UNIREG: JOINT MULTI-SOURCE LEARNING FOR UNIVERSAL IMAGE-TO-POINT-CLOUD REGISTRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Image-to-point-cloud registration recovers the 6-DoF pose of a query RGB image relative to a reference point cloud. While recent learning-based methods perform well on individual benchmarks, they remain largely tied to a fixed sensing process: outdoor sparse LiDAR and indoor RGB-D reconstructions differ substantially in density, scanning pattern, metric scale, completeness, and image–map overlap. As a result, models trained under one acquisition setup transfer poorly to another. We analyze this failure mode and attribute it to representations that entangle acquisition cues with scene geometry, pose pathways that are sensitive to density and scale, and the limited effectiveness of naively mixing heterogeneous datasets. We present UniReg, a single feed-forward network for I2P registration across indoor and outdoor scenes, LiDAR and RGB-D inputs, and fragment- to scene-level geometry. UniReg separates acquisition context from geometric reasoning through token-space conditioning, predicts sensor-robust ray-centric multi-geometry, and aggregates these predictions via confidence-aware fusion. To support this factorization, we introduce a scene-level ScanNet I2P dataset and a CARLA corpus with diverse camera–LiDAR configurations, and train jointly with standard benchmarks. With one set of weights, UniReg matches or surpasses specialist models across established indoor and outdoor evaluations.

1 INTRODUCTION

Image-to-point-cloud (I2P) registration estimates the six-degree-of-freedom (6-DoF) camera pose of a query RGB image with respect to a reference 3D point cloud. It serves as a cornerstone for LiDAR–camera calibration (Yuan et al., 2020), visual localization (Sarlin et al., 2023), simultaneous localization and mapping (SLAM) (Durrant-Whyte & Bailey, 2006), and augmented reality (Billinghurst et al., 2015). Beyond the modality gap between images and point clouds, a practical difficulty arises from the heterogeneity of the reference geometry itself. Outdoor systems typically acquire sparse LiDAR measurements whose density and coverage depend on the beam configuration, whereas indoor systems more commonly obtain point clouds by back-projecting RGB-D depth, either as local fragments or as fused scene-scale reconstructions. Although these inputs share a common point-set representation, they differ substantially in density, scanning pattern, metric scale, completeness, and image–map overlap. As a result, methods developed under one acquisition setup frequently degrade under another, limiting the deployability beyond individual benchmarks. Fig. 1(a) illustrates this heterogeneity across outdoor LiDAR and indoor RGB-D inputs.

Existing learning-based approaches typically address I2P registration by learning cross-modal features, establishing 2D–3D correspondences, and recovering the camera pose through geometric solvers (Ren et al., 2022; Zhou et al., 2023; Li et al., 2025). Subsequent work strengthens this pipeline with improved backbones, coarse-to-fine matching, and more robust outlier rejection (Li et al., 2026; Kang et al., 2024; Bie et al., 2025). Despite their effectiveness on individual benchmarks, these methods remain largely dataset-specific: a separate model is trained for each acquisition setup, and accuracy drops sharply when the sensing process changes. We attribute this limited generalization to three tightly coupled shortcomings: (1) **Entangled representations.** Under a fixed sensor and construction protocol, learned features absorb acquisition-dependent cues such as point density, scanning pattern, scale, and typical overlap, together with scene geometry. The model therefore fits registration under a particular sensing process, rather than learning a sensor-agnostic spatial mapping. (2) **Sensor-sensitive geometric pathways.** Pose is often recovered from sparse 2D–3D

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.

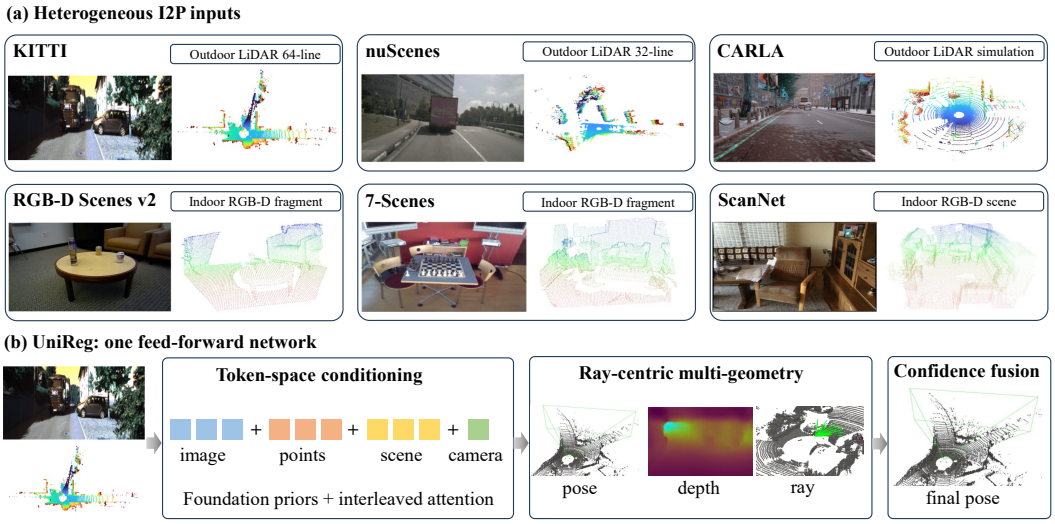

Figure 1: **UniReg** registers heterogeneous image–point cloud pairs with one feed-forward network. **(a)** Reference geometry varies across outdoor LiDAR (beam configuration and simulation) and indoor RGB-D (local fragments versus scene-scale reconstructions). **(b)** Token-space conditioning absorbs acquisition factors, ray-centric multi-geometry predicts pose, depth, and per-pixel rays, and confidence-aware fusion yields the final 6-DoF pose.

