH*, pp. 303–312, 1996.

Shoichiro Iwase et al. ZeroGrasp: Zero-shot shape reconstruction enabled robotic grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17405–17415, 2025.

Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics*, 32(3):29:1–29:13, 2013.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *PMLR*, pp. 3744–3753, 2019.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with MASt3R. In *Proceedings of the European Conference on Computer Vision*, 2024.

Zhe Liao and Steven L. Waslander. Multi-view 3d object reconstruction and uncertainty modelling with neural shape prior. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3098–3107, 2024.

Helen Oleynikova, Zachary Taylor, Marius Fehr, Roland Siegwart, and Juan Nieto. Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017.

Tim Rohe, Ann-Christine Ehlis, and Uta Noppeney. The neural dynamics of hierarchical bayesian causal inference in multisensory perception. *Nature Communications*, 10:1907, 2019.

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/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

Ming Su, Qian Yu, Zhi Wang, Yang Li, Chunhong Pan, Yiming Chen, Guofeng Wang, and Feng Zhu. GVGS: Gaussian visibility-aware multi-view geometry for accurate surface reconstruction. arXiv preprint arXiv:2601.20331, 2026.

V. K. Vats, S. Joshi, D. J. Crandall, M. A. Reza, and S. Jung. GC-MVSNet: Multi-view, multi-scale, geometrically-consistent multi-view stereo. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3242–3252, 2024.

Oleg Voynov et al. Multi-sensor large-scale dataset for multi-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21392–21403, 2023.

Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. PatchmatchNet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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*, pp. 5294–5306, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zhiwen Yang, Rui Zhang, Jing Shi, Zhaoyu Ai, Bo Zhao, Hujun Bao, Liu 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, pp. 9363–9372, 2025.

Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. In *Proceedings of the British Machine Vision Conference*, 2020.

Zhenyu Zhu, Weiyue Zhao, Ning Chen, Chen Tian, and Zhiwen Fan. Trust it or not: Evidential uncertainty for feed-forward 3d reconstruction with Trust3R. ICML accepted; arXiv:2605.19539, 2026.

# A Diagnostic Label Construction

We construct Where, Who, and Why targets from reference geometry and original observations using a common local-distance-field protocol. Reference geometry is used only for offline supervision and evaluation. Inference uses observations, fused surfaces, and calibration. Each task has its own validity mask; unsupported labels remain unknown rather than becoming negatives.

## A.1 Where: Local Geometric Quality

For fused query $q$, compute nearest distance $e_q$ to reference surface $S_{\text{gt}}$. A fixed 5 mm threshold defines Bad queries, with supervision restricted to valid reference distances:

$$e_q = d(q, S_{\text{gt}}), y_q^W = \mathbb{1}[e_q \geq 5\,\text{mm}].$$

The 5 mm threshold is identical across training, validation, and test splits, giving a consistent definition of local geometric distortion.

## A.2 Why: Multi-label Observation Factors

We project the reference correspondence of each query into input views and compare observed and reference depths after visibility and validity checks. The valid residual is $r_{q,v} = d_{q,v}^{\text{obs}} - d_{q,v}^{\text{ref}}$. Diagnostic thresholds are in millimeters.

Within each $32 \times 32$ pixel neighborhood, the residual median and median absolute deviation estimate location and scale. At least 32 valid samples are required:

$$b_v = median(r_v), \sigma_v = 1.4826 median(|r_v - b_v|).$$

The first three factors use single-view neighborhoods. Multi-view Conflict compares normal residuals at the same reference anchor. Let $t_{q,v}$ be the world-frame ray parameterized by camera depth, and $n_q^{\text{gt}}$ the reference normal. Then $\widetilde{r}_{q,v} = r_{q,v}(t_{q,v}^{\mathsf{T}} n_q^{\text{gt}})$. For at least two valid views, conflict is determined by the residual range:

$$\max_v \widetilde{r}_{q,v} - \min_v \widetilde{r}_{q,v} \geq 5\,\text{mm}.$$

Factors are ordered Bias, Residual Outlier, Noise, and Multi-view Conflict and predicted by independent sigmoids. For a Bad query, positive evidence from any valid observation supports a positive factor label. A negative label requires complete candidate-observation evidence; otherwise it remains unknown. Validity masks and confidence weights govern supervision. Factors are non-exclusive and do not establish physical faults: Bias and Conflict can co-occur.

Table A1: Diagnostic factors and fixed label rules

| Factor | Definition | Level | Rule |
|---|---|---|---|
| Bias | Persistent shift in local residuals | Single view | $\lvert b_v \rvert \geq 3\,\mathrm{mm}$ |
| Residual Outlier | Deviation from the robust local distribution | Single view | $\lvert r - b_v \rvert > max(5\,\mathrm{mm}, 3\sigma_v)$ |
| Noise | Excessive local residual dispersion | Single view | $\sigma_v \geq 3\,\mathrm{mm}$ |
| Multi-view Conflict | Inconsistent normal residuals at a shared reference anchor | Across views | At least two valid views; spread $\geq 5$ mm |

### A.3 WHO: LEAVE-ONE-VIEW-OUT EFFECT TARGETS

Leave-one-view-out (LOO) re-fusion provides continuous targets. At the same query location, we compare reference-relative local distance-field errors for full fusion and each view-removed fusion.

