OLD-SCHOOL WARPING SHINES FOR MODERN SELF-SUPERVISED 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans can develop a rich understanding of the 3D world with little to no explicit geometric supervision. Inspired by this ability, we show that a simple multi-view Transformer can learn strong 3D reconstruction capabilities through self-supervision with an "old-school" image warping loss. Our model, **Warp3R**, jointly predicts depth maps and camera poses and is trained using photometric and geometric consistency across views. It substantially outperforms state-of-the-art self-supervised methods on camera pose estimation across diverse datasets and improves self-supervised depth estimation. Extending supervised training with our objective on unlabeled data further improves both pose and depth estimation. Our results highlight the effectiveness of classical geometric principles for learning 3D reconstruction with modern networks. Code will be released to facilitate future research.

1 INTRODUCTION

Humans can develop a rich understanding of the 3D world with little to no explicit supervision of its geometry (Gibson & Walk, 1960; Kellman, 1984). They can draw on depth cues available in visual observations (Landy et al., 1995). This ability has long motivated a fundamental question: *how can machines learn 3D reconstruction from visual observations without explicit geometric labels?*

Classical 3D reconstruction methods demonstrate that geometry can be recovered from visual observations without explicit 3D labels (Hartley & Zisserman, 2004). Systems such as COLMAP recover camera poses and scene geometry through image correspondences and multi-view consistency (Schönberger & Frahm, 2016; Schönberger et al., 2016). These systems optimize geometry for each scene, but do not learn a reconstruction model across scenes.

Modern supervised learning methods have substantially advanced 3D reconstruction, with models such as DUSt3R (Wang et al., 2024), MASt3R (Leroy et al., 2024), and VGGT (Wang et al., 2025) predicting geometry from multiple images. Yet their reliance on large-scale 3D supervision contrasts with the human ability to perceive 3D structure without such labels. Recent methods, including E-RayZer (Zhao et al., 2026) and NAS3R (Huang et al., 2026), pursue self-supervised reconstruction by jointly learning geometry and appearance representations for differentiable view synthesis.

In this work, we introduce **Warp3R** (Figure 1) and demonstrate that "old-school" image warping provides powerful supervision for modern feed-forward 3D reconstruction. Our training objective is rooted in classical image-warping methods (Lucas & Kanade, 1981; Kanade & Okutomi, 1994; Scharstein & Szeliski, 2002; Alismail et al., 2016) and neural self-supervised 3D reconstruction models (Zhou et al., 2017; Godard et al., 2019; Bian et al., 2019). Through warping, the objective enforces photometric consistency across observed images and geometric consistency across predicted depth maps, without requiring additional appearance representations for view synthesis.

Conceptually, our method is designed as a self-supervised counterpart to VGGT (Wang et al., 2025). Our feed-forward network jointly predicts depth maps and camera poses across views, in contrast to pioneering self-supervised methods that make predictions for individual views or image pairs (Zhou et al., 2017; Godard et al., 2019). The warping losses are computed between pairs of views and aggregated within each scene. Our approach thus combines the multi-view reasoning capabilities of modern feed-forward networks with the simplicity of classical warping-based supervision.

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.

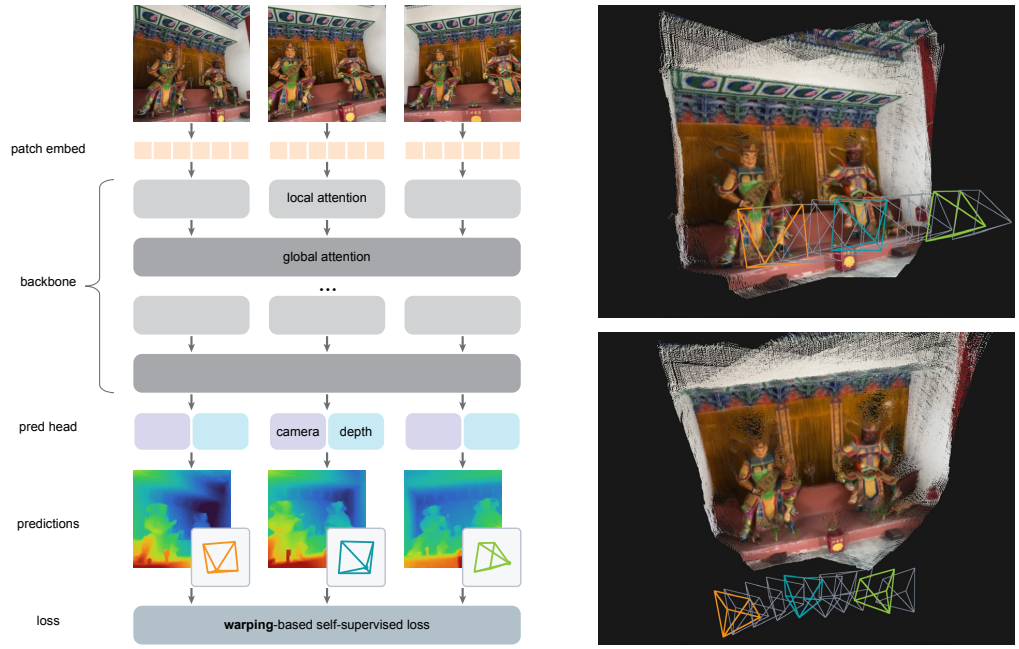

Figure 1: **Warp3R: self-supervised multi-view 3D reconstruction. Left:** A VGGT-inspired feed-forward Transformer with alternating local and global attention jointly processes multiple images, while lightweight heads predict depth maps and camera poses. Warping across views provides photometric and geometric supervision without explicit 3D labels. **Right:** Example 3D reconstructions and camera poses predicted by our model.

Our simple Warp3R model achieves strong results in self-supervised 3D reconstruction. It significantly outperforms the state-of-the-art methods, E-RayZer (Zhao et al., 2026) and NAS3R (Huang et al., 2026), on camera pose estimation across a wide range of datasets, including zero-shot evaluations. We also observe consistent improvements in self-supervised depth estimation. Moreover, we show that extending supervised training with our self-supervised objective on unlabeled data can improve both pose and depth estimation over purely supervised training. These results bring us closer to learning 3D reconstruction from visual observations themselves, a capability that humans naturally acquire as they learn to understand the 3D world.

2 Related Work

Self-supervised 3D Reconstruction. Classical image alignment (Lucas & Kanade, 1981) and photometric bundle adjustment (Alismail et al., 2016) establish image-space residuals as a basis for estimating motion and geometry. SfMLearner (Zhou et al., 2017) uses differentiable view synthesis to learn depth and ego-motion from video without geometric labels. Monodepth and Monodepth2 (Godard et al., 2017; 2019) develop photometric reconstruction losses and practical treatments of occlusion and invalid supervision, while SC-SfMLearner (Bian et al., 2019) couples depth predictions through scale-invariant geometric consistency. Visibility masking from predicted depth also has precedent (Schellevis, 2019). Our work revisits these established objectives for training a modern multi-view feed-forward reconstruction network. In fact, we show that warping-based self-supervision is robust to a wide variety of design choices.

Feed-forward Reconstruction. DUSt3R (Wang et al., 2024) and MASt3R (Leroy et al., 2024) perform feed-forward pairwise reconstruction, followed by optimization to align predictions across multiple views. VGGT (Wang et al., 2025) and subsequent developments (Wang et al., 2026b;a) enable feed-forward joint multi-view geometry prediction. Our architecture draws on this literature. Recent work also explores self-supervised training of such models. GPA-VGGT (Xu et al., 2026) uses photometric and geometric consistency to adapt a VGGT model pretrained with 3D supervision, while SS3D (Hariat et al., 2026) learns depth, pose, and intrinsics from web video through a staged

curriculum and self-supervised expert distillation. Our work studies a simple warping-based training objective without geometric labels or expert distillation.