matches or from dense outputs defined in the reference metric frame, such as depth or scene-frame point maps. Both pathways depend on the absolute layout and sparsity of a particular cloud: matches require points to exist at the right places, and scene-frame predictions must follow that cloud's scale and sampling, so the pose pathway changes whenever the sensing process changes. (3) **Limited benefit of naive multi-dataset training.** Pooling multiple datasets into the same specialist pipeline does not remove the above issues. Without an explicit separation of acquisition context from geometric reasoning, and without geometric representations that remain valid under density and scale changes, joint training tends to fit the dominant regime or to interfere across regimes. Recent work has sought to improve I2P generalization through pretrained cross-modal features, topology- or graph-based reasoning, and unified matching architectures (Wang et al., 2024; An et al., 2025; 2026; Goswami et al., 2026). These methods mainly strengthen cross-scene transfer within a similar sensing family, most notably indoor settings, where density, scale, and overlap statistics remain largely unchanged. Acquisition cues can still dominate the representation, match- or metric-frame pathways remain tied to the layout and sparsity of a particular cloud, and extending beyond a single sensing regime still requires per-sensor adaptation.

To overcome these limitations, we propose UniReg, a single feed-forward network that performs I2P registration across indoor and outdoor environments, across LiDAR and RGB-D sensing, and across fragment-level and scene-level geometry with one set of weights (Fig. 1(b)). We instantiate this goal with three key designs: (i) **Token-space conditioning.** To disentangle acquisition context from geometric features, UniReg initializes image and point-cloud encodings with foundation-model priors (Oquab et al., 2024; Wu et al., 2024; Wang et al., 2025) and fuses them by interleaved self- and cross-attention. A learnable camera token aggregates pose-related evidence, while a small set of scene tokens absorbs domain factors such as scale, density, overlap, and imaging style. Domain adaptation thus occurs in token space, and backbone features are kept more transferable across sensors. (ii) **Ray-centric multi-geometry.** To replace sensor-sensitive pose pathways, UniReg jointly predicts camera pose, dense depth, and a per-pixel ray map in the point-cloud frame. A ray encodes the projective imaging process of each image patch and depends on camera geometry rather than on the absolute metric layout or sparsity of a particular cloud. Unlike a point map, it remains meaningful when LiDAR density, RGB-D completeness, or scene extent changes. Pose, rays, and depth are coupled by projective geometry and provide complementary constraints under sensor shift. (iii) **Confidence-aware fusion and diverse supervision.** To make multi-source training effective rather than conflicting, UniReg fuses the above three predictions by confidence-aware fusion, turning projective agreement into a final pose. We further train jointly on standard benchmarks (KITTI, nuScenes, RGB-D Scenes V2, and 7-Scenes) together with two complementary sources that we introduce: a self-constructed scene-level ScanNet I2P dataset with complete, uncropped reconstruc-

tions, and a self-collected CARLA corpus covering diverse image and LiDAR configurations with exact poses and dense depth. Under this mixture, data diversity reinforces the factorization and ray-centric geometry above, instead of averaging specialist behaviors into one weight set. After joint training with standard benchmarks, a single UniReg model matches or surpasses per-dataset specialist methods across established indoor and outdoor evaluations.

Our main contributions are summarized as follows: (1) We identify cross-sensor generalization as the central open problem in I2P registration, analyze why existing correspondence-centric and dataset-specific pipelines fail to achieve it, and present UniReg, a single feed-forward framework that registers images to point clouds across indoor and outdoor scenes, LiDAR and RGB-D inputs, and fragment- to scene-level geometry with one set of weights. (2) We introduce a coherent design that separates acquisition context from geometric reasoning through token-space conditioning, replaces sensor-sensitive match- or metric-frame pose pathways with ray-centric multi-geometry prediction (pose, depth, and per-pixel rays), and combines these predictions into the final pose through confidence-aware fusion. (3) We construct complementary training resources for acquisition diversity, including a scene-level ScanNet I2P dataset and a large-scale CARLA corpus with diverse sensor configurations, and show that joint training with standard benchmarks (KITTI, nuScenes, RGB-D Scenes V2, and 7-Scenes) enables one UniReg model to match or surpass per-dataset specialist methods across established indoor and outdoor evaluations.

