LEVERAGING GEOMETRIC EVIDENCE ACROSS VIEWS FOR SCENE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

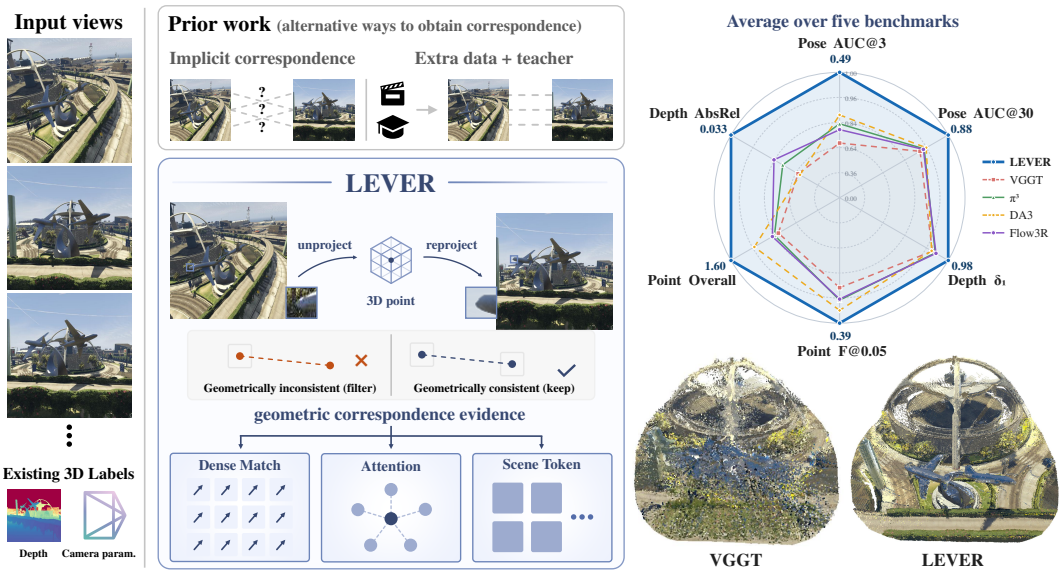


Figure 1: **Learning correspondence for 3D reconstruction.** Prior approaches learn correspondences implicitly or use teacher-generated flow from unlabeled videos. LEVER instead injects cross-view geometric evidence at three levels: output supervision, intermediate feature interactions, and global scene representation, using filtered correspondences from depth and camera annotations. In this extreme long-sequence case, LEVER matches more reliably and avoids inference collapse.

ABSTRACT

Cross-view correspondence provides geometric constraints for multi-view reconstruction by linking observations of the same physical points across views. Recent feed-forward models learn these associations through multi-view feature interactions, often using frame-wise and global attention. However, when matching supervision is confined to a dedicated prediction head, it shapes internal representations only indirectly. Cross-view interactions may therefore remain weakly constrained, limiting how effectively matching information supports reconstruction. We introduce LEVER, which injects cross-view geometric evidence at three levels: output supervision, intermediate feature interactions, and global scene representation. A dense Match Head learns pixel-level cross-view displacement fields, making fine-grained alignment an explicit learning objective. Within the backbone, selected cross-view attention responses are constrained by patch correspondences to encourage information exchange between geometrically related regions. Sequence-level scene tokens aggregate cross-view information under a correspondence-consistency constraint and provide global context to the prediction heads. Together, these components make correspondence a structural signal for reconstruction rather than an auxiliary output. Across five benchmarks, LEVER improves camera-pose estimation, depth estimation, and point-cloud reconstruction over its underlying baseline. Component ablations show that all three levels contribute to reconstruction quality.

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.

1 INTRODUCTION

Recovering scene geometry and camera parameters from images is a fundamental problem in computer vision. Relating observations of the same physical point across views is essential to coherent multi-view reconstruction. Classical structure-from-motion (SfM) makes this relationship explicit by establishing correspondences, estimating camera poses, and refining scene structure. Learning-based feed-forward methods fold these operations into direct prediction. DUSt3R formulates uncalibrated pairwise reconstruction as point-map regression, while MASt3R augments this representation with local descriptors for image matching (Wang et al., 2024b; Leroy et al., 2024). This shift makes cross-view interaction central to learned multi-view reconstruction.

Recent models typically rely on alternating attention for joint multi-view prediction (Wang et al., 2025a; 2026b; Keetha et al., 2026; Lin et al., 2026). Yet when matching is supervised at a dedicated prediction head, its associated loss reaches internal representations only indirectly, leaving correspondence structure to emerge from feature and attention patterns. This indirect supervision may leave cross-view interactions weakly constrained and prevent matching information from being fully exploited for reconstruction. Some prior work makes matching supervision more explicit, but requires an external teacher and additional unlabeled data, incurring teacher-inference and pseudo-label costs and potentially transferring teacher errors to the targets (Cong et al., 2026).

Motivated by this gap, we introduce LEVER, which injects cross-view geometric evidence at three levels: output supervision, intermediate feature interactions, and global scene representation. It derives reliability-filtered correspondence targets by geometrically reprojecting the depth and camera parameters (Yin & Shi, 2018; Zou et al., 2018), and applies this supervision from local prediction to global scene context.

First, a dense Match Head predicts pixel-level displacement fields for fine-grained cross-view alignment. Second, geometric correspondences supervise selected cross-view attention responses, shaping the intermediate features. Third, a correspondence-consistency constraint encourages corresponding patches to have compatible assignments to sequence-level scene tokens. These tokens aggregate information from all views and provide global context to downstream prediction heads.

Together, these components make correspondence a training-time structural signal for reconstruction. The Match Head and its auxiliary supervision are used only during training; at inference, the camera and depth heads provide the reconstruction outputs and the pair-specific Match Head is omitted. Across five benchmarks, LEVER improves reconstruction over its underlying feed-forward baseline, and ablations show that all three levels contribute.

Contributions. Our contributions are:

- We introduce LEVER, a geometry-guided architecture that injects cross-view geometric evidence into feed-forward 3D reconstruction through three complementary levels: explicit dense prediction, intermediate feature interactions, and sequence-level scene representation. To best exploit geometric correspondences, we reproject existing depth and camera annotations and strictly filter invalid, occluded, and dynamic-object inconsistencies, yielding clean cross-view targets without an external flow teacher or additional unlabeled data.
- Across five benchmarks, LEVER improves camera-pose estimation, depth estimation, and point-cloud reconstruction over its underlying feed-forward baseline. Component ablations show the contribution of all three levels.

2 RELATED WORK

2.1 FEED-FORWARD MULTI-VIEW RECONSTRUCTION

Classical structure-from-motion estimates camera poses and scene points from feature correspondences, then jointly refines them through bundle adjustment (Schönberger & Frahm, 2016). VGGSfM retains triangulation and bundle adjustment in a differentiable SfM pipeline (Wang et al., 2024a). DUSt3R instead regresses dense point maps from image pairs, and MASt3R augments this representation with local matching descriptors (Wang et al., 2024b; Leroy et al., 2024).

Feed-forward methods have since developed different ways to combine evidence across multiple views. MUSt3R uses multi-layer memory to reconstruct views in a shared coordinate frame, Fast3R processes more than 1,000 images in one forward pass, and CUT3R maintains a persistent state for online reconstruction (Cabon et al., 2025; Yang et al., 2025; Wang et al., 2025b). VGGT alternates frame-wise and global attention to jointly predict cameras, depth, point maps, and point tracks (Wang et al., 2025a). Other work explores permutation-equivariant prediction, optional geometric inputs, joint depth and ray prediction, and more efficient attention for larger models and training sets (Wang et al., 2026b; Keetha et al., 2026; Lin et al., 2026; Wang et al., 2026a).

These models connect observations across views through learned interactions. In VGGT, analyses find correspondence-like and epipolar patterns in global attention (Sun et al., 2026; Bratulić et al., 2025), suggesting that geometric associations are encoded implicitly. LEVER combines this implicit cross-view learning with explicit correspondence within the network.