Self-supervision via Novel-view Synthesis. RayZer (Jiang et al., 2025) learns reconstruction through a learned view-synthesis renderer, while E-RayZer (Zhao et al., 2026) and NAS3R (Huang et al., 2026) use explicit Gaussian Splatting representations (Kerbl et al., 2023) and differentiable rendering. SelfSplat (Kang et al., 2025) combines photometric reprojection with a rendering loss. These methods use additional scene-representation sub-networks to enable novel-view synthesis and construct the self-supervised loss. Our method instead trains depth and camera prediction directly through self-supervised image warping.

Self-supervised Visual Representation Learning. Visual self-supervision learns representations through contrastive learning (He et al., 2020; Chen et al., 2020), self-distillation (Caron et al., 2021), and masked autoencoding (He et al., 2022). CroCo (Weinzaepfel et al., 2022) extends masked autoencoding to multiple views by learning cross-view completion. Our method is related to multi-view masked autoencoding in its use of cross-view reconstruction as a learning signal. Rather than explicitly masking input patches, we constrain reconstruction through geometric warping, encouraging the network to learn depth and camera parameters that explain the observed views.

3 METHOD

Warp3R jointly predicts depth maps and cameras with a multi-view Transformer, using warping-based self-supervision to enforce photometric and geometric consistency across views (Figure 1).

3.1 FORMULATION

Given N RGB images $I_1, \ldots, I_N \in \mathbb{R}^{H\times W\times 3}$ of a single rigid scene, we jointly predict an $H \times W$ depth map z_i and camera parameters $c_i = (R_i, \tau_i, f_i)$ for each view in one feed-forward pass. Here, $R_i \in SO(3)$ maps camera coordinates to a shared world coordinate system, $\tau_i \in \mathbb{R}^3$ is the camera center in that system, and f_i is a scalar focal length.

Let $K_i \in \mathbb{R}^{3\times 3}$ denote the camera intrinsics matrix, which depends on f_i. For a pixel $p = (u, v)$, with homogeneous coordinates $\bar{p} = (u, v, 1)^\top$, the predicted depth $z_i(p)$ specifies its camera-space z-coordinate. Following the standard pinhole camera model (Hartley & Zisserman, 2004), we unproject the pixel into a 3D point $X_i(p) \in \mathbb{R}^3$ in the shared world coordinate system:

$$X_i(p) = R_i\big(z_i(p)K_i^{-1}\bar{p}\big) + \tau_i. \tag{1}$$

Here, the camera parameters (R_i, τ_i, K_i) and depth map z_i are all obtained from the network predictions. Appendix A.2 provides the camera parameterization details.

Gauge invariance. Unlike VGGT (Wang et al., 2025), which uses the first camera to define a canonical coordinate system, our self-supervised objective depends only on *relative* camera transformations, so we do not need to choose a canonical view or coordinate system. We also impose no canonical scale, reflecting the gauge freedom of 3D reconstruction (Triggs et al., 2000).

3.2 NETWORK ARCHITECTURE

Backbone. We use a multi-view Transformer to jointly encode N input images (Figure 1). The backbone alternates between *local attention* (*i.e.*, within each view) and *global attention* (*i.e.*, across all views), with all blocks initialized from a self-supervised model (*e.g.*, DINOv2 Oquab et al., 2024). Unlike VGGT (Wang et al., 2025), which appends newly initialized blocks to the pretrained encoder, we reuse the pretrained Transformer blocks without adding new blocks. We find that our simpler design is sufficient to achieve strong results.

Camera and depth prediction. A learnable token is attached to each view for camera prediction (Wang et al., 2025) and decoded into camera parameters by a small MLP. A depth map per view is predicted by an unpatchifying head that decodes each patch token into a patch of depth values and assembles these patches into a full-resolution depth map. These heads are lightweight by design, placing most of the model's capacity in the backbone. See Appendix A for details.

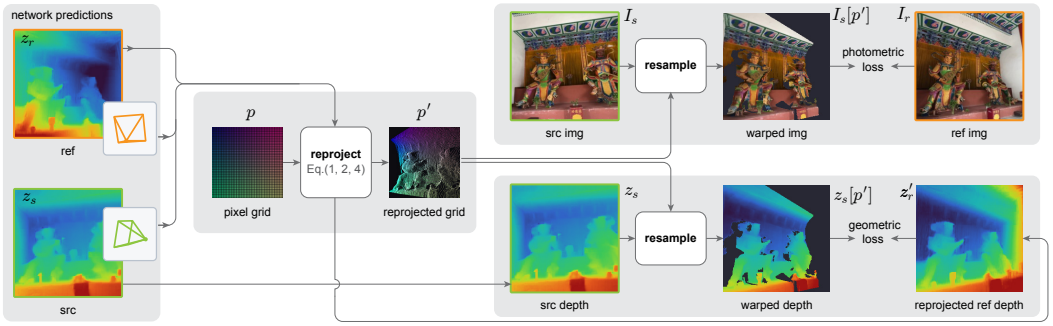

Figure 2: **Self-supervised objective.** We use image warping to encourage photometric and geometric consistency between views. Predicted cameras and reference depth define a reprojection grid for resampling the source image and depth map. The warped image is compared with the reference image, and the warped depth is compared with the reference depth reprojected into the source camera.

3.3 Self-supervised objective

Our self-supervised objective uses *warping* to enforce photometric and geometric consistency between pairs of views (Figure 2). Using the predicted cameras and depths, warping reprojects pixels from a reference view r into a source view s and differentiably resamples the source images or depth maps.

Reprojection. For a pixel p in reference view r, we obtain its 3D point $X_r(p)$ using Eq. (1) and project it into source view s:

$$p'_s = \pi\big(K_s R_s^\top (X_r(p) - \tau_s)\big), \tag{2}$$

where π denotes perspective division. Together with Eq. (1), this reprojection depends on the camera poses only through the relative rotation $R_s^\top R_r$ and translation $R_s^\top(\tau_r - \tau_s)$. Thus, a canonical coordinate system is not needed: applying the same arbitrary rotation and translation to all camera poses leaves the reprojection unchanged.

Photometric consistency. We use bilinear interpolation to sample the source image at p'_s and compute the photometric residual:

$$\rho_s = \rho\big(I_s[p'_s], I_r\big). \tag{3}$$

Here, $I_s[p'_s]$ denotes the warped source image, obtained by bilinear sampling at the reprojected coordinates p'_s.

Following prior work (Godard et al., 2017; 2019), we use $\rho = 0.85 \cdot \frac{1}{2}(1 - \mathrm{SSIM}) + 0.15 L_1$, with SSIM (Wang et al., 2004) computed over 3×3 windows and both terms averaged over color channels. Differentiable warping allows this loss to provide gradients to the predicted cameras in both views and the reference depth map.

Geometric consistency. We also enforce consistency between the predicted depth maps. The depth of the reference point $X_r(p)$ in the source camera is:

$$z'_r = \big[R_s^\top (X_r(p) - \tau_s)\big]_z \tag{4}$$

where $[\cdot]_z$ extracts the z-coordinate. We compare this depth with the source depth sampled at p'_s, using the normalized discrepancy of prior work (Bian et al., 2019):

$$g_s(p) = \frac{\big|z'_r - \mathrm{sg}\big(z_s[p'_s]\big)\big|}{z'_r + \mathrm{sg}\big(z_s[p'_s]\big)}, \tag{5}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. Here, $z_s[p'_s]$ denotes the warped source depth map.