2 Related Works

Specialist image-to-point-cloud registration. Image-to-point-cloud registration estimates the camera pose of an RGB image with respect to a reference point cloud, and is commonly solved by learning cross-modal features, establishing 2D–3D correspondences, and applying a geometric solver Lepetit et al. (2009). This paradigm inherits correspondence techniques developed for images DeTone et al. (2018); Sarlin et al. (2020); Sun et al. (2021) and point clouds Yu et al. (2021); Qin et al. (2023), but must additionally bridge a modality gap. Early cross-modal methods extract features independently for each modality Feng et al. (2019); Li & Lee (2021). Later designs predict overlap before matching Ren et al. (2022), insert a differentiable PnP solver Zhou et al. (2023), adapt coarse-to-fine matchers to 2D–3D features Kang et al. (2024), or augment correspondence with graph reasoning and auxiliary depth Bie et al. (2025); Li et al. (2026). Implicit regression Li et al. (2025) avoids an explicit match set, but still conditions geometric reasoning on the metric layout of the observed cloud. These methods are accurate when training and testing share the same acquisition setup. Their features, however, absorb the density, scanning pattern, scale, and overlap of that setup, so a model trained on sparse outdoor LiDAR transfers poorly to indoor RGB-D fragments or scene-scale reconstructions.

Generalizable image-to-point-cloud registration. A separate line of work aims to reduce this dataset dependence. Pretrained cross-modal features, topology-aware reasoning, and unified matching architectures Wang et al. (2024); An et al. (2025; 2026); Goswami et al. (2026) improve generalization across scenes, particularly in indoor RGB-D registration, where point density and metric scale remain similar. The underlying pose pathway is nevertheless still based on 2D–3D matches or on dense predictions defined in the reference metric frame. When the sensing process changes, those predictions become unreliable. Training one specialist architecture on a mixture of datasets does not resolve the issue: without a representation that separates acquisition context from geometry, optimization tends to follow the dominant regime or to interfere across regimes.

Foundation priors and pose representations. Pretrained visual and point encoders (Oquab et al., 2024; Wu et al., 2024; Wang et al., 2025) provide transferable features that reduce reliance on dataset-specific appearance and local geometry. Such priors have enabled models trained on one corpus to retain useful representations on another, and are therefore a practical basis for scaling I2P registration beyond a single benchmark. Recent geometric foundation models further suggest that large-scale pretraining can expose shared structure across scenes and sensing conditions, rather than only within one acquisition setup. These results motivate building cross-sensor I2P systems on foundation backbones; UniReg follows this line by adopting pretrained image and point encoders as the shared front end of a single feed-forward registration network.

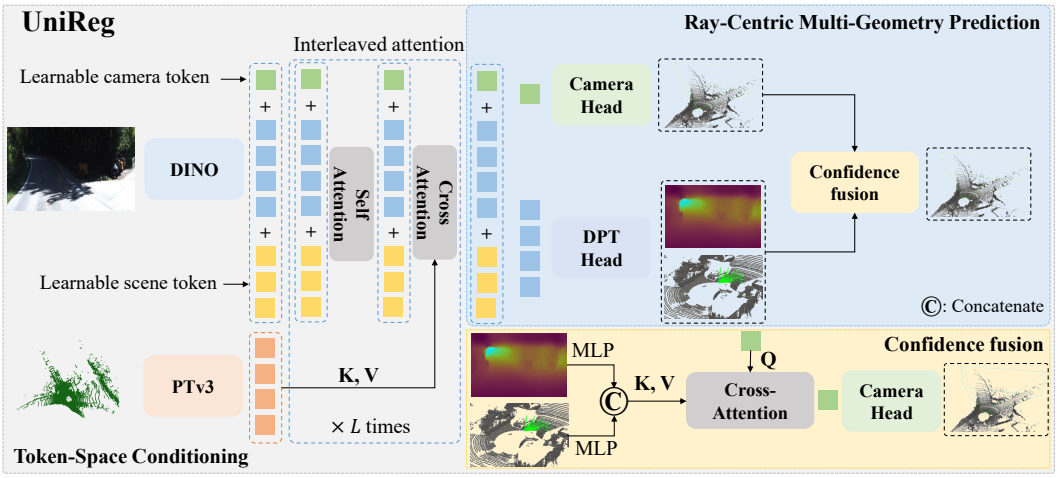

Figure 2: Overview of UniReg. Image and point-cloud encoders extract modality-specific tokens, which are fused by interleaved self- and cross-attention. A camera token aggregates pose evidence, and scene tokens represent acquisition factors. Ray-centric multi-geometry prediction decodes a camera pose, a dense ray map, and a dense depth map. Confidence-aware fusion updates the camera token with these predictions and regresses the final 6-DoF pose.

3 METHOD

3.1 OVERVIEW

Image-to-point-cloud registration estimates the rigid pose $\mathbf{T} = [\mathbf{R} \mid \mathbf{t}] \in \mathrm{SE}(3)$ of a query image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ with respect to a reference point cloud $\mathcal{P} = \{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^{N}$. The transformation maps a point from the point-cloud frame to the camera frame. The reference cloud may be sparse LiDAR or back-projected RGB-D geometry, and either a local fragment or a scene-level reconstruction. UniReg registers these inputs with one feed-forward network (Fig. 2).

To remain accurate across sensing regimes, the network follows three designs. Token-space conditioning keeps acquisition factors out of the geometric features. Ray-centric multi-geometry prediction replaces match- and metric-frame pathways with a camera pose, a dense ray map, and a dense depth map. Confidence-aware fusion turns their projective agreement into the final pose, so joint training does not average sensor-specific errors.