2.2 Correspondence Learning for Visual Geometry

Learned matching methods make cross-view correspondence an explicit prediction target. SuperGlue matches sparse local features (Sarlin et al., 2020), whereas LoFTR produces matches without a separate feature detector (Sun et al., 2021). DKM and RoMa estimate dense correspondences across wide baselines (Edstedt et al., 2023; 2024). Optical-flow and stereo methods likewise develop dense matching through recurrent refinement and global association (Teed & Deng, 2020; Xu et al., 2022; Lipson et al., 2021; Xu et al., 2023).

Geometry provides another source of supervision. GeoNet and DF-Net connect depth, camera motion, and optical flow through rigid reprojection and cross-task consistency (Yin & Shi, 2018; Zou et al., 2018). SuperGlue and LoFTR use depth and camera poses to supervise matching between image pairs (Sarlin et al., 2020; Sun et al., 2021); the resulting correspondences support downstream geometric estimation.

Within reconstruction networks, MASt3R learns matching descriptors and VGGT includes a point-tracking head (Leroy et al., 2024; Wang et al., 2025a). Flow3R uses a factored flow head and pseudo-flow from an external 2D teacher to supervise geometry learning, allowing additional unlabeled videos to be used alongside labeled 3D data (Cong et al., 2026). LEVER instead derives filtered correspondence targets from the depth and camera annotations without an external flow teacher or additional unlabeled videos.

2.3 Global Scene Representations

Auxiliary tokens offer additional capacity for aggregating information in Transformers. ViT registers reduce artifacts in patch feature maps (Darcet et al., 2024). VGGT uses frame-associated registers in its alternating attention blocks (Wang et al., 2025a), while VGGT-Ω routes cross-view communication through per-frame registers in selected layers to reduce attention cost (Wang et al., 2026a). NOVA3R aggregates multi-view evidence into a view-agnostic representation for non-pixel-aligned, amodal point-cloud reconstruction (Chen et al., 2026). By contrast, LEVER introduces sequence-level scene tokens with correspondence-consistency supervision, providing shared context to the prediction heads.

3 Method

3.1 Overview

Let $I = \{I_i\}_{i=0}^{S-1}$ denote an unposed sequence of S images. LEVER augments a VGGT-style alternating-attention Transformer (Wang et al., 2025a) with a frozen DINOv3 patch encoder (Siméoni et al., 2025). As illustrated in Fig. 2, the encoder produces patch tokens, which are augmented with a camera token and frame-associated register tokens (Darcet et al., 2024). The resulting tokens pass through alternating frame-wise and global attention, with a shared scene token prefix in the global blocks. Cached backbone features feed the camera, depth, and Match Heads.

During training, calibrated multi-view annotations provide correspondence evidence through geometric reprojection. We use this evidence at three levels: the Match Head explicitly predicts dense

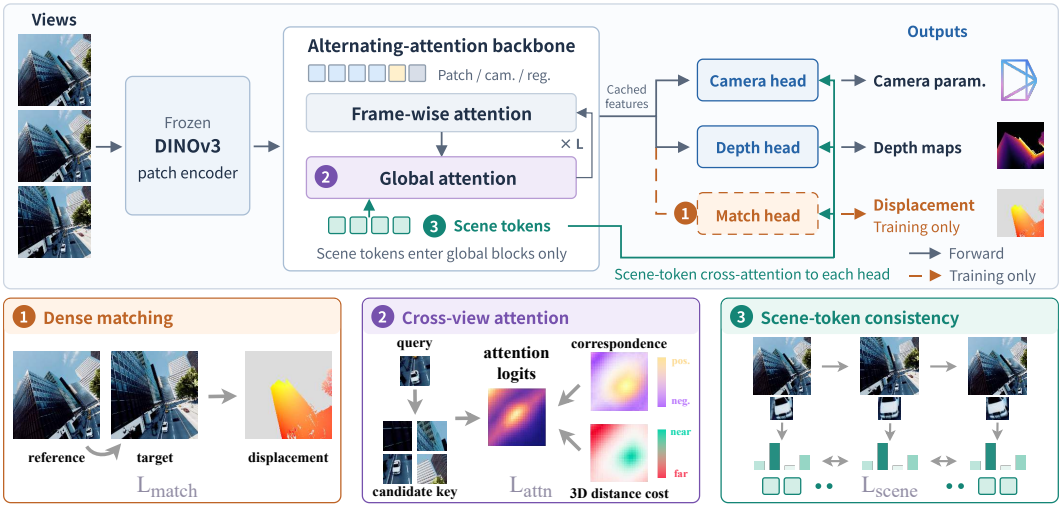


Figure 2: Overview of LEVER. A VGGT-style alternating-attention backbone encodes the complete image sequence. Geometric correspondence evidence supervise dense matching, selected cross-view attention, and a sequence-level scene-token representation. The camera and depth heads provide the reconstruction outputs; matching predictions serve as auxiliary training signals.

cross-view displacement, selected cross-view attention responses are supervised with patch correspondences, and scene tokens are trained to give corresponding patches compatible latent assignments. Native optical flow and stereo disparity, when available, also train the Match Head through the same displacement interface. Correspondence objectives are used during training; reconstruction inference uses the camera and depth outputs without pair-specific matching prediction.

3.2 Mining Cross-View Geometric Evidence

For an ordered view pair (i, j), let D_i and K_i be the reference z-depth and intrinsics, and let E_i^{h} and E_j^{h} be the homogeneous world-to-camera transforms after coordinate normalization. For $\mathbf{p} = (u, v)$, set $\widetilde{\mathbf{p}} = [u, v, 1]^{\mathsf{T}}$ and $\overline{\mathbf{X}}_i(\mathbf{p}) = [D_i(\mathbf{p})K_i^{-1}\widetilde{\mathbf{p}}; 1]$. The depth-induced displacement is

$$F_{i\to j}(\mathbf{p}) = \Pi_j\big(T_{i\to j}\overline{\mathbf{X}}_i(\mathbf{p})\big) - \mathbf{p}, \qquad T_{i\to j} = E_j^{\mathrm{h}}(E_i^{\mathrm{h}})^{-1}, \tag{1}$$

where Π_j applies the target intrinsics and perspective division. This depth-and-motion construction gives rigid flow (Yin & Shi, 2018; Zou et al., 2018). To suppress unreliable matches, we intersect masks for geometric validity, bidirectional consistency, and forward-splat coverage:

$$M_{i\to j} = M_{i\to j}^{\mathrm{geo}} \wedge M_{i\to j}^{\mathrm{bidir}} \wedge M_{i\to j}^{\mathrm{splat}}. \tag{2}$$

The mask controls displacement regression and patch sampling. For each calibrated sequence, we sample a bounded set of ordered view pairs with sufficient valid support for the Match Head and attention supervision. Scene token supervision uses a separate shuffled chain connecting all views. Figure 3 visualizes the filtered targets; Appendix B details the procedures.

3.3 Dense Correspondence Learning

We first use the geometric evidence for explicit dense correspondence prediction. The Match Head predicts two-dimensional displacement from multi-scale features cached by the backbone. Its factorized design follows the factored-flow principle explored by Flow3R (Cong et al., 2026): patch features and camera tokens are projected into geometry and camera branches, while reference-view features are fused with camera context from both views before refinement produces a full-resolution field. The backbone processes the sequence once, after which cached features of selected view pairs are gathered for pair-specific prediction.