Let $D_S(x)$ denote unsigned distance to geometry $S$. Sample probes $\mathcal{P}(q)$ within a 10 mm ball around $q$, and truncate distances at $\rho = 10\,\mathrm{mm}$. The local field error is defined below. Meshes use triangle distances; point sets use nearest-point distances. Distances are computed in meters and converted to millimeters.

$$E(S, q) = \frac{1000}{|\mathcal{P}(q)|} \sum_{x \in \mathcal{P}(q)} |min(D_S(x), \rho) - min(D_{\text{gt}}(x), \rho)|.$$

Let $E_{\text{all}}(q) = E(S_{\text{all}}, q)$ and $E_{-v}(q) = E(S_{-v}, q)$. Removal effect $\Delta_v(q) = E_{\text{all}}(q) - E_{-v}(q)$ is positive when removal reduces the defined local error. Effects above 0.1 mm retain their magnitude and are normalized into the target distribution:

$$w_v(q) = \Delta_v(q) \mathbb{1}[\Delta_v(q) > 0.1\,\mathrm{mm}], t_{q,v} = \frac{w_v(q)}{\sum_j w_j(q)}.$$

A valid target requires a Bad query, valid LOO effects for every input view, positive total retained effect, and effects representable by the input mask. Seven-view experiments require all seven counterfactuals. Effects are computed on 2 mm and 1 mm grids and must agree within 0.025 mm. Both probe sets must provide reference evidence within 2 mm of reference geometry. Fine-grid values define targets. This numerical check filters unstable labels but does not establish physical causality; queries with incomplete evidence are excluded from Who loss.

## B PNR FEATURES AND NUMERICAL TREATMENT

Each valid observation has an 11-dimensional PNR vector $x_{q,v} \in \mathbb{R}^{11}$, combining relative depth, viewing geometry, and surface orientation. The features retain one-to-one correspondence with query-view pairs for set-based relational modeling.

Let $d_s$ and $d_q$ be observed and camera-frame query depths, with residual $r = d_s - d_q$. We use fixed scale $\ell_0 = 0.010m$ and $\varepsilon = 10^{-6}m$. The three relative-depth quantities are $\frac{r}{\ell_0}$, $r/(d_q + \varepsilon)$, and $\log[max(d_s, \varepsilon)/max(d_q, \varepsilon)]$, expressing discrepancies relative to a physical scale, query depth, and depth ratio. The physical scale is independent of TSDF truncation.

Transform the query and fused-surface normal into camera coordinates $x_c = (x_c, y_c, z_c)$ and $n_c = (n_x^c, n_y^c, n_z^c)$. Together with normalized image coordinates $(u_n, v_n)$ and incidence cosine $\kappa$, they form the vector below. Camera-coordinate division uses $max(z_c, \varepsilon)$ as a safeguard.

$$x_{q,v} = [u_n, v_n, \kappa, \frac{r}{\ell_0}, \frac{r}{d_q + \varepsilon}, \log \frac{max(d_s, \varepsilon)}{max(d_q, \varepsilon)}, \frac{x_c}{max(z_c, \varepsilon)}, \frac{y_c}{max(z_c, \varepsilon)}, n_x^c, n_y^c, n_z^c].$$

Table B1: PNR feature dimensions

| Dim. | Feature | Expression | Role |
|---|---|---|---|
| 1–2 | Normalized image position | $u_n, v_n$ | Relative image location |
| 3 | Incidence cosine | $\kappa$ | Ray–normal relationship |
| 4 | Fixed-scale residual | $r/\ell_0, \ell_0 = 10\,\mathrm{mm}$ | Discrepancy at a physical scale |
| 5 | Query-depth-normalized residual | $r/(d_q + \varepsilon)$ | Discrepancy relative to scene depth |
| 6 | Depth ratio | $\log(d_s/d_q)$ | Scale-independent depth consistency |
| 7–8 | Normalized camera coordinates | $x_c/z_c, y_c/z_c$ | Camera-relative position |
| 9–11 | Camera-frame surface normal | $n_x^c, n_y^c, n_z^c$ | Local surface orientation |

All PNR inputs are available at inference: fused queries, surface normals, observed depth, and calibration. PNR excludes reference geometry, ground-truth distances, labels, and fixed view identifiers. Invalid observations are zeroed. Normalization statistics are fitted on valid training observations and frozen for validation and testing.

# C Implementation and Training

## C.1 Architecture and Boundary Handling

A masked PNR set $X_q$ is linearly embedded into 64-dimensional tokens. Two ISAB layers model view relations; PMA uses one learned seed, followed by rFF. Attention has four heads and 32 inducing points per ISAB. No LayerNorm is used. The rFF has widths $64 \to 128 \to 64$ with ReLU.

View features $H = \{h_v\}$ are permutation-equivariant; global context $z$ is invariant. Who outputs follow view permutations. Attention and pooling use validity masks. All-invalid inputs use safe placeholders to avoid numerical errors, produce zero Who weights, and mark the query unavailable.

Masked softmax over valid views $\mathcal{V}_q$ gives $\alpha_v = \exp(s_v)/\sum_{j\in\mathcal{V}_q}\exp(s_j)$; invalid-view weights are zero. The Who base MLP has widths $64 \to 64 \to 1$; its relational adapter has widths $129 \to 64 \to 1$. Both use GELU.

The Why base MLP has widths $64 \to 64 \to 4$, and its context adapter has widths $258 \to 64 \to 4$, both with GELU. Gated residual addition combines logits before four independent sigmoids.

## C.2 Losses and Supervision Masks

The notation W, A, and R below denotes Where, Who, and Why, respectively.

Where uses binary cross-entropy, Why uses four-label binary cross-entropy, and Who uses cross-entropy with the target view distribution. Operator $\langle\cdot\rangle_{M,C}$ is the weighted normalized mean under validity mask $M$ and confidence $C$. Terms without valid supervision are zero, and tasks have equal weights:

$$\mathcal{L} = \mathcal{L}_W + \mathcal{L}_A + \mathcal{L}_R.$$

$$\mathcal{L}_A = \langle -\sum_v t_{q,v} \log max(\alpha_{q,v}, 10^{-12})\rangle_{M_A,C_A}.$$

Where uses reference-distance validity and unit confidence. Who and Why additionally require available inputs and use their label masks and confidence weights. The current common Who protocol uses unit confidence. Unknown targets are masked, not treated as negatives.

### C.3 Training Protocol

Parallel Base and W3D train from scratch for 30 epochs, without pretrained Base weights or warm-up. The fixed-PNR seven-view TSDF comparison shares inputs, backbone, 71/15/15 scene split, losses, sampler, and optimizer. Seeds are 42, 2023, and 2024. AdamW uses learning rate 0.001, weight decay 0.0001, and gradient-norm clipping at 5. Batch size is 512. Each epoch visits 541,987 training queries without replacement and retains smaller final batches per scene. The earliest epoch achieving maximum validation Where AP is selected. Adapters add parameters, so the comparison controls backbone and training rather than total parameter count.