Validity and masking. We define a validity mask $V_s(p) \in \{0, 1\}$ that retains only points in front of the source camera and projected within its image boundaries. For photometric consistency, we additionally mask predicted occlusions using

$$O_s(p) = \mathbf{1}\big[z'_r \le (1 + \varepsilon) z_s[p'_s]\big], \tag{6}$$

with $\varepsilon = 0.05$ allowing a small relative tolerance in the depth comparison. This test retains points at or in front of the predicted source surface. Following prior work (Godard et al., 2019), we also apply auto-masking, retaining only pixels whose warped photometric error is lower than the unwarped one.

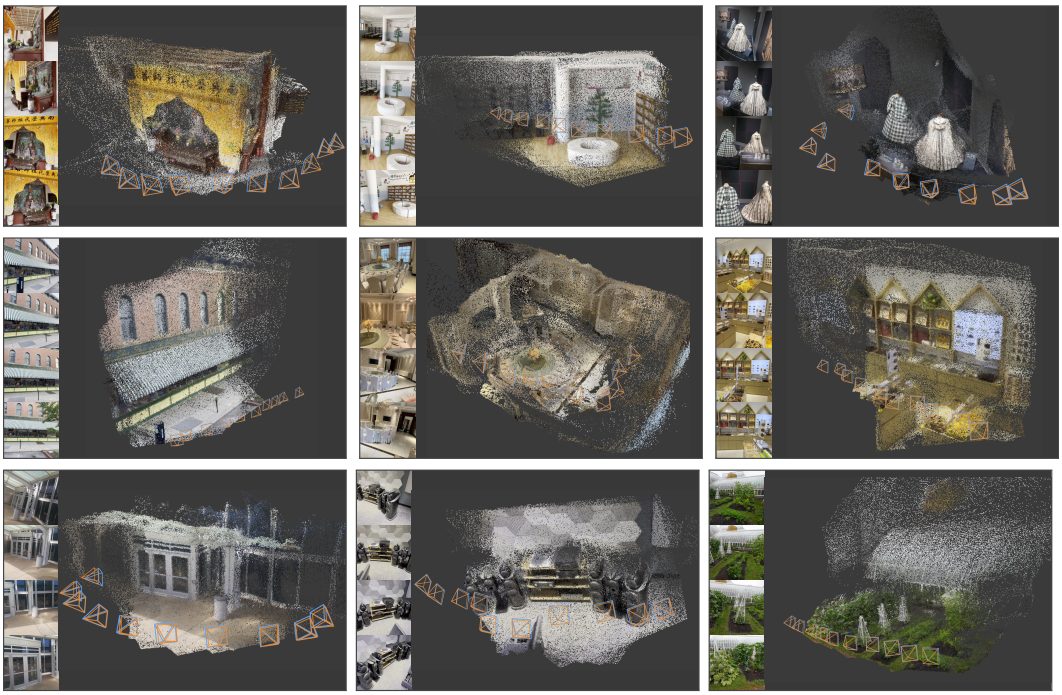

Figure 3: **Qualitative results of Warp3R**, predicted by our feedforward model on held-out DL3DV validation scenes. Point clouds are obtained by back-projecting predicted depths into the world coordinate system. Orange and blue frustums denote predicted and ground-truth cameras, respectively.

Loss aggregation. We use mean aggregation for photometric consistency and minimum aggregation for geometric consistency, adapting the minimum-reprojection principle of prior work (Godard et al., 2019). For a reference view r with temporal neighborhood N_r, the aggregated residuals are

$$L_{\text{photo}}(p) = \underset{s:\, V_s(p) O_s(p)=1}{\text{mean}} \rho_s(p), \qquad L_{\text{geo}}(p) = \underset{s:\, V_s(p)=1}{\min} g_s(p). \tag{7}$$

Both aggregations are over source views in N_r. Pixels with no eligible source are excluded from the corresponding term. We average $L_{\text{photo}}(p)$ over pixels retained by auto-masking and $L_{\text{geo}}(p)$ over pixels with a valid source, then average each term over reference views. The total loss is:

$$L_{\text{total}} = L_{\text{photo}} + 0.1 L_{\text{geo}}. \tag{8}$$

This objective jointly trains the backbone and heads using only the input images.

Temporal continuity as a prior. By default, the temporal neighborhood N_r contains the adjacent views: $s = r \pm 1$. This neighborhood exploits the spatial continuity of video, a data prior widely used in self-supervised reconstruction (Zhou et al., 2017; Godard et al., 2019; Bian et al., 2019; Jiang et al., 2025). We also encode temporal order through frame-index embeddings in the network (Jiang et al., 2025). This setting naturally fits static-scene video, a practical source of scalable unlabeled data for 3D reconstruction.

4 Experiments

We evaluate our method following the protocols of recent state-of-the-art methods for self-supervised 3D reconstruction: E-RayZer (Zhao et al., 2026) and NAS3R (Huang et al., 2026).

Protocols. We follow E-RayZer's 10-view evaluation protocol and NAS3R's 2-view pose evaluation protocol. We use their released checkpoints and, when available, the official evaluation code. We faithfully reproduced their published results. When evaluation code is unavailable, we implement the protocols described in the papers. Results depend on the frames sampled for evaluation, and differences in sampling may account for gaps between published and reproduced results. We therefore compare with both, denoted by "*orig*" and "*repro*" in Tables 1 and 2. The "*repro*" results use exactly the same evaluation frames as our method. Details are in A.4.

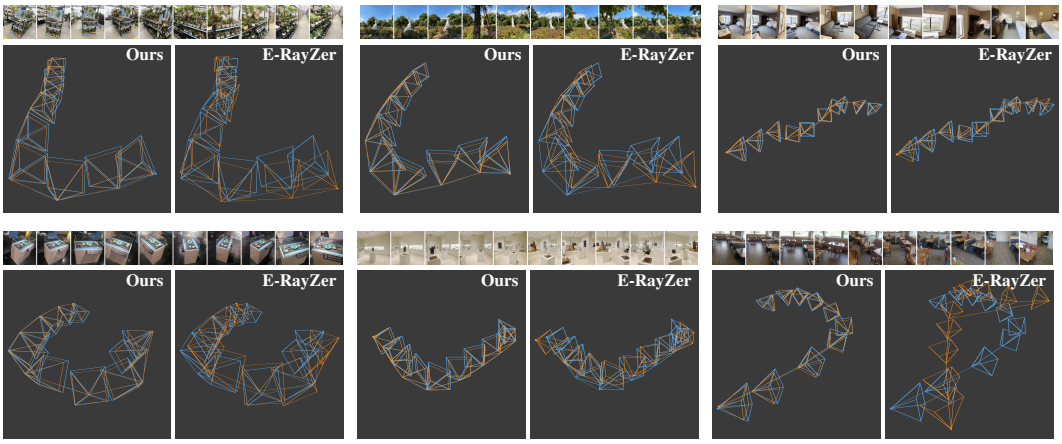

Figure 4: **Comparison between Warp3R and E-RayZer.** Warp3R predicts more accurate camera poses even where E-RayZer produces reasonable estimates. The bottom-right example shows a failure of E-RayZer. Orange and blue frustums denote predicted and ground-truth cameras, respectively.