3.2 TOKEN-SPACE CONDITIONING

Image and point-cloud encoders. A pretrained DINOv2 transformer (Oquab et al., 2024) maps $\mathbf{I}$ to M patch tokens,

$$\mathbf{F}_{\mathrm{img}}^{(0)} = \{\mathbf{f}_j^{(0)} \in \mathbb{R}^d\}_{j=1}^{M}. \tag{1}$$

A Point Transformer V3 backbone (Wu et al., 2024) maps $\mathcal{P}$ to

$$\mathbf{F}_{\mathrm{pc}} = \{\mathbf{g}_i \in \mathbb{R}^d\}_{i=1}^{N'}, \tag{2}$$

which provides the keys and values of cross-modal attention.

Camera and scene tokens. A learnable camera token $\mathbf{f}_{\mathrm{cam}}^{(0)} \in \mathbb{R}^d$ and R scene tokens $\mathbf{F}_{\mathrm{scene}}^{(0)} = \{\mathbf{s}_r^{(0)}\}_{r=1}^{R}$ are prepended to the image tokens,

$$\mathbf{F}^{(0)} = \left[\mathbf{f}_{\mathrm{cam}}^{(0)}; \mathbf{F}_{\mathrm{scene}}^{(0)}; \mathbf{F}_{\mathrm{img}}^{(0)}\right]. \tag{3}$$

The camera token is the representation from which the pose is regressed. The scene tokens represent acquisition factors such as metric scale, point density, overlap, and imaging style. Placing these factors in dedicated tokens prevents them from being entangled with the geometric features, which would otherwise fit one sensing process rather than a sensor-agnostic mapping.

Interleaved attention. Each layer applies self-attention over the token sequence and then cross-attention to the point-cloud features,

$$\begin{aligned}\tilde{\mathbf{F}}^{(l)} &= \mathrm{SelfAttn}\left(\mathbf{F}^{(l-1)}\right),\\ \mathbf{F}^{(l)} &= \mathrm{CrossAttn}\left(\tilde{\mathbf{F}}^{(l)},\, \mathbf{F}_{\mathrm{pc}}\right).\end{aligned} \tag{4}$$

Self-attention exchanges information among the camera token, the scene tokens, and the image tokens. Cross-attention injects reference geometry without constructing explicit 2D–3D correspondences. Let $\mathbf{F}^{(L)}$ denote the output of the last layer and $\mathbf{f}_{\mathrm{cam}}$ its camera token.

3.3 Ray-Centric Multi-Geometry Prediction

From $\mathbf{F}^{(L)}$, UniReg decodes a camera pose from the camera token and a dense ray map and depth map from the image tokens, replacing sparse 2D–3D matches and dense metric-frame predictions.

Camera head. An MLP regresses camera pose from the camera token,

$$(\mathbf{R}, \mathbf{t}) = \mathrm{MLP}(\mathbf{f}_{\mathrm{cam}}). \tag{5}$$

Ray head. A DPT (Ranftl et al., 2021) predicts a Plücker ray map (Plücker, 1828) $\boldsymbol{\ell} = (\mathbf{d}, \mathbf{m}) \in \mathbb{R}^{H\times W\times 6}$ in the point-cloud frame and a confidence $\mathbf{C}^{\mathrm{ray}} \in \mathbb{R}^{H\times W}$, where $\mathbf{d_u}$ and $\mathbf{m_u}$ are the direction and moment at pixel $\mathbf{u}$.

Depth head. A second DPT predicts depth $\mathbf{D} \in \mathbb{R}^{H\times W}$ and confidence $\mathbf{C}^{\mathrm{depth}} \in \mathbb{R}^{H\times W}$.

The ray map is the primary geometric cue: a Plücker ray is fixed by camera projection and does not depend on the density or sampling of the reference cloud, remaining valid when LiDAR sparsity, RGB-D completeness, or scene extent changes. Depth supplies the metric distance along the ray; scale variation is absorbed by the scene tokens rather than by the ray direction.

3.4 Confidence-Aware Fusion

The pose $(\mathbf{R}, \mathbf{t})$ summarizes the cross-modal tokens, but it does not yet enforce agreement with the dense ray and depth predictions. Those predictions are the projective constraints that remain meaningful across sensors. Confidence-aware fusion injects them into the camera token, which is used as the query. At each pixel, the ray feature and the depth feature are scaled by their own confidences and concatenated,

$$\mathbf{v_u} = \Big[C^{\mathrm{ray}}_{\mathbf{u}}\, \phi_{\mathrm{r}}\big([\mathbf{d_u}; \mathbf{m_u}]\big);\, C^{\mathrm{depth}}_{\mathbf{u}}\, \phi_{\mathrm{d}}(D_{\mathbf{u}})\Big], \tag{6}$$

where ϕ_{r} and ϕ_{d} are two-layer MLPs that project features into a shared channel dimension. The updated representation and the final pose are

$$\mathbf{z}_{\mathrm{fuse}} = \mathrm{CrossAttn}\left(\mathbf{f}_{\mathrm{cam}},\, \{\mathbf{v_u}\}\right), \qquad (\mathbf{R}^\star, \mathbf{t}^\star) = \mathrm{MLP}\left(\mathbf{z}_{\mathrm{fuse}}\right). \tag{7}$$

This step makes joint training across acquisition setups constructive rather than conflicting. The ray term supplies the primary projective constraint, because its direction is fixed by the camera and does not follow the sampling of the reference cloud. Depth adds the metric range of the same pixel, but that range depends on scene scale and is less stable when the sensor changes. Separate confidences weight the two features by their own reliability: an inaccurate depth estimate cannot displace the ray-based update, and an inaccurate ray is attenuated in the same way. The final pose is therefore supported by projective agreement among the camera token, the rays, and the depth, while the confidence identifies the pixels that support this agreement. Sensor-specific errors remain in the low-confidence predictions instead of being transferred into $(\mathbf{R}^\star, \mathbf{t}^\star)$.