### C.4 Fusion Geometry and Query Sampling

All four methods in Section 5.5 use the same seven calibrated depth maps, workspace-sphere crop, 0–5 m depth check, and valid central-difference normals. Normals are cross products of horizontal and vertical differences of backprojected points and face the observing camera. TSDF uses Open3D ScalableTSDFVolume with 2 mm voxels, 10 mm truncation, and depth stride 1. Voxel-centroid averages positions and normalizes summed normals within 2 mm voxels. Surfel merges nearest compatible surfels under fixed camera poses, using 4 mm distance and 45° normal-angle thresholds; it is not ElasticFusion. Screened Poisson uses raw oriented depth points, width 2 mm, scale 1.1, linear_fit disabled, one thread, and no density pruning.

Methods share per-scene query budgets, not spatial query locations. Point sets sample deduplicated native centers uniformly without replacement. Meshes sample triangles proportional to area and then continuous barycentric coordinates, excluding duplicate locations. Point sets retain native normals; meshes interpolate vertex normals, falling back to face normals for zero-norm results. SHA-256 hashes of scene and method names determine query seeds, and all training seeds share queries. Sampling uses neither reference geometry nor labels. Point sets use point-mass weights; meshes use surface-area weights.

## D Observation and Supervision Coverage

Each fusion method contains 783,557 queries in the complete constructed dataset. Who-valid counts for TSDF, Voxel-centroid, Surfel, and Screened Poisson are 119,076, 136,425, 138,072, and 43,746. Queries without valid inputs number 83, 0, 44, and 413,133, respectively. These full-dataset counts characterize observability and supervision coverage rather than test performance. Test metrics follow the validity rules in Section 5.1.

## E Supplementary Experimental Results

### E.1 Seed-Level Results

We report the selected checkpoint for every primary configuration and seed (42, 2023, 2024). Scores use the same valid-query masks and within-scene weights as the main tables. All 54 primary checkpoint evaluations reproduce the stored Where and four-factor AP values within $2 \times 10^{-6}$. Shared endpoints are listed once: A0 = W3D/raw7; A4 = W3D/PNR; seven-view TSDF endpoints also serve the view-count and fusion experiments. No checkpoint or score direction is selected using these test results.

Table E1: Selected-checkpoint test results

| Configuration | Seed | Where AP | Where AUROC | Who Top-1 | Who Mass | Why AP | Outlier AP |
|---|---|---|---|---|---|---|---|
| Base / raw7 | 42 | 0.90441 | 0.73779 | 0.38258 | 0.36099 | 0.87862 | 0.62196 |
| Base / raw7 | 2023 | 0.89818 | 0.67376 | 0.41953 | 0.38402 | 0.86508 | 0.54718 |
| Base / raw7 | 2024 | 0.88832 | 0.69136 | 0.39773 | 0.36415 | 0.84294 | 0.49295 |
| Base / PNR | 42 | 0.92160 | 0.81085 | 0.55440 | 0.49871 | 0.88917 | 0.60839 |
| Base / PNR | 2023 | 0.91317 | 0.79280 | 0.52254 | 0.46559 | 0.88610 | 0.62602 |

*Table E1 continued*

| Configuration | Seed | Where AP | Where AUROC | Who Top-1 | Who Mass | Why AP | Outlier AP |
|---|---|---|---|---|---|---|---|
| Base / PNR | 2024 | 0.91998 | 0.80679 | 0.56046 | 0.50257 | 0.89779 | 0.66664 |
| W3D / raw7 | 42 | 0.91436 | 0.73382 | 0.37643 | 0.35522 | 0.90195 | 0.68376 |
| W3D / raw7 | 2023 | 0.87482 | 0.62514 | 0.39453 | 0.36240 | 0.87747 | 0.59732 |
| W3D / raw7 | 2024 | 0.89612 | 0.72038 | 0.37386 | 0.34985 | 0.88406 | 0.61771 |
| W3D / PNR | 42 | 0.92832 | 0.80515 | 0.55715 | 0.49784 | 0.90049 | 0.64875 |
| W3D / PNR | 2023 | 0.92780 | 0.81778 | 0.55294 | 0.49681 | 0.91119 | 0.69138 |
| W3D / PNR | 2024 | 0.92517 | 0.80642 | 0.55929 | 0.50234 | 0.90882 | 0.68370 |
| A1 Relative depth | 42 | 0.92170 | 0.80869 | 0.57848 | 0.51534 | 0.90650 | 0.67165 |
| A1 Relative depth | 2023 | 0.91758 | 0.80353 | 0.55568 | 0.49703 | 0.90152 | 0.65164 |
| A1 Relative depth | 2024 | 0.92285 | 0.80666 | 0.56465 | 0.50737 | 0.89376 | 0.63684 |
| A2 No slots | 42 | 0.91851 | 0.79103 | 0.54383 | 0.48894 | 0.90088 | 0.65563 |
| A2 No slots | 2023 | 0.90823 | 0.77548 | 0.54735 | 0.49206 | 0.90884 | 0.68311 |
| A2 No slots | 2024 | 0.91420 | 0.77812 | 0.55563 | 0.49762 | 0.91284 | 0.69914 |
| A3 + camXY | 42 | 0.91931 | 0.79914 | 0.55276 | 0.49523 | 0.89922 | 0.64637 |
| A3 + camXY | 2023 | 0.91100 | 0.78172 | 0.52492 | 0.46897 | 0.90360 | 0.66128 |
| A3 + camXY | 2024 | 0.92149 | 0.79911 | 0.54364 | 0.48836 | 0.90275 | 0.65782 |
| Voxel-centroid | 42 | 0.90406 | 0.78454 | 0.53134 | 0.48749 | 0.89125 | 0.61485 |
| Voxel-centroid | 2023 | 0.90719 | 0.78814 | 0.51138 | 0.46807 | 0.90473 | 0.66725 |
| Voxel-centroid | 2024 | 0.90872 | 0.79811 | 0.54298 | 0.49295 | 0.88254 | 0.58953 |
| Surfel | 42 | 0.88820 | 0.74305 | 0.49082 | 0.46831 | 0.89918 | 0.63406 |
| Surfel | 2023 | 0.88606 | 0.76021 | 0.51371 | 0.48798 | 0.90034 | 0.63725 |
| Surfel | 2024 | 0.89574 | 0.76860 | 0.50666 | 0.48432 | 0.90083 | 0.64062 |
| Screened Poisson | 42 | 0.97288 | 0.81330 | 0.45093 | 0.41845 | 0.89601 | 0.67120 |
| Screened Poisson | 2023 | 0.97617 | 0.82547 | 0.41080 | 0.38573 | 0.89719 | 0.66124 |
| Screened Poisson | 2024 | 0.97609 | 0.82375 | 0.47793 | 0.44301 | 0.90808 | 0.70239 |