The same displacement interface accommodates depth-induced flow and native optical-flow annotations. Stereo disparity is converted to horizontal displacement: taking the left image as the reference

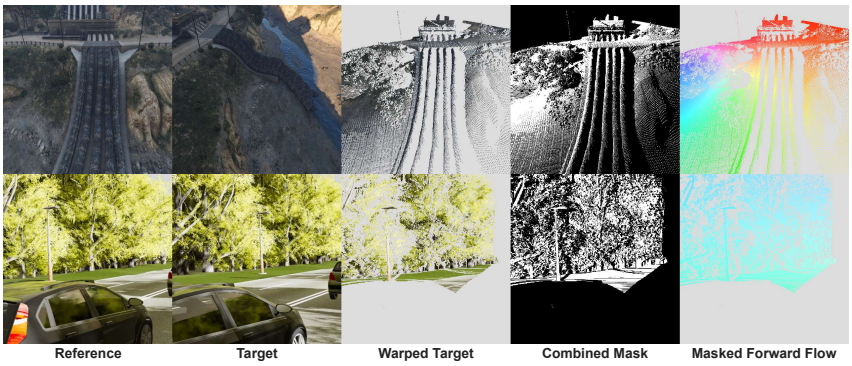


Figure 3: **Qualitative visualization of geometry-derived correspondence supervision.** From left to right, each row shows the reference image, target image, target image warped into the reference view, final combined validity mask, and masked forward geometric flow used for training. Filtering suppresses dynamic-object and occlusion effects to avoid erroneous supervision. Additional examples and details of the construction and filtering procedure are provided in Appendices B.3 and B.

view and the right image as the target view, $d = x_L - x_R$ gives $F_{\text{disp}}(x, y) = (-d(x, y), 0)$. All available matching targets are trained with masked displacement regression.

3.4 Geometry-Supervised Attention

Supervising the Match Head constrains its output, but shapes the backbone's cross-view interactions only indirectly. We therefore supervise sampled cross-view attention responses in a selected global-attention block. A valid pixel from Eq. equation 2 defines a query patch and its projected positive key; additional candidate keys are sampled from the patch pool of the complete sequence while preserving their view identities.

Let Q denote a query token and $\mathcal{K}_Q$ its sampled candidate-key set. We apply a learned projection g_θ to the per-head scaled dot products and normalize the resulting logits to obtain

$$\pi_\theta(K \mid Q) = \text{softmax}_{K \in \mathcal{K}_Q}\left(g_\theta\left(\left\{\frac{(q_Q^{(h)})^\top k_K^{(h)}}{\sqrt{d_h}}\right\}_h\right)\right). \tag{3}$$

With $K^*(Q)$ denoting the geometrically projected positive key, the correspondence term is

$$\mathcal{L}_{\text{corr}} = -\frac{1}{|\mathcal{Q}_{\text{valid}}|}\sum_{Q \in \mathcal{Q}_{\text{valid}}} \log \pi_\theta(K^*(Q) \mid Q)\,. \tag{4}$$

To account for 3D spatial distance, we add an expected distance cost over the sampled key distribution. For a query with a valid 3D point, c_{QK} is the clipped Euclidean distance between normalized world points, divided by the per-query mean distance over valid candidates; invalid candidate keys receive unit cost. With $\mathcal{Q}_{\text{3D}}$ denoting the queries with valid 3D points, the two terms define the attention objective:

$$\begin{aligned}\mathcal{L}_{\text{dist}} &= \frac{1}{|\mathcal{Q}_{\text{3D}}|}\sum_{Q \in \mathcal{Q}_{\text{3D}}}\sum_{K \in \mathcal{K}_Q} \pi_\theta(K \mid Q) c_{QK}, \\ \mathcal{L}_{\text{attn}} &= \mathcal{L}_{\text{corr}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}}.\end{aligned} \tag{5}$$

The correspondence term remains active for every valid query, whereas the distance term requires a valid 3D point. Together, they retain a discrete correspondence target while discouraging probability mass on spatially distant candidates.

3.5 Global Scene Representation

The camera token provides a view-level representation for camera prediction, and the frame-associated register tokens act as auxiliary slots in the aggregator. These registers are tied to individual views. LEVER therefore introduces learnable scene tokens that are replicated for each

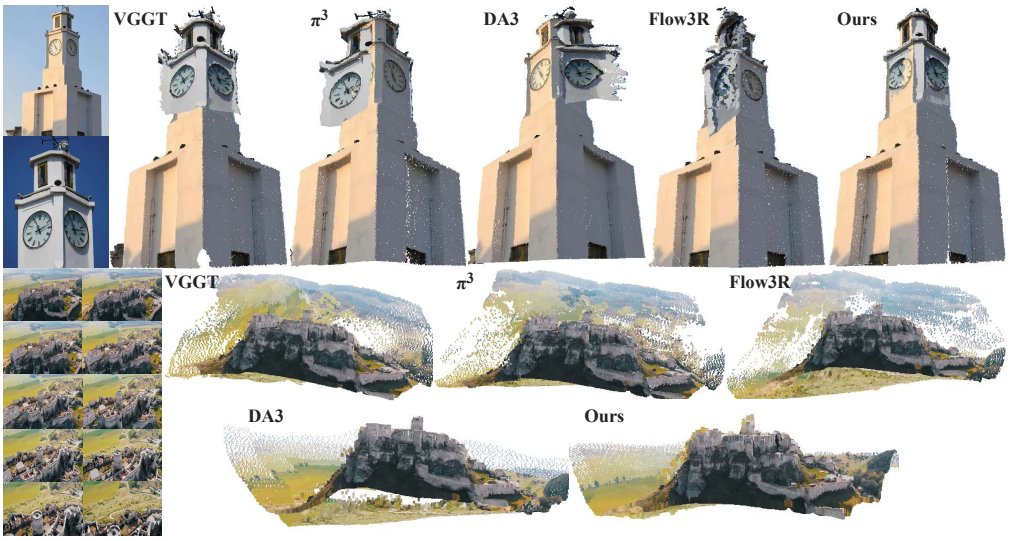


Figure 4: **Qualitative point-cloud reconstruction comparison.** Top: clock-tower views under differing illumination; bottom: an outdoor sequence. Input views are shown on the left, with reconstructions from the compared methods alongside. LEVER yields more coherent visible surfaces in these examples, complementing Table 3.

input sequence and prepended only to global attention. They are excluded from frame-wise attention and are updated across global blocks, allowing them to aggregate information from all views without selecting a reference frame.

Attention supervision constrains sampled patch-level interactions, while correspondence consistency shapes the scene tokens that provide sequence-level context for reconstruction.

At a selected intermediate block, Scene Token Assignment maps the patch representation $\mathbf{h}_{f,p}$ at patch p in frame f to the scene slots. With learned projections ϕ and ψ, the assignment probability for slot m is

$$A_{f,p,m} = \mathrm{softmax}_m \left(\frac{\phi(\mathbf{h}_{f,p})^\top \psi(\mathbf{s}_m)}{\sqrt{d_a}} \right). \tag{6}$$

Rather than prescribing a hard assignment, we encourage geometrically corresponding patches (f, p) and (g, q) to have compatible distributions over scene slots. Their agreement probability is

$$P_{\mathrm{agree}}((f,p),(g,q)) = \sum_m A_{f,p,m} A_{g,q,m}. \tag{7}$$

We minimize the negative log agreement over correspondences sampled from the shuffled-chain graph with edge-balanced averaging. Assignment-entropy and slot-usage regularizers encourage confident assignments while preventing collapse to a single slot; together, these terms define $\mathcal{L}_{\mathrm{scene}}$. Each prediction head receives a learned cross-attention residual from the scene tokens, providing a consistent sequence-level context to the independent camera, depth, and matching decoders.