Pixel confidence is trained by

$$\mathcal{L}_{\mathrm{conf}}(c, \hat{\mathbf{y}}, \mathbf{y}) = c\,\|\hat{\mathbf{y}} - \mathbf{y}\|_2 - \alpha \log c. \tag{8}$$

For a ray, $\hat{\mathbf{y}} = (\mathbf{d_u}, \mathbf{m_u})$; for depth, $\hat{\mathbf{y}} = D_{\mathbf{u}}$. The first term increases the contribution of accurate predictions to $\mathbf{v_u}$; the second prevents the confidence from collapsing to zero.

3.5 Training Objective

The overall training loss is

$$\mathcal{L} = \lambda_{\mathrm{p}}\mathcal{L}_{\mathrm{pose}} + \lambda_{\mathrm{r}}\mathcal{L}_{\mathrm{ray}} + \lambda_{\mathrm{d}}\mathcal{L}_{\mathrm{depth}}, \tag{9}$$

with $\lambda_{\mathrm{p}} = 5$, $\lambda_{\mathrm{r}} = 1$, and $\lambda_{\mathrm{d}} = 1$. The pose term supervises the camera-head output and the pose after confidence-aware fusion,

$$\mathcal{L}_{\mathrm{pose}} = \|\mathbf{R}_{\mathrm{gt}} - \mathbf{R}\|_2 + \|\mathbf{t}_{\mathrm{gt}} - \mathbf{t}\|_2 + \|\mathbf{R}_{\mathrm{gt}} - \mathbf{R}^{\star}\|_2 + \|\mathbf{t}_{\mathrm{gt}} - \mathbf{t}^{\star}\|_2. \tag{10}$$

The dense terms apply the confidence loss on valid pixels,

$$\mathcal{L}_{\mathrm{ray}} = \mathcal{L}_{\mathrm{conf}}\big(C^{\mathrm{ray}}, \boldsymbol{\ell}, \boldsymbol{\ell}_{\mathrm{gt}}\big), \qquad \mathcal{L}_{\mathrm{depth}} = \mathcal{L}_{\mathrm{conf}}\big(C^{\mathrm{depth}}, \mathbf{D}, \mathbf{D}_{\mathrm{gt}}\big). \tag{11}$$

Ground-truth rays and depth are obtained by projecting the reference geometry into the image with $\mathbf{T}_{\mathrm{gt}}$ and the camera intrinsics. Joint optimization couples the token representation, the ray-centric predictions, and the confidence-weighted pose, which is the training condition under which heterogeneous data reinforce a shared geometric mapping.

4 Experiments

4.1 Implementation Details

The image encoder is a pretrained DINOv2 ViT-L/14 (Oquab et al., 2024). Every image is resized to a width of 518 pixels with its aspect ratio preserved, and the height is rounded to a multiple of 14, the patch size. The same rule is used for outdoor and indoor inputs, and the intrinsics are rescaled accordingly. The point encoder is a Point Transformer V3 (Wu et al., 2024) with embedding dimension 1024 and a base voxel size of 0.03 m. Image and point tokens are fused by 8 transformer layers that alternate between self-attention and cross-attention to the point features. Each layer uses 16 heads. A camera token and 16 scene tokens are prepended to the image tokens. Depth and ray maps are decoded by DPT heads from layers $\{1, 3, 5, 7\}$. The camera head is an MLP on the camera token, and the final pose is decoded again after confidence-aware fusion.

Point coordinates, depth, and translation are divided by a fixed scale before they enter the network: 5 for outdoor sets, and 1 for the indoor sets. Reported translation errors are converted back to meters. The model is trained jointly on all six sources with one set of weights. We use AdamW (Loshchilov & Hutter, 2017) with batch size 4, weight decay 0.05, and peak learning rate 5×10^{-5}. The learning rate warms up linearly from 1×10^{-8} over the first 5% of training and then follows a cosine schedule back to 1×10^{-8}. Training runs for 40 epochs.

4.2 Datasets

UniReg is trained and evaluated on six image-to-point-cloud sources. Four are standard benchmarks. The other two are constructed to supply the acquisition factors that those benchmarks do not cover: scene-level indoor geometry, and outdoor LiDAR under changing camera and LiDAR configurations. A single weight set is used for every test split. Specialist baselines are trained separately on the benchmark for which they are evaluated.

KITTI Odometry (Geiger et al., 2013). KITTI contains urban driving sequences with calibrated cameras and a 64-beam LiDAR. Sequences 0–8 are used for training and 9–10 for testing. Each frame provides left and right color images, and the reference cloud is corresponding LiDAR scan.

nuScenes (Caesar et al., 2020). Image–LiDAR pairs are built by aggregating neighboring LiDAR sweeps and pairing the accumulated cloud with the current camera frame. We use the official split of 850 training scenes and 150 test scenes.