### E.1 Seed-Level Results Continued

Table E2: Selected-checkpoint test results

| Configuration | Seed | Where AP | Where AUROC | Who Top-1 | Who Mass | Why AP | Outlier AP |
|---|---|---|---|---|---|---|---|
| v3 Base / raw7 | 42 | 0.87252 | 0.67339 | 0.65269 | 0.61333 | 0.76081 | 0.29208 |
| v3 Base / raw7 | 2023 | 0.85708 | 0.62490 | 0.74033 | 0.69238 | 0.76832 | 0.29023 |
| v3 Base / raw7 | 2024 | 0.85759 | 0.62539 | 0.69300 | 0.65296 | 0.77407 | 0.32299 |
| v3 W3D / PNR | 42 | 0.90519 | 0.77043 | 0.77369 | 0.72578 | 0.80987 | 0.40476 |
| v3 W3D / PNR | 2023 | 0.91056 | 0.77676 | 0.77917 | 0.72866 | 0.81422 | 0.42711 |
| v3 W3D / PNR | 2024 | 0.90549 | 0.76838 | 0.76779 | 0.71318 | 0.80410 | 0.39402 |
| v5 Base / raw7 | 42 | 0.88973 | 0.72746 | 0.49943 | 0.45753 | 0.81032 | 0.42379 |
| v5 Base / raw7 | 2023 | 0.86854 | 0.66625 | 0.46161 | 0.42165 | 0.81895 | 0.44312 |
| v5 Base / raw7 | 2024 | 0.85989 | 0.62411 | 0.44218 | 0.40243 | 0.81419 | 0.44592 |
| v5 W3D / PNR | 42 | 0.90810 | 0.77760 | 0.58162 | 0.54184 | 0.86167 | 0.54705 |
| v5 W3D / PNR | 2023 | 0.91011 | 0.79118 | 0.57639 | 0.53501 | 0.86683 | 0.57485 |
| v5 W3D / PNR | 2024 | 0.91315 | 0.80274 | 0.58641 | 0.54256 | 0.83509 | 0.44455 |
| v9 Base / raw7 | 42 | 0.92024 | 0.75566 | 0.41581 | 0.38878 | 0.91302 | 0.69134 |
| v9 Base / raw7 | 2023 | 0.92149 | 0.75254 | 0.33637 | 0.31742 | 0.92246 | 0.71696 |
| v9 Base / raw7 | 2024 | 0.91960 | 0.75845 | 0.42396 | 0.39414 | 0.91867 | 0.70477 |
| v9 W3D / PNR | 42 | 0.93804 | 0.81273 | 0.53461 | 0.49292 | 0.94430 | 0.79509 |
| v9 W3D / PNR | 2023 | 0.93161 | 0.80733 | 0.52422 | 0.48371 | 0.93852 | 0.76919 |
| v9 W3D / PNR | 2024 | 0.93595 | 0.81113 | 0.53452 | 0.49287 | 0.92615 | 0.72312 |
| v11 Base / raw7 | 42 | 0.92390 | 0.76004 | 0.42498 | 0.39463 | 0.92960 | 0.74626 |
| v11 Base / raw7 | 2023 | 0.91813 | 0.71858 | 0.38307 | 0.35927 | 0.92270 | 0.71633 |
| v11 Base / raw7 | 2024 | 0.91837 | 0.74432 | 0.31574 | 0.29782 | 0.93076 | 0.75138 |

Table E3: Seven-view geometric quality scores

| Score | Where AP | Where AUROC |
|---|---|---|
| Mean absolute residual | 0.88389 | 0.65274 |
| Residual MAD | 0.87956 | 0.64378 |
| Mean grazing angle | 0.87151 | 0.65175 |
| W3D | 0.92710 ± 0.00169 | 0.80978 ± 0.00696 |

Table E4: View selection and signed local error improvement

| Selector | Who Top-1 | Who Mass | Who-valid Δ (mm) | Complete Bad Δ (mm) |
|---|---|---|---|---|
| Max residual | 0.24133 | 0.21515 | 0.07643 | 0.05308 |
| Residual deviation | 0.39934 | 0.35492 | 0.22586 | 0.16501 |
| Grazing angle | 0.23525 | 0.23136 | 0.11885 | 0.08211 |
| W3D | 0.55646 ± 0.00323 | 0.49900 ± 0.00294 | 0.30832 ± 0.00405 | 0.22616 ± 0.00316 |
| Oracle | 1.00000 | 0.77231 | 0.45566 | 0.35351 |
| Random | 0.18304 | 0.18304 | 0.06613 | 0.04145 |

*Table E2 continued*

| Configuration | Seed | Where AP | Where AUROC | Who Top-1 | Who Mass | Why AP | Outlier AP |
|---|---|---|---|---|---|---|---|
| v11 W3D / PNR | 42 | 0.93058 | 0.81485 | 0.51217 | 0.46685 | 0.94674 | 0.79880 |
| v11 W3D / PNR | 2023 | 0.93431 | 0.80848 | 0.50651 | 0.46340 | 0.94586 | 0.79670 |
| v11 W3D / PNR | 2024 | 0.94097 | 0.82800 | 0.50707 | 0.46135 | 0.94591 | 0.79299 |

### E.2 Geometric Controls and Local Selection Utility

Geometric baselines use inference-time quantities only. Mean absolute observation-to-query depth residual, residual median absolute deviation (MAD), and mean grazing angle rank Where queries. For Who, we select the maximum absolute residual, maximum deviation from the per-query median residual, or maximum grazing angle. Grazing angle is one minus the absolute incidence cosine. Scores are fixed before evaluation; no reference geometry, labels, fitted thresholds, or test-dependent sign reversal constructs these predictions. Aggregates with no valid observation are set to zero for Where; Who supervision requires available views.