3.6 Joint Training Objective

The total objective combines reconstruction and correspondence-related losses:

$$\begin{aligned} \mathcal{L}_{\mathrm{total}} = {} & \lambda_{\mathrm{cam}}\mathcal{L}_{\mathrm{cam}} + \lambda_{\mathrm{depth}}\mathcal{L}_{\mathrm{depth}} + \lambda_{\mathrm{point}}\mathcal{L}_{\mathrm{point}} \\ & + \lambda_{\mathrm{match}}\mathcal{L}_{\mathrm{match}} + \lambda_{\mathrm{attn}}\mathcal{L}_{\mathrm{attn}} + \lambda_{\mathrm{scene}}\mathcal{L}_{\mathrm{scene}}. \end{aligned} \tag{8}$$

Here $\mathcal{L}_{\mathrm{match}}$ is a masked displacement loss: geometric reprojection supplies targets when depth, intrinsics, and poses are available, while native flow or disparity supplies them otherwise. The terms $\mathcal{L}_{\mathrm{cam}}$, $\mathcal{L}_{\mathrm{depth}}$, and $\mathcal{L}_{\mathrm{point}}$ are the camera, depth, and unprojected world-point losses; $\mathcal{L}_{\mathrm{attn}}$ and $\mathcal{L}_{\mathrm{scene}}$ are defined above. Implementation-specific weights, sampling budgets, and architectural hyperparameters are deferred to the supplementary material.

Table 1: Camera pose estimation on five benchmarks. AUC@3 and AUC@30 are the areas under the relative-pose accuracy curve up to 3° and 30°; higher is better.

	CO3D		TartanAir		7-Scenes		DTU		TUM-Dynamic	
Method	AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑
CUT3R	0.116	0.615	0.271	0.800	0.040	0.587	0.061	0.726	0.056	0.426
DUSt3R	0.249	0.799	0.379	0.855	0.039	0.602	0.089	0.727	0.028	0.281
Fast3R	0.320	0.826	0.038	0.493	0.067	0.651	0.057	0.722	0.046	0.429
VGGT	0.480	0.890	0.111	0.771	0.168	0.743	0.659	0.956	0.238	0.724
π^3	0.532	0.902	0.646	0.933	0.131	0.761	0.542	0.940	0.168	0.629
Depth Anything 3	0.487	0.881	**0.706**	**0.943**	0.159	0.772	0.590	0.947	0.206	0.656
Flow3R	0.512	0.901	0.582	0.917	0.172	0.791	0.504	0.934	0.151	0.607
LEVER	**0.603**	**0.920**	0.697	0.923	**0.189**	**0.817**	**0.698**	**0.961**	**0.241**	**0.758**

Table 2: Multi-view depth estimation on five benchmarks. We report AbsRel (↓) and threshold accuracy δ_1 (↑).

	CO3D		TartanAir		7-Scenes		DTU		TUM-Dynamic	
Method	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
CUT3R	0.027	0.993	0.095	0.920	0.062	0.953	0.026	0.991	0.112	0.859
DUSt3R	0.041	0.984	0.240	0.720	0.063	0.953	0.023	0.991	0.190	0.810
Fast3R	0.037	0.986	0.287	0.671	0.077	0.934	0.030	0.990	0.194	0.780
VGGT	0.013	0.997	0.076	0.940	0.065	0.942	0.013	0.992	0.103	0.918
π^3	0.012	0.997	0.074	0.929	0.069	0.940	**0.009**	**0.993**	0.052	0.965
Depth Anything 3	0.019	0.993	0.095	0.913	0.062	0.950	0.010	0.992	0.094	0.932
Flow3R	0.013	0.997	0.076	0.927	0.033	0.982	**0.009**	**0.993**	0.067	0.933
LEVER	**0.010**	**0.998**	**0.071**	**0.951**	**0.027**	**0.984**	0.010	**0.993**	**0.049**	**0.968**

Table 3: Point-cloud reconstruction on five benchmarks. Overall is the mean of accuracy and completeness, and F@0.05 is the F-score at a distance threshold of 0.05. Lower Overall and higher F-score are better.

	CO3D		TartanAir		7-Scenes		DTU		TUM-Dynamic	
Method	Overall↓	F@0.05↑	Overall↓	F@0.05↑	Overall↓	F@0.05↑	Overall↓	F@0.05↑	Overall↓	F@0.05↑
CUT3R	3.504	0.033	3.096	0.051	0.296	0.272	6.286	0.713	1.102	0.103
DUSt3R	4.220	0.064	4.614	0.026	0.259	0.327	7.437	0.782	1.124	0.057
Fast3R	4.156	0.085	4.405	0.030	0.232	0.347	6.547	0.718	0.952	0.144
VGGT	3.335	0.101	2.302	0.051	0.159	0.431	3.737	**0.918**	0.353	0.278
π^3	3.946	0.110	1.676	0.166	0.177	0.466	3.175	0.889	0.556	0.238
Depth Anything 3	3.712	0.108	1.603	**0.189**	0.147	0.498	**2.375**	0.896	0.557	0.223
Flow3R	4.059	0.100	1.811	0.134	0.151	0.485	2.899	0.884	0.444	0.260
LEVER	**2.971**	**0.121**	**1.587**	0.182	**0.137**	**0.533**	2.959	0.800	**0.352**	**0.298**

4 EXPERIMENTS

We assess LEVER on camera pose estimation, depth prediction, and point-cloud reconstruction, then examine the contributions of its three components.

4.1 EXPERIMENTAL SETUP

Evaluation benchmarks. We evaluate relative camera pose, depth, and point-cloud reconstruction on CO3D (Reizenstein et al., 2021), TartanAir (Wang et al., 2020), 7-Scenes (Shotton et al., 2013), DTU (Jensen et al., 2014), and TUM-Dynamic (Sturm et al., 2012). We use the CO3D and TartanAir test splits for in-domain evaluation and others for out-of-domain evaluation.

Compared methods. We compare LEVER with CUT3R (Wang et al., 2025b), DUSt3R (Wang et al., 2024b), Fast3R (Yang et al., 2025), VGGT (Wang et al., 2025a), π^3 (Wang et al., 2026b), Depth Anything 3-L (Lin et al., 2026), and Flow3R (Cong et al., 2026), using their respective checkpoints and a common task-specific pipeline. Component effects are examined in the ablations below.

Evaluation metrics. For camera pose, we report AUC@3 and AUC@30, the areas under the relative-pose accuracy curve up to 3° and 30°. For depth, we report absolute relative error (AbsRel) and threshold accuracy δ_1. For point-cloud reconstruction, we report Overall and F-score at a distance threshold of 0.05 (F@0.05).

Table 4: Component ablations using the reduced LEVER reference model. Pose AUC@30, depth AbsRel, and point-cloud F-score are averaged over the evaluation datasets used for the ablations. The reference model retains all three components.

Variant	Pose AUC@30↑	Depth AbsRel↓	Point cloud F@0.05↑
Dense matching			
w/o Geometric flow supervision	0.458	0.081	0.069
w/o Matching supervision	0.378	0.087	0.062
Attention supervision			
w/o Distance term	0.590	0.077	0.095
w/o Attention supervision	0.543	0.079	0.084
Scene representation			
w/o Scene-token supervision	0.616	0.077	0.085
w/o Scene tokens	0.598	0.080	0.085
Ablation reference	**0.644**	**0.072**	**0.102**

Training configurations. Benchmark results use the full LEVER configuration. To limit the cost of repeated training, component ablations use a smaller reference model and a reduced training mixture; their scores are compared only within the ablation.

4.2 Benchmark Results

Camera pose estimation. As shown in Table 1, LEVER leads at both thresholds on four of the five benchmarks. It reaches $0.603/0.920$ on CO3D versus $0.532/0.902$ for π^3, and $0.189/0.817$ on 7-Scenes versus $0.172/0.791$ for Flow3R. It also leads both thresholds on DTU and TUM-Dynamic, while Depth Anything 3-L remains ahead on TartanAir.

Depth estimation. Table 2 shows that LEVER leads on both depth metrics on four benchmarks. On 7-Scenes, it records an AbsRel of 0.027 and δ_1 of 0.984; on DTU, LEVER matches the best δ_1 of 0.993, while its AbsRel of 0.010 remains above the 0.009 of π^3 and Flow3R.

Point-cloud reconstruction. Table 3 shows that LEVER achieves the lowest Overall on four benchmarks and the highest F-score on three. On CO3D, it reaches an Overall of 2.971 and an F-score of 0.121; Depth Anything 3-L retains a slightly higher F-score on TartanAir, while other methods lead both DTU metrics.