Table 1: **Camera pose estimation *vs.* E-RayZer** Zhao et al. (2026). This table follows the evaluation suite of E-RayZer (their Table 2). 10-view relative pose accuracy (RPA, in %) at 5°, 15°, and 30° is reported. "E-RayZer (orig)" denotes the published results, and "repro" denotes our reproduced evaluation of theirs. Input resizes are 224 for ours and 256 for E-RayZer.

Method	In-domain			Out-of-domain (Zero-shot Generalization)																			
	DL3DV			RE10K			CO3Dv2			WildRGB-D			7-Scenes			CamLand			BlendedMVS			NAVI			ScanNet++		
	@5	@15	@30	@5	@15	@30	@5	@15	@30	@5	@15	@30	@5	@15	@30	@5	@15	@30	@5	@15	@30	@5	@15	@30	@5	@15	@30
E-RayZer (orig)	72.1	88.4	93.5	83.0	96.8	–	19.1	61.8	78.8	51.1	82.3	–	38.8	78.0	–	18.1	62.9	–	22.9	46.8	–	20.7	57.8	–	7.7	33.6	63.0
E-RayZer (repro)	72.1	88.5	93.5	81.9	96.6	98.5	18.9	59.0	76.3	51.3	82.3	90.6	38.6	78.8	94.6	20.4	61.1	86.5	23.8	**47.6**	**69.2**	19.1	55.5	68.6	13.3	41.7	61.5
Warp3R (ours)	**92.5**	**97.2**	**98.4**	**92.5**	**98.9**	**99.6**	**30.1**	**75.0**	**87.3**	**70.5**	**95.3**	**98.5**	**53.0**	**91.4**	**96.8**	**22.4**	**68.5**	**90.4**	**24.1**	44.4	53.3	**28.4**	**67.7**	**79.3**	**22.1**	**61.6**	**76.3**

Metrics. We follow the metrics used in E-RayZer and NAS3R. For camera pose estimation, we report relative pose accuracy (RPA) at 5°, 15°, and 30°, and normalized area under the error-recall curve (AUC) at 10° and 20° for rotation, translation direction, and their maximum (overall pose).

Implementation. We initialize from DINOv2 ViT-L/14 (Oquab et al., 2024) and train all parameters using 10-frame clips from the DL3DV (Ling et al., 2024) training set. For faster prototyping, our ablations use 10M training clips at 140×140, which is about $2\times$ faster than training at 224×224. We further fine-tune at 224×224 for 0.7M clips to compare with E-RayZer and NAS3R at comparable input resolutions. Details are in Appendix A.

4.1 Self-supervised camera pose estimation

Figure 3 shows qualitative reconstructions and camera predictions of Warp3R on held-out DL3DV scenes. We discuss quantitative comparisons below.

Comparison with E-RayZer. In Table 1, we report that Warp3R substantially improves camera pose accuracy over E-RayZer (Zhao et al., 2026), particularly at the strict 5° threshold. Large gains are observed across datasets, including in-domain DL3DV (**+20.4** percentage points at RPA@5) and zero-shot WildRGB-D (**+19.2**), 7-Scenes (**+14.4**), CO3Dv2 (**+11.2**), and RE10K (**+10.6**).

Figure 4 shows a qualitative comparison with E-RayZer on held-out DL3DV scenes. Warp3R predicts camera poses that more closely match the ground truth, even in scenes where E-RayZer produces reasonable estimates. Warp3R can produce accurate predictions even when E-RayZer fails.

Comparison with NAS3R. In Table 2, we compare with NAS3R on RE10K (Zhou et al., 2018) following its 2-view evaluation protocol; we also report 10-view RPA and include E-RayZer for comparison. To match NAS3R's training data, we train another Warp3R model on RE10K only. Both our models outperform NAS3R and E-RayZer across all reported metrics.

Table 2: **Camera pose estimation *vs.* NAS3R** (Huang et al., 2026) on RealEstate10K. We evaluate two settings: 2-view pose AUC (NAS3R's protocol) and 10-view RPA (E-RayZer's protocol), in %. Input resizes are 224 for NAS3R and ours, and 256 for E-RayZer.

Method	Data	Train views	**2-view eval** (NAS3R protocol)						**10-view eval**		
			AUC rotation		AUC translation		AUC overall		RPA		
			@10	@20	@10	@20	@10	@20	@5	@15	@30
NAS3R (orig)	RE10K	2	69.9	78.4	52.7	66.3	51.0	64.9	–	–	–
NAS3R (repro)	RE10K	2	70.1	78.4	52.7	66.4	51.3	65.0	12.9	34.4	50.5
NAS3R (repro)	RE10K	2–10	67.8	76.7	50.0	64.3	48.3	62.6	44.4	69.3	79.6
E-RayZer (repro)	DL3DV	10	48.4	63.4	23.5	43.5	21.6	41.1	81.9	96.6	98.5
Warp3R (ours)	RE10K	10	75.6	83.1	60.3	71.5	59.0	70.3	**94.1**	**99.1**	**99.7**
Warp3R (ours)	DL3DV	10	**77.0**	**84.7**	**63.4**	**75.1**	**61.4**	**73.6**	92.5	98.9	99.6

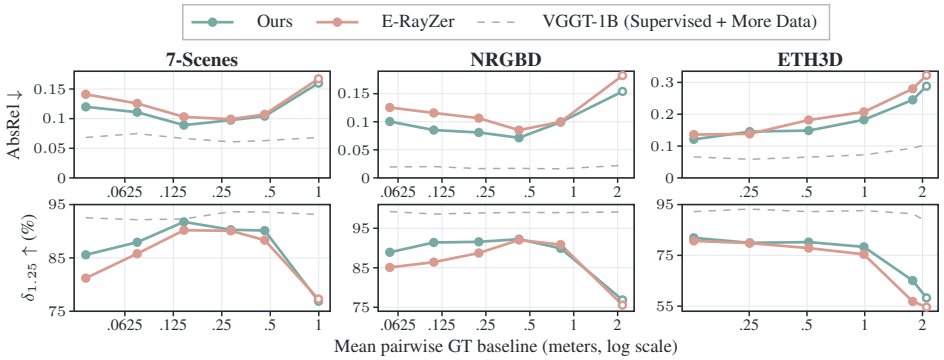

Figure 5: **Depth estimation.** Warp3R generally improves over E-RayZer across a range of camera baselines. VGGT-1B, trained on more data with supervision, serves as a reference. Figure details: All methods are evaluated on 10-view inputs. AbsRel (top, lower is better) and $\delta_{1.25}$ (bottom, higher is better) are plotted against the average pairwise ground-truth camera distance. Open markers (rightmost) denote random 10-view selections.

Our method exhibits strong generalization behavior. Our DL3DV-trained model outperforms the RE10K-trained NAS3R models despite being evaluated out of domain. In addition, Warp3R generalizes directly from 10-view training to 2-view evaluation without adaptation.

4.2 Self-supervised depth estimation

Next, we evaluate self-supervised depth estimation, with results shown in Figure 5.

Setting. We evaluate depth estimation on 7-Scenes, Neural RGB-D, and ETH3D. We use 10-view inputs sampled at different temporal strides or sampled randomly. We report absolute relative error (AbsRel) and $\delta_{1.25}$ accuracy. Warp3R and E-RayZer are trained with self-supervision on DL3DV. VGGT-1B (Wang et al., 2025), trained on more data with supervision, serves as a reference. All methods use the same evaluation crops and per-window scale alignment. Details are in A.4.