Random selection is the exact uniform expectation over valid input views. Oracle selects the largest signed removal effect and serves only as an evaluation reference. Its Who Mass is the maximum target probability and need not equal one. Top-1 uses the main protocol's argmax convention. All selectors share the same queries and area weights.

The Who-valid cohort contains 17,885 queries. The expanded cohort contains 23,496 Bad queries with complete, reference-supported and representable effects, including queries with no beneficial removal. Improvement is full-fusion local field error minus the selected LOO error; negative values retain harmful deletions. W3D's mean improvement remains positive in the expanded cohort (0.22616 mm), compared with 0.16501 mm for residual deviation and 0.04145 mm for random selection. These measurements concern the local field discrepancy defined in Appendix A, not query-distance error or whole-scene repair.

### E.3 Transfer across View Counts with Frozen Weights

We freeze each seven-view W3D checkpoint and its training normalization. The 3-, 5-, 9-, and 11-view tests use the native fusion geometry, observations, queries and labels constructed for that count. We do not delete tokens while retaining seven-view targets. Scene splits remain fixed, and no target-count retraining, calibration or checkpoint selection is performed.

Table E5: Seven-view W3D transferred to other view counts

| Test views | Where AP | Where AUROC | Who Top-1 | Who Mass | Why AP | Outlier AP |
|---|---|---|---|---|---|---|
| 3 | 0.91017 ± 0.00038 | 0.78379 ± 0.00546 | 0.75980 ± 0.00792 | 0.71784 ± 0.00607 | 0.80645 ± 0.00202 | 0.40878 ± 0.01417 |
| 5 | 0.91678 ± 0.00200 | 0.79478 ± 0.00852 | 0.58826 ± 0.00146 | 0.54413 ± 0.00171 | 0.86217 ± 0.00339 | 0.55416 ± 0.01565 |
| 9 | 0.93697 ± 0.00227 | 0.81015 ± 0.00973 | 0.52916 ± 0.00227 | 0.49071 ± 0.00251 | 0.93569 ± 0.00294 | 0.75965 ± 0.01147 |
| 11 | 0.93922 ± 0.00319 | 0.81645 ± 0.01153 | 0.50036 ± 0.00345 | 0.45897 ± 0.00430 | 0.94690 ± 0.00154 | 0.80014 ± 0.00659 |

Table E6: Random and oracle references under each count's targets

| Views | Random Top-1 | Random Mass | Oracle Top-1 | Oracle Mass |
|---|---|---|---|---|
| 3 | 0.38315 | 0.38315 | 1.00000 | 0.88792 |
| 5 | 0.24044 | 0.24044 | 1.00000 | 0.82166 |
| 7 | 0.18304 | 0.18304 | 1.00000 | 0.77231 |
| 9 | 0.14944 | 0.14944 | 1.00000 | 0.78008 |
| 11 | 0.12956 | 0.12956 | 1.00000 | 0.78694 |

Frozen transfer preserves strong diagnostic performance across all tested counts. Relative to count-specific W3D training, the largest absolute difference is 1.38 pp for Who Top-1 and 0.49 pp for Who Mass. This supports cross-count transfer within the tested TSDF setting, but does not establish transfer across fusion methods or arbitrary view configurations.

Random references depend on the number of valid observations at each query, rather than nominal view count alone. Changes in the target distribution also change the oracle Mass ceiling. Accordingly, absolute Who scores across view counts should be interpreted relative to these references.

### E.4 Factor-Level Performance and Class Priors

AP averages scenes containing positives, including all-positive scenes. AUROC averages only scenes containing both classes; undefined terms are excluded rather than replaced by zero. In Tables E7–E8 each cell lists AP on the first line and AUROC on the second, as seed mean ± sample SD. Configuration-specific scene counts and class priors appear in Table E9. All four factors are evaluated independently.

Near-unit Bias and Noise AP reflects the strongly positive-skewed valid labels. On seven-view TSDF, their macro area-weighted positive proportions are 99.55% and 99.9916%; only 6 and 3 of 15 scenes, respectively, contain both classes. Residual Outlier and Conflict each have 15 two-class scenes. Factor-level AUROC therefore characterizes different evaluable scene subsets and should not be read as a shared-sample ranking.

### E.4 Factor-Level Performance Continued

### E.4 Test Supervision Coverage

Counts below are pooled raw test-query counts, while positive percentage is the mean of per-scene area-weighted proportions over valid labels. These quantities summarize different aspects of the data. AP/AUC denotes the number of contributing scenes. Labels and coverage are shared across seeds and architectures at a fixed fusion/view configuration; all seven-view TSDF feature ablations share the same labels.