Figure 4 shows that, across clock-tower and outdoor sequences with varied appearance, LEVER produces more coherent surfaces and a less fragmented hilltop structure than the alternatives, supporting its integration of corresponding evidence across views.

4.3 Ablation Studies

Table 4 reports component variants trained with the reduced configuration described in Section 4.1. We probe how dense matching, cross-view attention supervision, and global scene representation each contribute to reconstruction; these scores are compared within the reduced-model setting.

Dense correspondence learning. Removing matching supervision reduces pose AUC@30 from 0.644 to 0.378 and point-cloud F-score from 0.102 to 0.062. Retaining native flow and disparity targets while removing geometry-derived flow gives an intermediate AUC@30 of 0.458, indicating that explicit Match Head supervision provides the largest gain.

Geometry-supervised attention. Removing only $\mathcal{L}_{\text{dist}}$ yields 0.590 pose AUC@30 and 0.095 F-score, compared with 0.543 and 0.084 without attention supervision; the full objective reaches 0.644 and 0.102. Thus, the 3D distance term improves correspondence quality beyond selecting a positive key. Across supervision layers 12–24, layer 24 gives the best F-score (0.102) and is used in the main model. Figure 5(a) also shows more focused attention responses with the full loss.

Global scene representation. Scene tokens without correspondence supervision improve pose AUC@30 from 0.598 to 0.616 over removing the tokens, while F-score remains 0.085; adding the

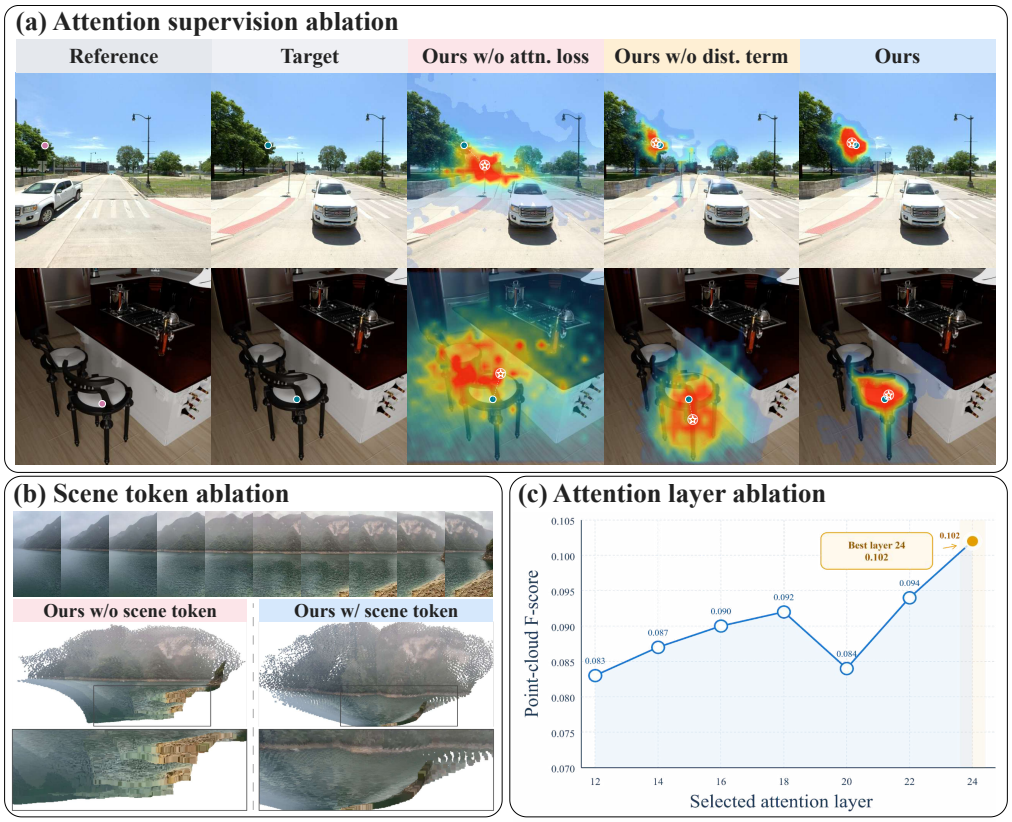


Figure 5: **Qualitative and quantitative ablations of correspondence supervision.** (a) Cross-view attention for reference–target pairs with no attention supervision, no 3D distance term, or the full objective. Magenta dots, cyan dots, and orange stars denote queries, geometric targets, and top-1 attended target keys; dashed lines connect predictions to targets. (b) Scene token ablation on the same sequence; highlighted crops show the shoreline and water surface. (c) Point-cloud F-score for attention supervision at layers 12–24 (step 2); layer 24 is best (0.102) and used in the main model.

supervision raises the reference to 0.644, 0.072 AbsRel, and 0.102 F-score. Figure 5(b) shows a more continuous shoreline and water surface and fewer visible seams in the highlighted region when scene tokens are used, suggesting improved cross-view consistency.

5 Conclusion

LEVER uses cross-view correspondence as geometric supervision for feed-forward 3D reconstruction. It derives reliability-filtered correspondence targets from depth and camera annotations through geometric reprojection and consistency filtering, then applies the resulting evidence at three levels: dense displacement prediction, selected cross-view attention responses, and sequence-level scene representation. This design extends correspondence supervision beyond a separate prediction output to the interactions and representations used for reconstruction.

Across five benchmarks, LEVER improves camera-pose estimation, depth prediction, and point-cloud reconstruction over its underlying feed-forward baseline. Component ablations assess the contributions of the three supervision levels. At inference, reconstruction relies on the camera and depth outputs without a pair-specific matching pass. The method may remain sensitive to severe occlusion and non-rigid motion, motivating future work on more robust cross-view correspondence modeling in challenging scenes.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the manuscript. The authors reviewed and edited the AI-assisted material and take responsibility for the final content of this work, including its text and scientific claims.

REFERENCES

Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *Computer Vision – ECCV*, 2022.

Gilad Baruch, Zhaoyang Chen, Daniel DeTone, et al. ARKitScenes: A diverse real-world dataset for 3d indoor scene understanding using mobile RGB-D data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Jelena Bratulić, Sudhanshu Mittal, Thomas Brox, and Christian Rupprecht. On geometric understanding and learned priors in feed-forward 3d reconstruction models, 2025. URL https://arxiv.org/abs/2512.11508.

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision – ECCV*, 2012.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020.

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome 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 (CVPR)*, pp. 1050–1060, 2025. doi: 10.1109/CVPR52734.2025.00106. URL https://doi.org/10.1109/CVPR52734.2025.00106.

Angel X. Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, and Shuran Song. Matterport3D: Learning from RGB-D data in indoor environments. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2017.

Weirong Chen, Chuanxia Zheng, Ganlin Zhang, Andrea Vedaldi, and Daniel Cremers. NOVA3R: Non-pixel-aligned visual transformer for amodal 3d reconstruction. In *International Conference on Learning Representations (ICLR)*, pp. 98279–98297, 2026.

Zhongxiao Cong, Qitao Zhao, Minsik Jeon, and Shubham Tulsiani. Flow3r: Factored flow prediction for scalable visual geometry learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 438–447, 2026.

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

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, pp. 2632–2652, 2024.

Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17765–17775, 2023. doi: 10.1109/CVPR52729.2023.01704. URL https://doi.org/10.1109/CVPR52729.2023.01704.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19790–19800, 2024. doi: 10.1109/CVPR52733.2024.01871. URL https://doi.org/10.1109/CVPR52733.2024.01871.

Michael Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *Proceedings of the IEEE/CVF Conference on Computer Vision Workshops*, 2019.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Po-Han Huang, Kevin Matzen, Narendra Kopf, Narendra Ahuja, and Jia-Bin Huang. DeepMVS: Learning multi-view stereopsis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Zhaoyang Huang et al. MatrixCity: A large-scale city dataset for city-scale neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3d reconstruction. In *2026 International Conference on 3D Vision (3DV)*, pp. 499–509, 2026. doi: 10.1109/3DV69130.2026.00054. URL https://doi.org/10.1109/3dv69130.2026.00054.