Results. Figure 5 shows that Warp3R generally improves depth estimation over E-RayZer across a range of camera baselines. The gains are most evident on 7-Scenes and Neural RGB-D, where Warp3R achieves lower AbsRel and higher $\delta_{1.25}$ accuracy at small to moderate baselines. A gap remains to supervised VGGT-1B, which may be partly attributed to its larger model size, more training data, and access to ground-truth depth supervision.

4.3 Ablation studies

In Table 3, we ablate Warp3R by changing one design choice at a time. We report changes relative to the default baseline at 140×140 resolution, denoted by **(*)**. For this baseline, we also report the standard deviation across 5 independent training runs to help interpret the observed deltas relative to run-to-run variation. We make the following observations:

- Removing the geometric loss **(a)**, *i.e.*, using L_{photo} alone, reduces accuracy but remains strong.

Table 3: **Ablations of Warp3R**: our method remains strong and robust to a variety of design choices. Table details: All are trained on DL3DV. We report 10-view camera pose estimation on DL3DV and, zero-shot, on RE10K and 7-Scenes, together with depth estimation on 7-Scenes. Baseline (*) reports absolute results as mean ± standard deviation over five independent runs. Each other row changes a single setting relative to (*) and reports single-run deltas from the baseline mean. See A.1 for details.

4.0% degradation — 4.0% improvement

Setting	DL3DV, pose		RE10K, pose		7-Scenes, pose		7-Scenes, depth	
	RPA@5↑	RPA@15↑	RPA@5↑	RPA@15↑	RPA@5↑	RPA@15↑	AbsRel↓	$\delta_{1.25}$ ↑
baseline								
(*) default @ 140×140	87.6 ±1.1	95.6 ±0.2	91.1 ±0.5	98.8 ±0.1	53.7 ±2.7	89.7 ±1.5	.111 ±.003	88.5 ±0.8
loss terms								
(a) no geometric loss	−2.1	−0.4	−0.6	+0.0	−5.5	−1.1	+0.011	−1.3
(b) + smoothness loss	+0.1	+0.8	−0.6	+0.1	+1.4	+0.8	−0.010	+0.8
masking								
(c) no auto-mask	−2.8	−0.3	−1.0	−0.1	−4.0	−2.3	+0.005	+0.0
(d) no occlusion mask	+0.3	+0.5	+0.0	+0.2	−1.0	−0.9	−0.002	−0.5
stop-grad								
(e) no depth stop-grad	+1.2	+0.7	+0.2	+0.1	+0.4	+0.0	−0.007	+0.4
source neighborhood (default: ±1)								
(f) ±2 neighbors	+2.8	+0.9	−0.5	+0.0	+0.7	−0.8	−0.004	−0.1
(g) ±3 neighbors	+0.2	+0.2	−1.4	−0.2	+0.6	+1.2	−0.002	−0.8
depth head								
(h) conv-refine head	+1.6	+0.5	+0.2	+0.0	−0.2	+1.3	+0.006	−0.8
global attention (default: every 2)								
(i) every 4	−0.8	−0.4	−1.7	−0.2	−1.9	−1.9	+0.002	−0.4
(j) every 6	−1.4	−0.2	−1.3	−0.2	−4.5	−3.2	−0.002	−0.6
pre-trained init (default: DINOv2)								
(k) MoCo v3 init	−0.2	+0.0	+0.2	+0.1	−5.2	−0.3	−0.002	−0.3
(l) MAE init	+0.7	+0.9	−0.6	+0.1	−1.5	−0.9	−0.003	+0.3
(m) PIXIO init	+0.3	+0.4	−1.6	−0.1	−3.0	−0.1	−0.007	+0.7
frame indexing								
(n) no frame emb	−3.2	−3.0	−2.1	−0.5	−2.0	+0.3	+0.000	−0.1
high-resolution tuning								
(o) @ 224×224	+4.9	+1.6	+1.4	+0.1	−0.7	+1.7	+0.005	−0.2

- Adding a classical smoothness loss (Godard et al., 2019) **(b)** is slightly beneficial but not necessary.
- Removing auto-masking **(c)** reduces pose accuracy, while removing occlusion masking (Eq. 6) **(d)** retains similar performance.
- Removing stop-gradient through the source depth (Eq. 5) **(e)** maintains strong performance. However, we keep stop-gradient because we observed improved training stability in other ablations.
- Using more neighboring views **(f, g)** improves results on some datasets but not all.
- Replacing our simple unpatchifying depth head with a conv-refine head (Wang et al., 2025) **(h)** improves several pose metrics, but not depth metrics in our case.
- Using fewer global attention blocks **(i, j)** generally harms accuracy.
- Initializing with MoCo v3 (Chen et al., 2021), MAE (He et al., 2022), or Pixio (Yang et al., 2026) **(k, l, m)** also yields strong in-domain results though may degrade out-of-domain results. We have not yet achieved strong results with random initialization and leave this exploration to future work.
- Removing frame-index embeddings **(n)** generally reduces pose accuracy, with little effect on depth.
- Fine-tuning at 224 **(o)** improves in-domain pose accuracy, with smaller or no gains in other cases.

Overall, Warp3R remains strong across a variety of design choices, demonstrating the robustness of our warping-based self-supervised approach. In fact, all ablated configurations still outperform E-RayZer by large margins in RPA@5: its results in Table 1 correspond to deltas of −15.5, −9.2, and −15.1 percentage points relative to **(*)** on DL3DV, RE10K, and 7-Scenes.

4.4 Extending supervised training with unlabeled data

We next study whether our self-supervised objective can complement *supervised* training using unlabeled data, a practically useful scenario.

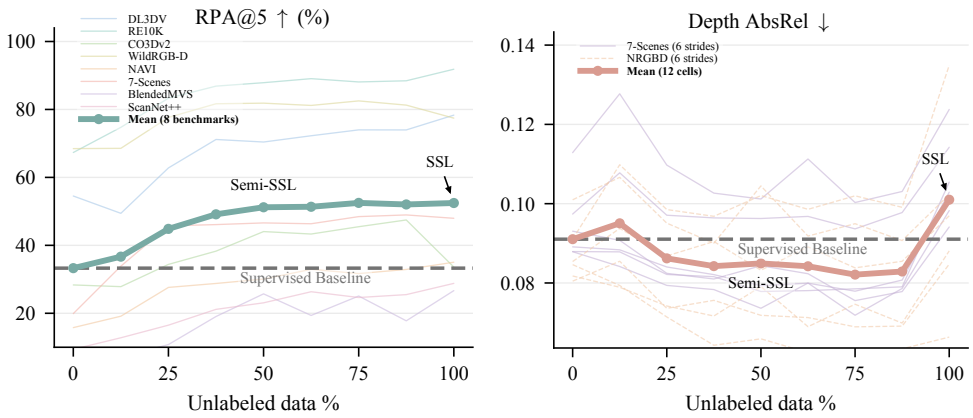

Figure 6: **Extending supervised training with unlabeled data.** Our self-supervised objective can *improve* pose and depth estimation over purely supervised training. Figure details: We vary the fraction of unlabeled training samples under a fixed training budget. We report average RPA@5 over 8 pose benchmarks (left) and AbsRel over 12 depth settings (2 datasets by 6 strides; right). Dashed lines denote purely supervised baselines.