Table E7: Four-factor AP and AUROC

| Configuration | Bias | Residual Outlier | Noise | Conflict |
|---|---|---|---|---|
| | 0.99700 ± 0.00086 | 0.55403 ± 0.06478 | 0.99996 ± 0.00003 | 0.89788 ± 0.01648 |
| Base / raw7 | 0.65886 ± 0.14440 | 0.60089 ± 0.07112 | 0.68927 ± 0.10678 | 0.52191 ± 0.05550 |
| | 0.99802 ± 0.00053 | 0.63369 ± 0.02987 | 0.99998 ± 0.00001 | 0.93239 ± 0.01629 |
| Base / PNR | 0.81482 ± 0.04130 | 0.67453 ± 0.03051 | 0.81205 ± 0.14579 | 0.66493 ± 0.04939 |
| | 0.99639 ± 0.00097 | 0.63293 ± 0.04519 | 0.99992 ± 0.00007 | 0.92207 ± 0.00547 |
| W3D / raw7 | 0.60874 ± 0.06433 | 0.64766 ± 0.05709 | 0.67823 ± 0.18871 | 0.57943 ± 0.06122 |
| | 0.99678 ± 0.00089 | 0.67461 ± 0.02272 | 0.99998 ± 0.00001 | 0.95596 ± 0.00016 |
| W3D / PNR | 0.82215 ± 0.07712 | 0.71691 ± 0.01985 | 0.87335 ± 0.06405 | 0.77315 ± 0.00563 |
| | 0.99799 ± 0.00036 | 0.65338 ± 0.01747 | 0.99999 ± 0.00001 | 0.95101 ± 0.00970 |
| A1 Relative depth | 0.92279 ± 0.01132 | 0.70183 ± 0.02010 | 0.86916 ± 0.10421 | 0.75380 ± 0.03379 |
| | 0.99814 ± 0.00022 | 0.67930 ± 0.02201 | 0.99998 ± 0.00001 | 0.95267 ± 0.00272 |
| A2 No slots | 0.90085 ± 0.01963 | 0.72655 ± 0.01877 | 0.86949 ± 0.05521 | 0.74533 ± 0.00876 |
| | 0.99846 ± 0.00011 | 0.65516 ± 0.00780 | 0.99999 ± 0.00000 | 0.95382 ± 0.00150 |
| A3 + camXY | 0.87636 ± 0.04419 | 0.71020 ± 0.00263 | 0.92519 ± 0.00370 | 0.74707 ± 0.00291 |
| | 0.99417 ± 0.00039 | 0.62388 ± 0.03964 | 0.99992 ± 0.00002 | 0.95338 ± 0.00618 |
| Voxel-centroid | 0.79021 ± 0.02927 | 0.75328 ± 0.01853 | 0.71784 ± 0.06834 | 0.75811 ± 0.01717 |
| | 0.99649 ± 0.00042 | 0.63731 ± 0.00328 | 0.99994 ± 0.00001 | 0.96673 ± 0.00040 |
| Surfel | 0.84741 ± 0.04624 | 0.77017 ± 0.00052 | 0.83515 ± 0.04037 | 0.77924 ± 0.00452 |
| | 0.98181 ± 0.00998 | 0.67828 ± 0.02147 | 0.99829 ± 0.00017 | 0.94333 ± 0.01900 |
| Screened Poisson | 0.79640 ± 0.05066 | 0.77588 ± 0.01079 | 0.69404 ± 0.08028 | 0.75697 ± 0.03897 |

Table E8: Four-factor AP and AUROC

| Configuration | Bias | Residual Outlier | Noise | Conflict |
|---|---|---|---|---|
| | 0.99026 ± 0.00447 | 0.30177 ± 0.01840 | 0.99860 ± 0.00028 | 0.78032 ± 0.01247 |
| v3 Base / raw7 | 0.59715 ± 0.08876 | 0.52364 ± 0.06319 | 0.55314 ± 0.02668 | 0.55265 ± 0.01618 |
| | 0.98897 ± 0.00095 | 0.40863 ± 0.01688 | 0.99931 ± 0.00049 | 0.84067 ± 0.00592 |
| v3 W3D / PNR | 0.78251 ± 0.04571 | 0.66595 ± 0.01226 | 0.66262 ± 0.13023 | 0.69187 ± 0.01034 |
| | 0.99396 ± 0.00129 | 0.43761 ± 0.01205 | 0.99981 ± 0.00008 | 0.82656 ± 0.01071 |
| v5 Base / raw7 | 0.61740 ± 0.12207 | 0.58712 ± 0.01226 | 0.64139 ± 0.11981 | 0.48926 ± 0.03965 |
| | 0.99416 ± 0.00097 | 0.52215 ± 0.06863 | 0.99951 ± 0.00045 | 0.90231 ± 0.00429 |
| v5 W3D / PNR | 0.80896 ± 0.12158 | 0.61583 ± 0.07579 | 0.64230 ± 0.23208 | 0.67803 ± 0.02129 |
| | 0.99882 ± 0.00008 | 0.70435 ± 0.01281 | 0.99993 ± 0.00003 | 0.96910 ± 0.00639 |
| v9 Base / raw7 | 0.67508 ± 0.06311 | 0.67059 ± 0.02509 | 0.83538 ± 0.06863 | 0.61672 ± 0.03780 |
| | 0.99851 ± 0.00013 | 0.76247 ± 0.03645 | 0.99987 ± 0.00016 | 0.98446 ± 0.00175 |
| v9 W3D / PNR | 0.70326 ± 0.09291 | 0.75738 ± 0.02214 | 0.78242 ± 0.15727 | 0.76793 ± 0.02299 |
| | 0.99885 ± 0.00023 | 0.73799 ± 0.01893 | 0.99991 ± 0.00002 | 0.97399 ± 0.00171 |
| v11 Base / raw7 | 0.72881 ± 0.09485 | 0.68250 ± 0.03575 | 0.72058 ± 0.05757 | 0.58449 ± 0.05379 |
| | 0.99909 ± 0.00038 | 0.79616 ± 0.00294 | 0.99995 ± 0.00003 | 0.98947 ± 0.00165 |
| v11 W3D / PNR | 0.65474 ± 0.08861 | 0.75800 ± 0.01393 | 0.81629 ± 0.15702 | 0.78642 ± 0.04054 |