Hoang-An Le, Thomas Mensink, Partha Das, Sezer Karaoglu, and Theo Gevers. EDEN: Multimodal synthetic dataset of enclosed garden scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with MASt3R. In *Computer Vision – ECCV 2024*, pp. 71–91, 2024. doi: 10.1007/978-3-031-73220-1_5. URL https://doi.org/10.1007/978-3-031-73220-1_5.

Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, pp. 141261–141285, 2026. URL https://iclr.cc/virtual/2026/poster/10006531.

Lahav O. Lipson, Zachary Teed, and Jia Deng. RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pp. 218–227, 2021. doi: 10.1109/3DV53792.2021.00032. URL https://doi.org/10.1109/3dv53792.2021.00032.

Kun Luo et al. InStereo2K: A large-scale dataset for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Mapillary. Metropolis dataset. Dataset, 2025. URL https://www.mapillary.com/dataset/metro.

Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4981–4991, 2023. doi: 10.1109/CVPR52729.2023.00482.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Open challenges in deep stereo: the Booster dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19513–19522, 2022. doi: 10.1109/CVPR52688.2022.02049.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4938–4947, 2020. doi: 10.1109/CVPR42600.2020.00499. URL https://doi.org/10.1109/CVPR42600.2020.00499.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016. doi: 10.1109/CVPR.2016.445. URL https://doi.org/10.1109/CVPR.2016.445.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Jurgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8922–8931, 2021. doi: 10.1109/CVPR46437.2021.00881. URL https://doi.org/10.1109/CVPR46437.2021.00881.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Xianbing Sun, Zhikai Zhu, Zhengyu Lou, Bo Yang, Jinyang Tang, Liqing Zhang, He Wang, and Jianfu Zhang. AVGGT: Rethinking global attention for accelerating VGGT. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 251–260, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Sun_AVGGT_Rethinking_Global_Attention_for_Accelerating_VGGT_CVPR_2026_paper.pdf.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020*, pp. 402–419, 2020. doi: 10.1007/978-3-030-58536-5_24. URL https://doi.org/10.1007/978-3-030-58536-5_24.

Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. SMD-Nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. VGGSfM: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21686–21697, 2024a. doi: 10.1109/CVPR52733.2024.02049. URL https://doi.org/10.1109/CVPR52733.2024.02049.

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)*, pp. 5294–5306, 2025a. doi: 10.1109/CVPR52734.2025.00499. URL https://doi.org/10.1109/CVPR52734.2025.00499.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21486–21499, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Wang_VGGT-ohm_CVPR_2026_paper.pdf.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522, 2025b. doi: 10.1109/CVPR52734.2025.00983. URL https://arxiv.org/abs/2501.12387.

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 (CVPR)*, pp. 20697–20709, 2024b. doi: 10.1109/CVPR52733.2024.01956. URL https://doi.org/10.1109/CVPR52733.2024.01956.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual SLAM. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

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. In *International Conference on Learning Representations (ICLR)*, pp. 10481–10497, 2026b.

Cho-Ying Wu, Jialiang Wang, Michael Hall, Ulrich Neumann, and Shuochen Su. Toward practical monocular indoor depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3804–3814, 2022. doi: 10.1109/CVPR52688.2022.00379.

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

Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21919–21928, 2023. doi: 10.1109/CVPR52729.2023.02099. URL https://doi.org/10.1109/CVPR52729.2023.02099.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. GMFlow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8121–8130, 2022. doi: 10.1109/CVPR52688.2022.00795. URL https://doi.org/10.1109/CVPR52688.2022.00795.

Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5510–5519, 2019. doi: 10.1109/CVPR.2019.00566.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935, 2025. doi: 10.1109/CVPR52734.2025.02042. URL https://doi.org/10.1109/CVPR52734.2025.02042.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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.

Zhichao Yin and Jianping Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1983–1992, 2018. doi: 10.1109/CVPR.2018.00212. URL https://doi.org/10.1109/CVPR.2018.00212.

Amir Roshan Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yuliang Zou, Zelun Luo, and Jia-Bin Huang. DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Computer Vision – ECCV 2018*, pp. 38–55, 2018. doi: 10.1007/978-3-030-01228-1_3. URL https://doi.org/10.1007/978-3-030-01228-1_3.

APPENDIX

A IMPLEMENTATION DETAILS

A.1 IMPLEMENTATION DETAIL

Following VGGT (Wang et al., 2025a), LEVER uses an alternating-attention transformer that exchanges information within each frame and across the input views. Our main model has approximately 1.2B parameters. Camera and frame-register tokens provide view-level outputs, while 16 learnable scene tokens aggregate sequence-level context for the camera, dense, and matching heads. The attention-based matching supervision is applied at the 24th transformer layer.

Training uses 512×512 inputs and 2–24 views per sample. We augment each sequence with shared geometric perturbations, randomly varying the aspect ratio within $[0.33, 1.33]$ while keeping the image area approximately 512×512, and apply scale/color jitter and occasional grayscale conversion. We use AdamW with learning rate $1\mathrm{e}{-4}$, weight decay 0.05, a linear warm-up followed by cosine decay, and gradient clipping at 1.0. Training runs for 200K iterations on 128 96-GB H20 GPUs with bfloat16 mixed precision and gradient checkpointing. All experiments use random seed 42. The camera, depth, point, flow, sampled-attention, and scene-token losses use weights 5.0, 1.0, 0.5, 1.0, 0.01, and 0.005, respectively.

A.2 ABLATION REFERENCE CONFIGURATION

The ablation model uses approximately $0.6\times$ the parameter count of the main model. Each variant changes one correspondence-related component while keeping the remaining recipe fixed.

B GEOMETRIC CORRESPONDENCE CONSTRUCTION

We construct dense correspondence targets online for calibrated training sequences. Depths and camera poses are first expressed in a common sequence coordinate frame and normalized by a scene scale. For an ordered reference-target pair (i, j), the image-space target is then obtained by back-projecting the reference depth and reprojecting the points with the relative camera transform.

B.1 PROJECTION

For pixel $p = (u, v)$ with homogeneous coordinate $\tilde{p}$, the reference point and its target projection are

$$X_i(p) = D_i(p) K_i^{-1} \tilde{p}, \tag{9}$$

$$X_j(p) = R_{ij} X_i(p) + t_{ij}, \qquad p_j = \pi(K_j X_j(p)), \tag{10}$$

where K_i, K_j are intrinsics and (R_{ij}, t_{ij}) is the relative pose. The forward geometric flow is

$$F_{i\to j}(p) = p_j - (u, v). \tag{11}$$

The reverse field $F_{j\to i}$ is computed in the same way.

B.2 FILTERING AND PAIR SELECTION

We form three masks on the reference-view pixel lattice. The geometric-validity mask $M^{\mathrm{geo}}_{i\to j}$ keeps pixels with positive source depth, positive depth after transformation to the target camera, and projected coordinates inside the target image.