CARLA (Dosovitskiy et al., 2017). Standard outdoor benchmarks fix the camera and the LiDAR. To vary the sensing process within the LiDAR regime, we collect a CARLA corpus of 3,072 runs over six towns. Each run freezes one sensor configuration and records 100 frames with the camera pose, the LiDAR pose, metric depth, and the LiDAR cloud. The configurations cross four horizontal fields of view ($60°$, $75°$, $90°$, $120°$), four resolutions (640×480, 1280×720, 1280×960, $1920 \times$

1080), two LiDAR beam counts (32 and 64), and four weather conditions (ClearNoon, ClearSunset, CloudyNoon, and SoftRainNoon).

RGB-D Scenes V2 (Lai et al., 2014). RGB-D Scenes V2 contains 11,427 RGB-D frames from 14 indoor scenes. For each scene, a point-cloud fragment is fused from every 25 consecutive depth frames, and one RGB image is sampled every 25 frames. Pairs with overlap below 30% are removed. Scenes 1–8 are used for training, 9–10 for validation, and 11–14 for testing. The reference geometry is a local fragment, so density is high but the map is incomplete outside that window.

7-Scenes (Glocker et al., 2013). 7-Scenes contains 46 RGB-D sequences from 7 indoor rooms. Pairs are built as for RGB-D Scenes V2 and are kept when their overlap is at least 30%. We follow the official sequence split. Relative to RGB-D Scenes V2, the viewpoint changes are larger and the retained overlap is often smaller.

ScanNet, scene level (Dai et al., 2017). Fragment benchmarks never ask the model to register an image to a complete room. We therefore pair each ScanNet color frame with the reconstructed scene mesh, rather than with a short-window depth fragment. The mesh is the cleaned reconstruction released with the scene, voxel-downsampled at 0.05 m, and is not cropped to the camera frustum. Frames with invalid poses are discarded, and the remaining frames are taken with a stride of 10. Scenes are sorted by name and split at a ratio of 85/15 into training and evaluation, with no shared scene between the two sides. The visible overlap is then whatever portion of the full room projects into the image. This set complements RGB-D Scenes V2 and 7-Scenes: the sensor is still RGB-D, but the reference changes from a local fragment to scene-level geometry.

4.3 Evaluation Metrics

We follow VP2P-Match (Zhou et al., 2023) and report the relative translation error (RTE, in meters), the relative rotation error (RRE, in degrees), and registration accuracy (Acc.). On outdoor datasets, including CARLA, Acc. is the percentage of pairs with RTE < 2 m and RRE $< 5^\circ$. No test pair is removed because its error is large.

Table 1: Image-to-point-cloud registration on KITTI and nuScenes. UniReg is trained jointly on all six sources and uses one set of weights for both columns. Other methods are trained and tested on the same individual dataset. Acc. is the percentage of pairs with RTE < 2 m and RRE $< 5^\circ$. † denotes methods that use an external depth estimator.

Method	KITTI			nuScenes		
	RTE(m)↓	RRE(°)↓	Acc(%)↑	RTE(m)↓	RRE(°)↓	Acc(%)↑
DeepI2P(3D) (Li & Lee, 2021)	4.06±3.54	24.73±31.69	3.77	2.88±2.12	20.65±12.24	2.26
DeepI2P(2D) (Li & Lee, 2021)	3.59±3.21	11.66±18.16	25.95	2.78±1.99	4.80±6.21	38.10
CorrI2P (Ren et al., 2022)	3.78±65.16	5.89±20.34	72.42	3.04±60.76	3.73±9.03	49.00
VP2P-Match (Zhou et al., 2023)	0.75±1.13	3.29±7.99	83.04	0.89±1.44	2.15±7.03	88.33
CFI2P (Yao et al., 2024)	0.63±1.19	1.72±2.58	97.15	0.86±1.47	1.83±1.28	95.23
FreeReg† (Wang et al., 2024)	0.95±1.05	2.06±3.21	91.68	-	-	-
CoFiI2P (Kang et al., 2024)	0.31±0.20	1.24±0.84	-	1.21±9.55	2.54±8.93	-
ICL (Li et al., 2025)	0.20±0.21	1.24±2.34	97.49	0.63±0.44	2.13±3.75	90.94
GraphI2P† (Bie et al., 2025)	0.32±0.81	1.65±1.32	99.61	0.49±1.22	1.73±1.63	99.48
UniReg	**0.10±0.06**	**0.59±1.58**	**99.91**	**0.39±0.35**	**0.69±1.37**	**99.54**

4.4 Comparison with Prior Methods

Outdoor datasets. Table 1 compares UniReg with prior I2P methods on real outdoor LiDAR. Each baseline is trained and tested on the same individual dataset, whereas UniReg uses one jointly trained weight set for both columns. On KITTI, UniReg reaches 0.10 m RTE, 0.59° RRE, and 99.91% Acc., halving ICL's translation error (0.20 m→0.10 m) and reducing RRE from 1.24° to 0.59°. On nuScenes, the same weights give 0.39 m RTE and 0.69° RRE, below CoFiI2P (1.21 m), ICL (0.63 m), and GraphI2P (0.49 m), with a clear rotation lead (0.69° against 1.73°–2.54°).