Table E9: Factor supervision and evaluable scenes

| Cohort | Factor | Valid | Positive | Positive % | AP/AUC |
|---|---|---|---|---|---|
| 7v TSDF | Bias | 60,952 | 60,718 | 99.5514% | 15/6 |
| 7v TSDF | Residual Outlier | 28,626 | 12,897 | 47.4296% | 15/15 |
| 7v TSDF | Noise | 61,433 | 61,428 | 99.9916% | 15/3 |
| 7v TSDF | Conflict | 35,664 | 31,645 | 89.5725% | 15/15 |
| 7v Voxel | Bias | 66,401 | 66,038 | 99.3765% | 15/6 |
| 7v Voxel | Residual Outlier | 34,364 | 12,730 | 37.6525% | 15/15 |
| 7v Voxel | Noise | 66,207 | 66,190 | 99.9739% | 15/6 |
| 7v Voxel | Conflict | 43,503 | 38,755 | 89.3953% | 15/15 |
| 7v Surfel | Bias | 59,576 | 59,269 | 99.3700% | 15/7 |
| 7v Surfel | Residual Outlier | 31,205 | 9,646 | 33.7056% | 15/15 |
| 7v Surfel | Noise | 59,468 | 59,450 | 99.9739% | 15/3 |
| 7v Surfel | Conflict | 39,627 | 36,008 | 90.8205% | 15/14 |

*Table E9 continued*

| Cohort | Factor | Valid | Positive | Positive % | AP/AUC |
|---|---|---|---|---|---|
| 7v Poisson | Bias | 17,323 | 17,051 | 98.2471% | 15/8 |
| 7v Poisson | Residual Outlier | 9,203 | 3,266 | 36.8482% | 15/15 |
| 7v Poisson | Noise | 17,476 | 17,449 | 99.8370% | 15/6 |
| 7v Poisson | Conflict | 10,069 | 8,887 | 87.8968% | 15/14 |
| 3v TSDF | Bias | 57,350 | 56,683 | 98.7743% | 15/8 |
| 3v TSDF | Residual Outlier | 34,537 | 7,218 | 24.3072% | 15/15 |
| 3v TSDF | Noise | 58,426 | 58,348 | 99.8741% | 15/5 |
| 3v TSDF | Conflict | 33,973 | 24,676 | 73.4152% | 15/15 |
| 5v TSDF | Bias | 58,436 | 58,045 | 99.2421% | 15/7 |
| 5v TSDF | Residual Outlier | 31,014 | 10,036 | 35.9753% | 15/15 |
| 5v TSDF | Noise | 59,010 | 58,981 | 99.9571% | 15/4 |
| 5v TSDF | Conflict | 34,864 | 28,463 | 82.3317% | 15/15 |
| 9v TSDF | Bias | 65,064 | 64,988 | 99.8577% | 15/5 |
| 9v TSDF | Residual Outlier | 26,997 | 15,243 | 59.7724% | 15/15 |
| 9v TSDF | Noise | 65,022 | 64,999 | 99.9655% | 15/4 |
| 9v TSDF | Conflict | 38,991 | 37,376 | 96.0576% | 15/15 |
| 11v TSDF | Bias | 64,802 | 64,735 | 99.8798% | 15/3 |
| 11v TSDF | Residual Outlier | 27,560 | 17,050 | 63.9573% | 15/15 |
| 11v TSDF | Noise | 64,916 | 64,894 | 99.9669% | 15/3 |
| 11v TSDF | Conflict | 40,885 | 39,612 | 97.0062% | 15/15 |

### E.5 Scene-Level Results for Seven-View W3D

The machine-readable supplement retains each seed and scene separately, including coverage, factor AP/AUROC and signed removal effects. The scene means above are not independent training replicates; uncertainty in the main tables is calculated over training seeds.

### E.6 Native-Score Transfer and the ETH3D Boundary Case

DTU score semantics. Vis-MVSNet's pairwise log Laplace uncertainty is used directly; PatchMatchNet's captured source-view weight is converted to one minus that weight. Each score is sampled at the projected query and averaged across supporting reference views after rotating the seven references. Higher values rank a view as a proxy for harmful-view relevance. These transformations come from the stored adapter, not test-set selection, and neither native quantity is a removal-effect estimate. The original Vis-MVSNet formulation uses matching uncertainty to guide visibility-aware cost fusion (Zhang et al., 2020).

The common-score subset contains 625 of 1,586 Who-valid queries from all 22 held-out scans. Its area coverage is 39.79%. W3D uses target-domain diagnostic supervision, whereas the two external models use fixed pretrained weights; the comparison neither isolates architecture nor ensures identical training exposure.

Table E11: DTU coverage and Top-1 / Mass by test scan

| Scan | Common / Who-valid | Area coverage | Vis-MVSNet | PatchMatchNet | W3D |
|---|---|---|---|---|---|
| 1 | 35/66 | 55.5% | 0.6801/0.6342 | 0.4083/0.3708 | 0.5651/0.5395 |
| 4 | 14/39 | 34.7% | 0.1821/0.1821 | 0.0162/0.0162 | 0.6322/0.5979 |
| 9 | 39/87 | 46.2% | 0.1162/0.0866 | 0.2306/0.2306 | 0.4100/0.3901 |
| 10 | 16/36 | 57.7% | 0.3381/0.3365 | 0.2809/0.2867 | 0.6390/0.6306 |
| 11 | 47/196 | 22.1% | 0.1935/0.1864 | 0.0231/0.0231 | 0.5444/0.5660 |
| 12 | 54/97 | 58.8% | 0.7357/0.7357 | 0.2689/0.2674 | 0.7123/0.7113 |
| 13 | 37/157 | 26.9% | 0.4462/0.4198 | 0.4831/0.4503 | 0.4595/0.4553 |
| 15 | 12/33 | 37.6% | 0.3013/0.3013 | 0.0000/0.0000 | 0.5216/0.5491 |
| 23 | 10/43 | 22.8% | 0.1637/0.1747 | 0.0174/0.0284 | 0.3342/0.3808 |