The bidirectional-consistency mask samples the reverse field at the forward-warped location. With $e(p)$ denoting the squared round-trip error, $m(p)$ the sum of the forward and reverse squared magnitudes, and $(\alpha, \beta) = (10^{-5}, 0.01)$, we use

$$M^{\mathrm{bidir}}_{i\to j}(p) = \mathbf{1}[e(p) \le \alpha m(p) + \beta], \tag{12}$$

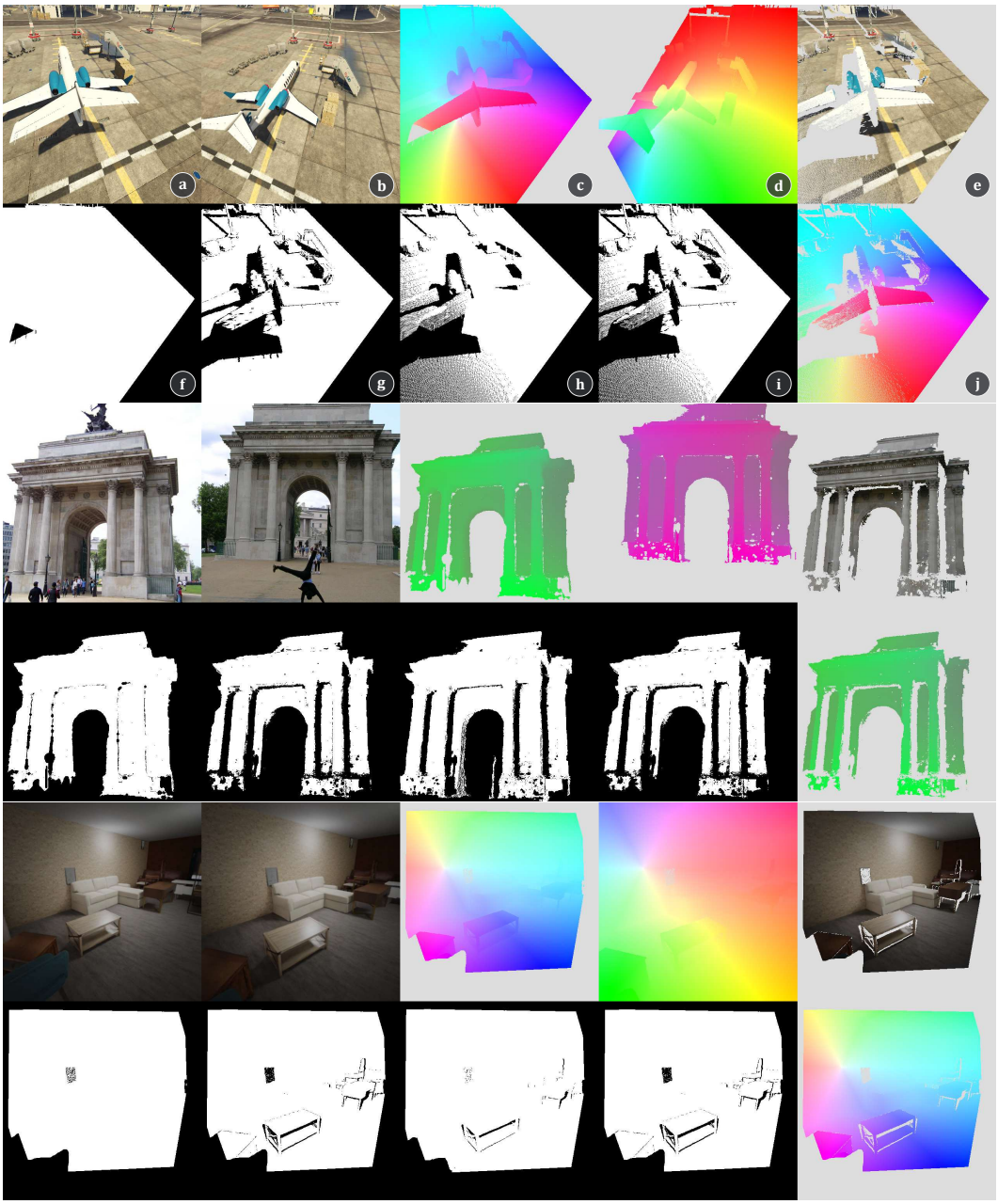


Figure 6: **Generation and filtering process for correspondence supervision.** Each example contains (a) the reference image, (b) the target image, (c) forward geometric flow, (d) backward geometric flow, (e) the target image warped into the reference view, (f) geometric validity, (g) bidirectional consistency, (h) forward-splat coverage, (i) the combined validity mask, and (j) the masked forward flow retained for supervision. Geometric validity removes invalid projections, while bidirectional consistency and splat coverage suppress inconsistent, occluded, and uncovered correspondences. Their intersection retains reliable dense supervision. Gray regions denote invalid or filtered pixels.

Finally, the forward-splat mask uses the reverse field to map valid target pixels back to the reference view. Bilinear contributions form a coverage map $C_{j\to i}$, and the mask retains pixels whose coverage exceeds $\tau = 0.3$:

$$M_{i\to j}^{\text{splat}}(p) = \mathbf{1}[C_{j\to i}(p) \geq \tau]\,. \tag{13}$$

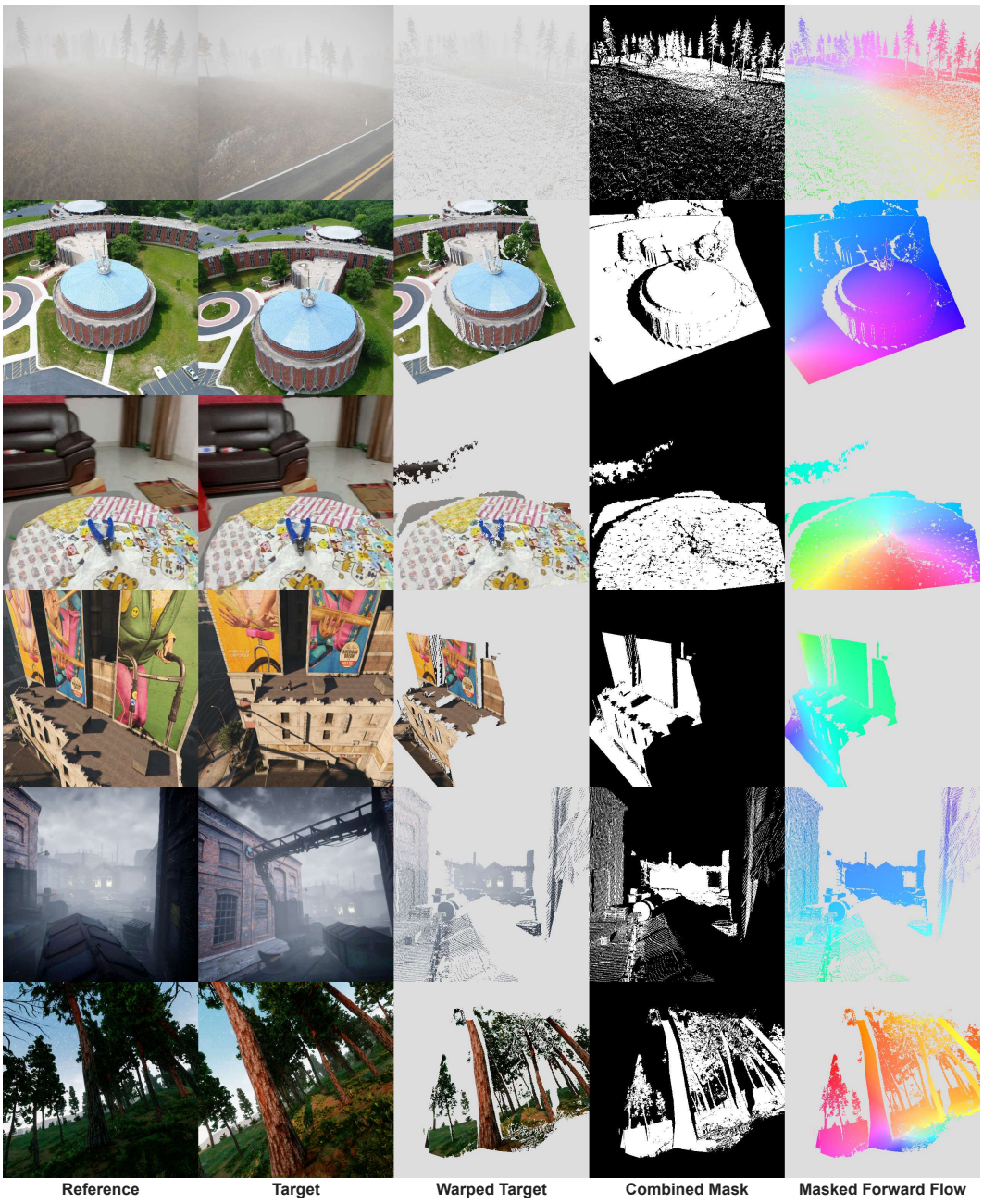


Figure 7: **Additional qualitative examples of geometry-derived correspondence supervision.** From left to right, each row shows the reference image, target image, warped target, combined validity mask, and masked forward geometric flow, complementing the examples in the main paper.