Setting. We construct a *supervised* counterpart to Warp3R using the same backbone as ours and the same supervised losses as VGGT (Wang et al., 2025). All training runs use the *same* number of batches and the same batch size, *i.e.*, semi-supervised training does not benefit from longer training. We vary the proportion of labeled and unlabeled samples drawn from their respective data pools. Labeled samples use only the supervised loss, while unlabeled samples use only our self-supervised warping loss. The purely supervised model serves as the baseline (dashed lines in Figure 6). All settings use the same architecture. Details are in A.3.

Results. Our self-supervised objective substantially strengthens supervised training (Figure 6).

For camera pose estimation (Figure 6, left), *all* semi-supervised settings improve average RPA@5 over the supervised baseline; even purely self-supervised training outperforms this baseline. These gains arise even though unlabeled samples replace part of the labeled training input under the same training budget.

For depth estimation (Figure 6, right), a wide range of mixture ratios yields lower average AbsRel error than the supervised baseline. Purely self-supervised training still yields higher error than all semi-supervised settings, suggesting that depth labels provide additional information that our self-supervised objective does not fully recover.

Together, these results demonstrate the value of *unlabeled* data and warping-based self-supervision for complementing supervised 3D reconstruction.

5 Conclusion

We introduced Warp3R, a simple self-supervised 3D reconstruction method combining classical warping losses with a VGGT-inspired multi-view feed-forward backbone. Despite its simplicity, Warp3R significantly outperforms state-of-the-art self-supervised methods in camera pose estimation and improves depth estimation. Our self-supervised objective can also complement supervised training by exploiting unlabeled data, suggesting potential for real-world applications where unlabeled data are more readily available than geometric labels.

Our study offers a computational approach toward mimicking the human ability to understand the 3D world with little to no explicit geometric supervision. Humans, and many other animals, perceive the 3D world through movement and observations from changing viewpoints. Our results suggest that consistency across such observations can provide a powerful learning signal, bringing us closer to machines that acquire 3D understanding from visual experience.

AI USE STATEMENT

In this work, we used generative AI tools for identifying some related works, polishing writing, early drafting, and to provide paper feedback. We additionally used generative AI tools to critique our ideas, and for coding and launching experiments. We have not used generative AI tools for creation of synthetic datasets, and uses in proving mathematical claims are not applicable to this work. We have reviewed all AI-assisted work, including writing for falsehoods or hallucinations, and benchmarks were always run with adjacent reproductions and controls to ensure correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Hatem Alismail, Brett Browning, and Simon Lucey. Photometric bundle adjustment for vision-based SLAM. In *ACCV*, 2016. URL https://arxiv.org/abs/1608.02026.

Dejan Azinović, Ricardo Martin-Brualla, Dan B. Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *CVPR*, 2022. URL https://arxiv.org/abs/2104.04532.

Jia-Wang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *NeurIPS*, 2019. URL https://arxiv.org/abs/1908.10553.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. URL https://arxiv.org/abs/2104.02057.

Eleanor J. Gibson and Richard D. Walk. The "visual cliff". *Scientific American*, 202(4):64–71, 1960. doi: 10.1038/scientificamerican0460-64. URL https://www.scientificamerican.com/article/the-visual-cliff/.

Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. URL https://arxiv.org/abs/1609.03677.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. URL https://arxiv.org/abs/1806.01260.

Marwane Hariat, Gianni Franchi, David Filliat, and Antoine Manzanera. SS3D: End2End self-supervised 3D from web videos. *arXiv preprint arXiv:2604.22686*, 2026. URL https://arxiv.org/abs/2604.22686.

Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004. URL https://doi.org/10.1017/cbo9780511811685.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. URL https://arxiv.org/abs/2111.06377.

Ranran Huang, Weixun Luo, Ye Mao, and Krystian Mikolajczyk. From none to all: Self-supervised 3D reconstruction via novel view synthesis. *arXiv preprint arXiv:2603.27455*, 2026. URL https://arxiv.org/abs/2603.27455.

Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, André Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. NAVI: Category-agnostic image collections with high-quality 3D shape and pose annotations. In *NeurIPS Datasets and Benchmarks Track*, 2023.

Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, and Georgios Pavlakos. RayZer: A self-supervised large view synthesis model. In *ICCV*, 2025. URL https://arxiv.org/abs/2505.00702.

Takeo Kanade and Masatoshi Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9): 920–932, 1994. URL https://doi.org/10.1109/34.310690.

Gyeongjin Kang, Jisang Yoo, Jihyeon Park, Seungtae Nam, Hyeonsoo Im, Sangheon Shin, Sangpil Kim, and Eunbyung Park. SelfSplat: Pose-free and 3D prior-free generalizable 3D Gaussian splatting. In *CVPR*, pp. 22012–22022, 2025. URL https://arxiv.org/abs/2411.17190.

Philip J. Kellman. Perception of three-dimensional form by human infants. *Perception & Psychophysics*, 36(4):353–358, 1984. doi: 10.3758/BF03202789. URL https://link.springer.com/article/10.3758/BF03202789.

Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6555–6564, 2017. doi: 10.1109/CVPR.2017.694.

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

Michael S. Landy, Laurence T. Maloney, Elizabeth B. Johnston, and Mark Young. Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research*, 35(3):389–412, 1995. URL https://doi.org/10.1016/0042-6989(94)00176-m.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *ECCV*, 2024. URL https://arxiv.org/abs/2406.09756.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *CVPR*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. URL https://arxiv.org/abs/1711.05101.

Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981. URL https://www.ijcai.org/Proceedings/81-2/Papers/017.pdf.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction, 2021. URL https://arxiv.org/abs/2103.13413.

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 *ICCV*, 2021.

Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47:7–42, 2002. URL https://link.springer.com/article/10.1023/A:1014573219977.

Maarten Schellevis. Improving self-supervised single view depth estimation by masking occlusion. *arXiv preprint arXiv:1908.11112*, 2019. URL https://arxiv.org/abs/1908.11112.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Schonberger_Structure-From-Motion_Revisited_CVPR_2016_paper.html.

Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. URL https://link.springer.com/chapter/10.1007/978-3-319-46487-9_31.

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 *CVPR*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Schops_A_Multi-View_Stereo_CVPR_2017_paper.html.

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 *CVPR*, 2013. URL https://openaccess.thecvf.com/content_cvpr_2013/html/Shotton_Scene_Coordinate_Regression_2013_CVPR_paper.html.

Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice*, pp. 298–372. Springer, 2000. URL https://link.springer.com/chapter/10.1007/3-540-44480-7_21.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025. URL https://arxiv.org/abs/2503.11651.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *CVPR*, 2026a. URL https://arxiv.org/abs/2605.15195.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *CVPR*, 2024. URL https://arxiv.org/abs/2312.14132.

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 *ICLR*, 2026b. URL https://arxiv.org/abs/2507.13347.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. URL https://doi.org/10.1109/tip.2003.819861.

Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. CroCo: Self-supervised pre-training for 3D vision tasks by cross-view completion. In *NeurIPS*, 2022.

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 *CVPR*, 2024.

Yangfan Xu, Lilian Zhang, Xiaofeng He, Pengdong Wu, Wenqi Wu, and Jun Mao. GPA-VGGT: Adapting VGGT to large scale localization by self-supervised learning with geometry and physics aware loss. *arXiv preprint arXiv:2601.16885*, 2026. URL https://arxiv.org/abs/2601.16885.

Lihe Yang, Shang-Wen Li, Yang Li, Xinjie Lei, Dong Wang, Abdelrahman Mohamed, Saining Xie, Hengshuang Zhao, Kaiming He, and Hu Xu. In pursuit of pixel supervision for visual pre-training. In *CVPR*, 2026. URL https://arxiv.org/abs/2512.15715.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, 2020.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *ICCV*, 2023.