*Table E11 continued*

| Scan | Common / Who-valid | Area coverage | Vis-MVSNet | PatchMatchNet | W3D |
|---|---|---|---|---|---|
| 24 | 25/64 | 40.7% | 0.0010/0.0010 | 0.3497/0.3497 | 0.5436/0.5436 |
| 29 | 8/31 | 25.8% | 0.0238/0.0238 | 0.0212/0.0212 | 0.4850/0.4850 |
| 32 | 10/72 | 12.7% | 0.2921/0.3109 | 0.3275/0.3275 | 0.4955/0.4952 |
| 33 | 14/47 | 30.7% | 0.1158/0.0762 | 0.0934/0.1186 | 0.3242/0.3676 |
| 34 | 9/34 | 27.0% | 0.6060/0.6060 | 0.3458/0.3458 | 0.7513/0.7513 |
| 48 | 44/104 | 46.5% | 0.1913/0.1769 | 0.0744/0.0601 | 0.5425/0.5287 |
| 49 | 58/87 | 69.4% | 0.1339/0.1308 | 0.1931/0.1989 | 0.6468/0.6423 |
| 62 | 29/84 | 33.2% | 0.0104/0.0104 | 0.0757/0.0763 | 0.3137/0.3134 |
| 75 | 33/67 | 45.9% | 0.2644/0.2644 | 0.0261/0.0261 | 0.3468/0.3400 |
| 77 | 42/74 | 53.6% | 0.5240/0.4267 | 0.4302/0.3358 | 0.6270/0.5380 |
| 110 | 39/77 | 56.1% | 0.3704/0.3487 | 0.3572/0.3800 | 0.5995/0.5990 |
| 114 | 25/39 | 68.3% | 0.0817/0.0491 | 0.1466/0.1140 | 0.5669/0.5359 |
| 118 | 25/52 | 50.8% | 0.0000/0.0000 | 0.0799/0.0799 | 0.6731/0.6731 |

### E.6 ETH3D Score and Split Audit

Trust3R's native NIW epistemic trace is averaged over valid views, with larger uncertainty assigned to Bad. W3D predicts Bad probability. Both directions remain fixed. W3D is trained on seven DSLR scenes and selected on two DSLR scenes; five rig scenes are reserved for testing. Training/validation queries originate from predicted depths, whereas test queries lie on the frozen reconstructed surface. This setting combines acquisition and query-construction shifts, so the audit does not isolate a single cause.

Labels measure distance to a supported local laser plane with a fixed 5 mm threshold. Unsupported or unstable planes are unknown. Test AP is close to the 76.70% macro Bad prevalence, and both test AUROCs are below 0.5. Thus, neither usage protocol establishes reliable ranking in this setting. Modest training and validation AUROC also rules out describing the issue solely as a sudden test-only collapse. We therefore retain these results as a boundary analysis and do not reverse score directions after observing test outcomes.

Table E10: Per-scene seed means

| Scene | Where AP | Where AUROC | Who Top-1 | Who Mass | Why AP | Outlier AP |
|---|---|---|---|---|---|---|
| golden snail | 0.93213 | 0.86309 | 0.64528 | 0.60295 | 0.81567 | 0.37597 |
| enamel bowl | 0.92138 | 0.77679 | 0.40604 | 0.40604 | 0.96532 | 0.88032 |
| jewelry box | 0.73919 | 0.74929 | 0.60366 | 0.50733 | 0.85313 | 0.41819 |
| ribbed flowerpot | 0.93975 | 0.86205 | 0.69007 | 0.65082 | 0.95574 | 0.89517 |
| pink boot | 0.94945 | 0.80186 | 0.52864 | 0.49229 | 0.80990 | 0.33534 |
| chicken toy | 0.99634 | 0.76580 | 0.30143 | 0.26602 | 0.81452 | 0.35690 |
| plush bear | 0.87288 | 0.75728 | 0.64821 | 0.57465 | 0.95024 | 0.81525 |
| colanders | 0.99328 | 0.81737 | 0.51375 | 0.43557 | 0.99608 | 0.98442 |
| christmas tree stand | 0.66295 | 0.56995 | 0.50211 | 0.44727 | 0.84369 | 0.37682 |
| rubiks cube | 0.99978 | 0.97135 | 0.62132 | 0.52434 | 0.88004 | 0.62293 |
| jin chan | 0.98812 | 0.92109 | 0.58819 | 0.54316 | 0.93816 | 0.82657 |
| yellow coffee mill | 0.99594 | 0.90502 | 0.54816 | 0.49024 | 0.90301 | 0.65526 |
| round box | 0.96818 | 0.87418 | 0.68657 | 0.62776 | 0.98822 | 0.95592 |
| white castle land | 0.95339 | 0.56342 | 0.41884 | 0.36916 | 0.94231 | 0.77781 |
| brown relief pot | 0.99374 | 0.94821 | 0.64465 | 0.54737 | 0.94649 | 0.84228 |

Table E12: ETH3D local-plane proxy evaluation

| Protocol | Method | Where AP | Where AUROC |
|---|---|---|---|
| Fixed pretrained uncertainty | Trust3R | 0.76285 | 0.46907 |
| DSLR diagnostic supervision | W3D | $0.76831 \pm 0.00415$ | $0.48429 \pm 0.00348$ |

Table E13: Selected-checkpoint W3D performance by split

| Split | Scenes | Where AP | Where AUROC |
|---|---|---|---|
| train | 7 | $0.83512 \pm 0.00409$ | $0.56966 \pm 0.00709$ |
| val | 2 | $0.78143 \pm 0.01398$ | $0.56336 \pm 0.02789$ |
| test | 5 | $0.76831 \pm 0.00415$ | $0.48429 \pm 0.00348$ |

Table E14: ETH3D test support and quality ranking

| Scene | Supported | Coverage | Bad % | Trust3R AUROC | W3D AUROC |
|---|---|---|---|---|---|
| delivery area | 1592 | 1.59% | 80.94% | 0.39748 | 0.46105 |
| electro | 795 | 0.80% | 76.93% | 0.50018 | 0.48934 |
| forest | 477 | 0.48% | 74.07% | 0.48924 | 0.48445 |
| playground | 878 | 0.88% | 78.53% | 0.50641 | 0.51275 |
| terrains | 49536 | 49.54% | 73.03% | 0.45203 | 0.47389 |