The point-validity mask is used only to select splatting sources; this test checks coverage and does not perform depth ordering or a z-buffer. The final supervision mask is the intersection of the three masks:

$$M_{i\to j} = M^{\text{geo}}_{i\to j} \wedge M^{\text{bidir}}_{i\to j} \wedge M^{\text{splat}}_{i\to j}. \tag{14}$$

For each depth batch, ordered non-self pairs are randomly permuted. We select up to eight pairs, accept a candidate only when at least 10% of pixels are valid. Scene-token supervision uses a separately shuffled chain of sequence edges.

B.3 Qualitative Visualization

Figure 6 traces the target-construction pipeline for a representative calibrated pair. The visualization follows the forward and backward geometric fields through geometric-validity, bidirectional-consistency, and splat-coverage filtering, and ends with the masked flow used for supervision.

Figure 7 shows additional examples across different scenes. The alignment between the warped target, combined validity mask, and retained flow illustrates how the filters remove unsupported or inconsistent pixels while preserving reliable correspondences.

C Datasets and Evaluation Protocols

C.1 Training Data Mixture

The full model is trained with three supervised data pools. Depth and point supervision uses Virtual KITTI 2 (Cabon et al., 2020), TartanAir V1 and V2 (Wang et al., 2020), MVS-Synth/GTAV (Huang et al., 2018), CO3D (Reizenstein et al., 2021), WildRGB-D (Xia et al., 2024), UnrealStereo4K (Tosi et al., 2021), MapFree (Arnold et al., 2022), Waymo (Sun et al., 2020), ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), ARKitScenes (Baruch et al., 2021), Hypersim (Roberts et al., 2021), BlendedMVS (Yao et al., 2020), MatrixCity (Huang et al., 2023), Spring (Mehl et al., 2023), Matterport3D (Chang et al., 2017), MegaDepth (Li & Snavely, 2018), EDEN (Le et al., 2021), Taskonomy (Zamir et al., 2018), Aria Synthetic Environments (Pan et al., 2023), Mid-Air (Fonder & Van Droogenbroeck, 2019), Mapillary Metropolis (Mapillary, 2025), and GTA-SfM (Wang et al., 2024a).

Disparity supervision uses Virtual KITTI 2 (Cabon et al., 2020), TartanAir V1 and V2 (Wang et al., 2020), UnrealStereo4K (Tosi et al., 2021), KITTI 2012 and KITTI 2015 (Geiger et al., 2012), Virtual Apartment (VA) (Wu et al., 2022), Booster (Ramirez et al., 2022), CARLA-Highres (Yang et al., 2019), InStereo2K (Luo et al., 2020), Sintel (Butler et al., 2012), Hypersim (Roberts et al., 2021), BlendedMVS (Yao et al., 2020), Spring (Mehl et al., 2023), EDEN (Le et al., 2021), and Mid-Air (Fonder & Van Droogenbroeck, 2019).

Optical-flow supervision uses TartanAir V1 and V2 (Wang et al., 2020), KITTI 2012 and KITTI 2015 (Geiger et al., 2012), Sintel (Butler et al., 2012), and Spring (Mehl et al., 2023).

Component ablations use a reduced mixture of TartanAir V1 (Wang et al., 2020), CO3D (Reizenstein et al., 2021), UnrealStereo4K (Tosi et al., 2021), TartanAir V2 (Wang et al., 2020), Spring (Mehl et al., 2023), EDEN (Le et al., 2021), and Mid-Air (Fonder & Van Droogenbroeck, 2019).

C.2 Evaluation Splits and Windows

We evaluate on CO3D (Reizenstein et al., 2021), TartanAir V1 (Wang et al., 2020), 7-Scenes (Shotton et al., 2013), DTU (Jensen et al., 2014), and TUM-Dynamic (Sturm et al., 2012). CO3D uses its test split; sequences contain at least 50 images, ten views are sampled with seed 42, and bundle adjustment is disabled. TartanAir V1 uses its test split with eight-frame windows, interval 2, stride 50, and at most ten windows. For 7-Scenes and TUM-Dynamic, we sample one ten-frame clip per sequence with a temporal interval chosen from $\{5, 10, 20, 40\}$. For DTU, we use the 22 evaluation scans and all available views, following prior works (Lin et al., 2026). CO3D and TartanAir V1 are used for in-domain evaluation, while 7-Scenes, DTU, and TUM-Dynamic provide out-of-domain evaluation.

C.3 Metrics

We follow the standard protocols used in the main paper. Camera pose is measured by AUC@3 and AUC@30, computed from the maximum of the relative rotation and translation-direction errors. Depth is evaluated after per-view median scaling with AbsRel and δ_1, where δ_1 is the fraction of valid pixels whose depth error is within a factor of 1.25. For point-cloud reconstruction, predicted points are aligned to the ground truth with a similarity transform and evaluated by Overall, the mean of accuracy and completeness, and F-score at a distance threshold of 0.05. Higher AUC, δ_1, and F-score, and lower AbsRel and Overall indicate better performance.

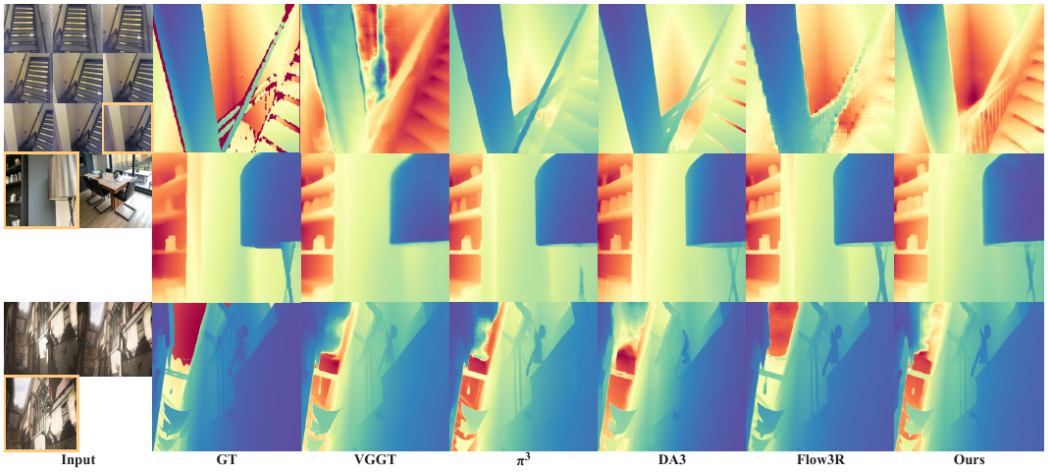


Figure 8: Qualitative depth visualization across the compared methods.

D Qualitative Depth Comparison

We provide supplementary qualitative depth visualizations comparing LEVER with VGGT, π^3, Depth Anything 3, and Flow3R under the same multi-view inputs.

E Limitations

Rigid-scene and visibility assumptions. The geometric target is based on a rigid relative-pose approximation. The geometric-validity, bidirectional-consistency, and splat-coverage masks qualitatively filter out most unreliable correspondences across our experiments. Nevertheless, a small number of erroneous correspondences may remain in extreme scenes, such as those with fast motion, repeated texture, thin structures, or severe occlusion.

Ablation scope. Given the substantial cost of full-scale retraining, we use an ablation model with approximately $0.6\times$ the parameters of the primary model to target competitive performance while isolating component changes.