Indoor datasets. Table 2 evaluates the same UniReg weights on local RGB-D fragments and scene-level ScanNet meshes. On 7-Scenes, UniReg records 0.071 m RTE and 1.816° RRE against 0.076 m / 3.200° for CA-I2P. On ScanNet, RTE/RRE are 0.045 m / 1.116° versus 0.213 m / 4.798° for 2D3D-MATR. On RGB-D Scenes V2, UniReg is second in translation (0.069 m vs. 0.059 m for B2-3Dnet) but first in rotation (1.908° vs. 2.559°).

Table 2: Indoor registration with the same UniReg weights as Table 1. RGB-D Scenes V2 and 7-Scenes use local fragments. ScanNet uses the full scene mesh, not a depth fragment.

Method	RGB-D Scenes V2		7-Scenes		ScanNet	
	RTE(m)↓	RRE(°)↓	RTE(m)↓	RRE(°)↓	RTE(m)↓	RRE(°)↓
2D3D-MATR (Li et al., 2023)	0.077±0.066	3.026±2.649	0.079±0.076	3.291±3.058	0.213±0.224	4.798±10.253
B2-3Dnet (Cheng et al., 2025a)	**0.059±0.055**	2.600±2.410	0.076±0.073	3.232±2.988	-	-
CA-I2P (Cheng et al., 2025b)	0.061±**0.055**	2.559±2.190	0.076±0.069	3.200±2.825	-	-
UniReg	0.069±0.046	**1.908±1.324**	**0.071±0.068**	**1.816±2.785**	**0.045±0.035**	**1.116±2.497**

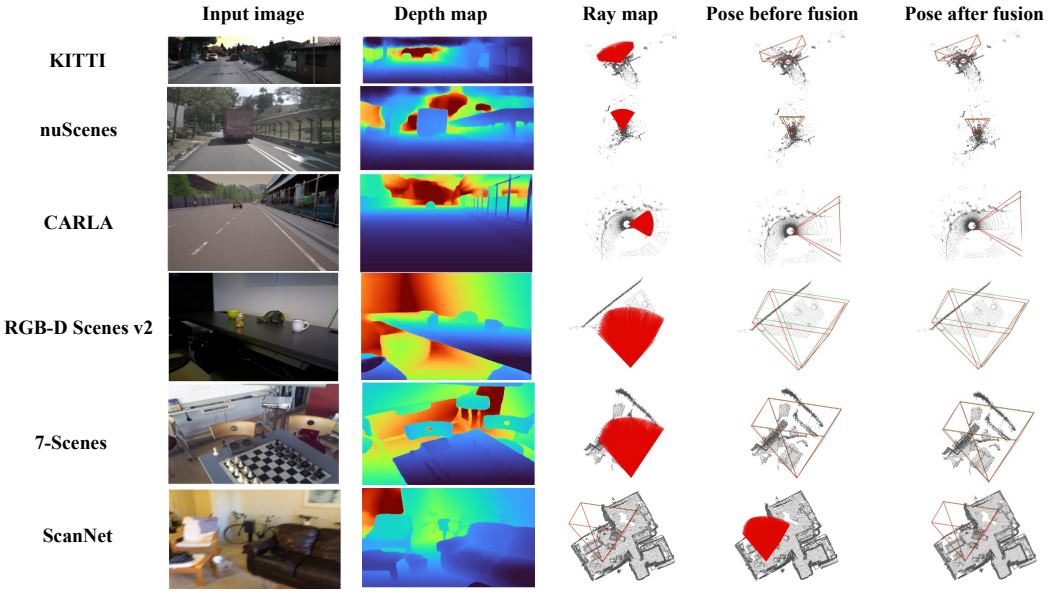

Figure 3: Qualitative results with one jointly trained UniReg model. Depth and rays are predicted densely; green/red frustums are GT/prediction before and after confidence-aware fusion. The fused pose remains tightly aligned across all six acquisition regimes.

Discussion. Across both tables, one jointly trained model matches or surpasses per-dataset specialists. Joint training therefore does not reduce to averaging specialist behaviors: the shared weights remain competitive wherever a column-specific model is allowed to overfit its own sensing setup. We attribute this to the three designs above. Scene tokens absorb acquisition context, ray-centric geometry provides projective constraints under changing density and overlap, and confidence-aware fusion keeps source-specific residuals out of the final pose.

4.5 Qualitative Results

Figure 3 shows qualitative results of the jointly trained UniReg model on all six sources. For each example we visualize the input image, the predicted depth, the predicted rays on the reference point cloud, and the camera frustums before and after fusion, with the ground truth in green and the prediction in red. Across outdoor LiDAR, CARLA, RGB-D fragments, and scene-level ScanNet, the predicted depth and rays are well aligned with the scene, and the fused pose closely matches the ground truth. No dataset-specific alignment is used to produce the figure.

4.6 Ablation Study

Effect of Ray and Depth Supervision. We first study the roles of ray and depth supervision by removing them individually or jointly, while keeping pose supervision in every row. As reported in Table 3, using neither ray nor depth increases KITTI RTE from 0.10 m to 0.27 m and RRE from 0.59° to 1.84°, with similar degradation on 7-Scenes and ScanNet. Rays alone recover most of the rotation (KITTI 0.73°, 7-Scenes 2.013°, ScanNet 1.348°), whereas depth alone is more helpful for translation on the larger maps (KITTI 0.13 m, ScanNet 0.058 m) but leaves higher RRE. On 7-Scenes, rays remain better on both metrics. The full objective with both heads is best in every column. These results indicate that rays and depth are complementary rather than interchangeable.