Qitao Zhao, Hao Tan, Qianqian Wang, Sai Bi, Kai Zhang, Kalyan Sunkavalli, Shubham Tulsiani, and Hanwen Jiang. E-RayZer: Self-supervised 3D reconstruction as spatial visual pre-training. In *CVPR*, 2026. URL https://arxiv.org/abs/2512.10950.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. URL https://arxiv.org/abs/1704.07813.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4), 2018.

Table 4: **Configuration summary.** Settings for the model used in the main pose comparisons. The 224×224 stage is our main configuration, and resumes from the 140×140 checkpoint.

	140×140	224×224
architecture		
backbone initialization	DINOv2 ViT-L/14 with registers (Oquab et al., 2024)	
patch size	14	
depth head	linear unpatchify decoder	
camera head	2-layer MLP	
data and sampling		
training data	DL3DV (Ling et al., 2024) video	
clip length	10 frames	
batch	16 clips (160 frames) per step	
evaluation resolution	224×224	
optimization		
optimizer	AdamW (Loshchilov & Hutter, 2019), β_1=0.9, β_2=0.95	
learning rate	10^{-4}	
weight decay	0.01	
lr warmup	10k linear warmup	no warmup
lr decay	no decay	
gradient clipping	1.0	
model precision	bfloat16, fp32 accumulation	
loss precision	fp32	
ema decay (eval)	0.999	
schedule (main pose model)		
iterations	625k	extra 44.5k
clips	10M	extra 0.7M

A Additional implementation details

A.1 Configurations

Table 4 summarizes our architectural choices and training hyperparameters. We train with self-supervision on DL3DV (Ling et al., 2024) for 10M clips at 140×140, then fine-tune for 0.7M clips at 224×224 to compare with E-RayZer (Zhao et al., 2026) and NAS3R (Huang et al., 2026) at comparable input resolutions. Each clip contains 10 frames, sampled using temporal strides $\{2, 4, 6, 8\}$ during the first stage and stride 8 during fine-tuning. The patch size remains 14; weights carry over and the positional embedding is re-interpolated.

The default model uses DINOv2 (Oquab et al., 2024) initialization, global attention every two Transformer blocks, and learned frame-index embeddings. Its unpatchifying depth head uses features from layers 6, 12, 18, and 24, following the use of intermediate features in prior work (Ranftl et al., 2021); we sum these features before the depth head. On a v6e-16 TPU, training takes 0.11 s/step at 140×140 and 0.21 s/step at 224×224. At these step times, the schedules in Table 4 correspond to approximately 19.1 hours and 2.6 hours.

A.2 Geometry and loss implementation

The camera head predicts an 8-vector comprising a quaternion, translation, and focal length. We normalize the scalar-first quaternion $q = (w, x, y, z)$ and convert it to a camera-to-world rotation R_i.

Square pixels and a centered principal point give

$$K_i = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{9}$$

$$f_x = f_y = f_i W, \qquad c_x = W/2, \quad c_y = H/2. \tag{10}$$

A softplus produces positive z-depth z_i and focal lengths f_i.

Table 5: **Benchmarks statistics.** We report the length of the evaluation clips over which we evenly subsample. Exceptions to frame selection process are stated. Pair counts follow the 5% baseline filter except for CamLand. *Released* indicates that both the selection rule and scene list were recovered from published code or split files; *reconstructed* indicates that at least one was unavailable and had to be inferred.

Benchmark	Clip Length	Scenes	Pairs	Protocol source
DL3DV (Ling et al., 2024)	96	140	6,293	released (method split)
RE10K (Zhou et al., 2018)	256	1,088	48,956	released (method rule)
CO3Dv2 (Reizenstein et al., 2021)	first 96 frames	510	22,856	reconstructed
WildRGB-D (Xia et al., 2024)	192	460	20,634	released (dataset split)
7-Scenes (Shotton et al., 2013)	256	18	809	released (dataset split)
CamLand (Kendall & Cipolla, 2017)	96-frame windows	5	225	reconstructed
BlendedMVS (Yao et al., 2020)	overlap set, then 24	7	315	reconstructed
NAVI (Jampani et al., 2023)	24	136	6,109	reconstructed
ScanNet++ (Yeshwanth et al., 2023)	48	50	2,236	released (dataset split)

A.3 Co-training protocol

We train with nine batch compositions, with $0, 2, \ldots, 16$ unlabeled sequences per batch of 16, training each for 629k steps at 140×140 from DINOv2 initialization. The labeled and unlabeled pools each contain frames from the training sets of DL3DV, RealEstate10K, CO3Dv2, ARKitScenes, ScanNet, WildRGB-D, and MegaDepth. For both our labeled and unlabeled pools, we use relative sequence-sampling weights $2 : 1 : 2 : 2 : 2 : 1 : 1$. The VGGT loss sets scene scale based on ground truth camera translation and not ground truth points. Labeled sequences receive only the supervised VGGT loss, and unlabeled sequences receive only our self-supervised loss. DL3DV and RE10K are pose-only and receive no depth supervision within the labeled pool.

The pose mean weights evaluation datasets equally, and the depth mean weights all stride combinations equally, using strides $\{1, 2, 4, 8, 12, 16\}$.

A.4 Evaluation protocols

For our primary pose metrics, we follow the 10-view evaluation suite in E-RayZer's Table 2 (Zhao et al., 2026). Where the original protocol is not fully recoverable, we reconstruct it from released code and data and apply it consistently to all methods. Each reconstructed protocol is checked against the original method's reported results and applied identically to both methods. Table 5 overviews each of these datasets.

Preprocessing. Pose inputs use an aspect-preserving short-side resize and center crop, sampled at each model's native resolution: 224×224 for ours and 256×256 for E-RayZer (Zhao et al., 2026). Crops are otherwise identical, with no anisotropic rescaling or model-specific correction.

Two-view evaluation. Table 2 uses NAS3R's released RealEstate10K context pairs for $5{,}601$ scenes. Rotation, translation-direction, and maximum pose errors follow the definitions above; AUC integrates each error-recall curve to 10° or 20° and divides by that threshold. We evaluate without baseline filtering, or outlier rejection.

Depth estimation. We sample ten-frame windows at strides $s \in \{1, 2, 4, 8, 16\}$ plus random ten-frame selections: 36 windows per stride on 7-Scenes Shotton et al. (2013), 36 on Neural RGB-D Azinović et al. (2022), and 32 on ETH3D Schöps et al. (2017). All methods receive identical center-cropped frames at their native resolutions and are scored on a common 256×256 ground-truth grid. We fit one median scale per window, floor ground-truth depth at 0.1 m, and average errors over frames and then windows. ETH3D uses raw laser depth without hole filling and tests occlusion against the reconstructed surface mesh; approximately 7–8% of pixels are valid, and units with none are skipped. This replaces mesh raycasting, which assigns depth to unscanned surfaces: the correction halved the supervised reference's error but left both self-supervised models nearly unchanged.

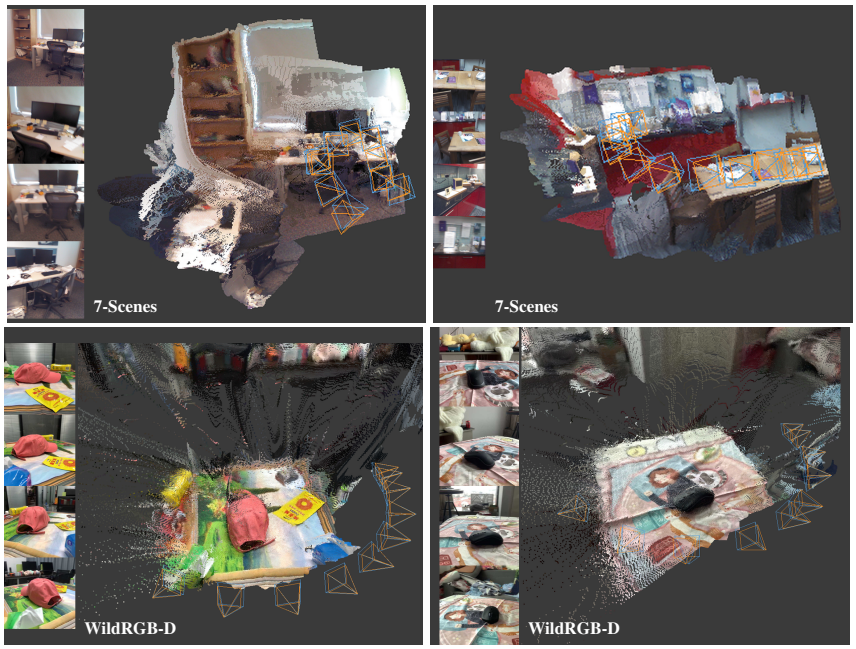

Figure 7: **Out-of-domain reconstruction.** Warp3R reconstructs scene structure on datasets unseen during training. We show examples from 7-Scenes and WildRGB-D, with input views beside each predicted point cloud. Orange and blue frustums denote predicted and ground-truth cameras.

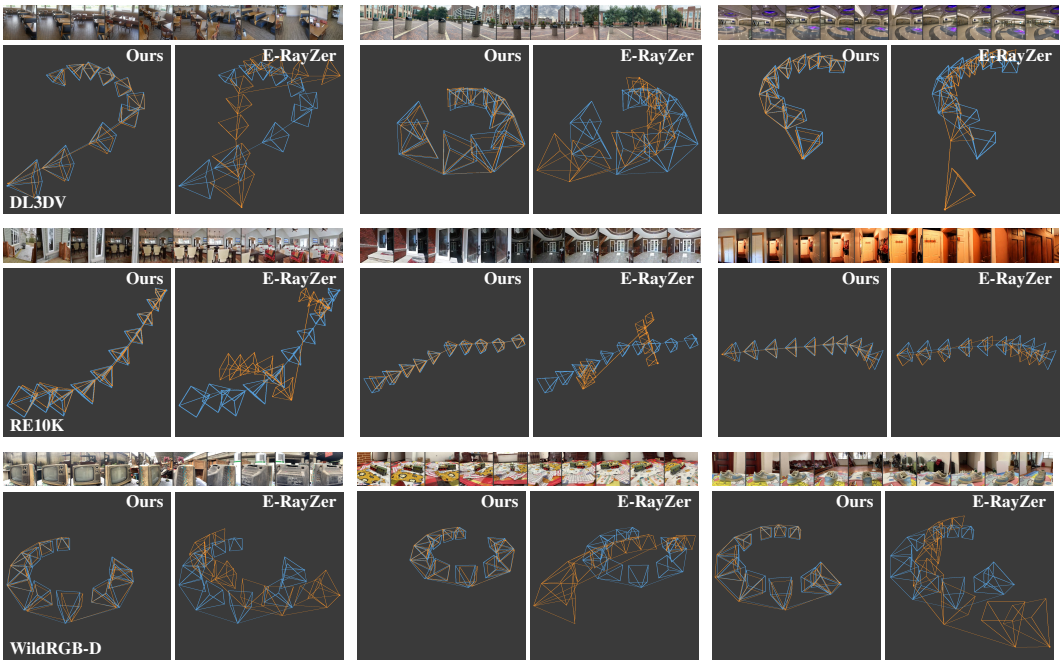

Figure 8: **More comparisons with E-RayZer.** Warp3R predicts accurate camera poses in examples where E-RayZer (Zhao et al., 2026) fails. We show three held-out clips each from DL3DV and the out-of-domain datasets RE10K and WildRGB-D. Input views appear above each comparison. Orange and blue frustums denote predicted and ground-truth cameras.

B Additional Qualitative Results

We show out-of-domain reconstructions in Figure 7 and further comparisons with E-RayZer in Figure 8. Figures 9 and 10 show camera pose estimation in textureless and glossy scenes.

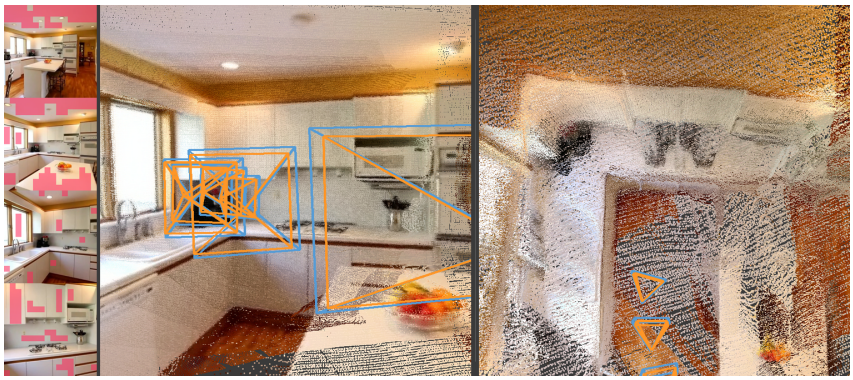

Figure 9: **Camera pose estimation of a textureless scene.** Warp3R recovers coherent scene structure despite large textureless regions. Input views from RE10K are shown on the left, with textureless regions (local variance below a threshold) highlighted in pink; the predicted point cloud is shown from two viewpoints. Orange and blue frustums denote predicted and ground-truth cameras.

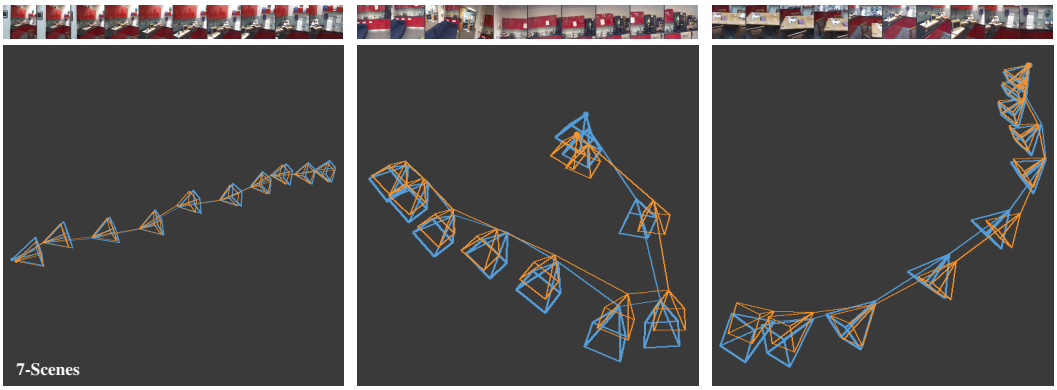

Figure 10: **Camera pose estimation in glossy scenes.** Warp3R predicts accurate camera poses despite reflective surfaces. Input views from three 7-Scenes clips are shown above their predicted camera poses. Orange and blue frustums denote predicted and ground-truth cameras.