A ray is a line of sight and therefore provides a dense projective constraint on orientation, which remains valid when the reference sampling is sparse or incomplete. Depth supplies metric range along those rays and mainly improves translation once the orientation is constrained. Supervising both heads therefore yields accurate and stable pose estimates under joint indoor–outdoor training.

Table 3: Ablation of ray (R) and depth (D) supervision. Pose supervision is kept in every row. Rays carry the projective constraint; depth adds metric range.

R	D	KITTI		7-Scenes		ScanNet	
		RTE(m)↓	RRE(°)↓	RTE(m)↓	RRE(°)↓	RTE(m)↓	RRE(°)↓
✗	✗	0.27±0.18	1.84±3.12	0.134±0.109	3.528±4.417	0.112±0.086	2.714±3.806
✓	✗	0.16±0.11	0.73±1.79	0.088±0.076	2.013±2.941	0.064±0.050	1.348±2.591
✗	✓	0.13±0.09	1.26±2.35	0.097±0.081	2.706±3.512	0.058±0.047	1.926±3.114
✓	✓	**0.10±0.06**	**0.59±1.58**	**0.071±0.068**	**1.816±2.785**	**0.045±0.035**	**1.116±2.497**

Table 4: Ablation of scene tokens. "w/o" removes scene tokens; "w/" is the full model.

Scene tokens	KITTI		7-Scenes		ScanNet	
	RTE(m)↓	RRE(°)↓	RTE(m)↓	RRE(°)↓	RTE(m)↓	RRE(°)↓
w/o	0.23±0.13	1.34±2.39	0.123±0.092	2.962±3.751	0.103±0.078	2.288±3.471
w/	**0.10±0.06**	**0.59±1.58**	**0.071±0.068**	**1.816±2.785**	**0.045±0.035**	**1.116±2.497**

Effect of Scene Tokens. We next examine the contribution of scene tokens by removing them from the full model. As shown in Table 4, KITTI RTE/RRE rise from 0.10 m / 0.59° to 0.23 m / 1.34°, and 7-Scenes/ScanNet RTE from 0.071 m / 0.045 m to 0.123 m / 0.103 m. These results highlight the importance of scene tokens in UniReg. Under joint indoor–outdoor training, density, scale, overlap, and imaging style otherwise remain entangled in shared features and interfere with pose estimation. Scene tokens provide a dedicated pathway for this context, so the ray and depth heads can focus on projective geometry and registration improves on both outdoor and indoor benchmarks.

Table 5: Ablation of CARLA in the training mixture. "w/o" trains on the other five sources; "w/" is the full six-source model.

CARLA	KITTI		7-Scenes		ScanNet	
	RTE(m)↓	RRE(°)↓	RTE(m)↓	RRE(°)↓	RTE(m)↓	RRE(°)↓
w/o	0.17±0.12	1.08±2.16	0.080±0.072	2.047±2.912	0.052±0.040	1.271±2.548
w/	**0.10±0.06**	**0.59±1.58**	**0.071±0.068**	**1.816±2.785**	**0.045±0.035**	**1.116±2.497**

Effect of CARLA Training Data. Table 5 compares the full six-source model with a variant trained on the other five sources under the same optimizer budget. Removing CARLA increases KITTI RTE/RRE from 0.10 m / 0.59° to 0.17 m / 1.08°, while indoor changes remain small (7-Scenes 0.071→0.080 m, ScanNet 0.045→0.052 m). Varying FoV, resolution, and LiDAR beam count therefore strengthens the outdoor pathway without harming indoor accuracy, supporting that multi-source training helps once acquisition context is separated from the geometric pose pathway.

5 Conclusion

We presented UniReg, a single feed-forward network for image-to-point-cloud registration across outdoor LiDAR, indoor RGB-D fragments, and scene-level meshes. Instead of learning a specialist for each acquisition setup, UniReg factors the problem into three components: scene tokens that absorb density, scale, and sensing style; ray-centric multi-geometry prediction that uses projective rays as the primary cue and depth as complementary metric range; and confidence-aware fusion that turns agreement among rays, depth, and the camera token into the final pose. Trained jointly on six sources with one set of weights, UniReg outperforms per-dataset specialists on KITTI, nuScenes, RGB-D Scenes V2, 7-Scenes, and scene-level ScanNet. Ablations further show that ray and depth supervision are complementary, that scene tokens are essential for joint training, and that sensor-varying CARLA data strengthens the outdoor pathway without harming indoor accuracy. These results suggest that cross-sensor I2P registration is better addressed by disentangling acquisition context from geometry than by training a separate model for every sensing regime.

AI USE STATEMENT

In preparing this manuscript, we used AI-assisted writing tools (including large language models) to help refine English wording, improve clarity, and reorganize draft text in the introduction, method, experiments, and conclusion. All technical content, including the problem formulation, method design, implementation details, experimental protocols, reported results, and scientific claims, was conceived, verified, and finalized by the authors. AI tools were not used to generate experimental numbers, fabricate citations, or replace author responsibility for the correctness of the work. The authors have carefully checked the final manuscript and take full responsibility for its content.

REFERENCES

Pei An, Jiaqi Yang, Muyao Peng, You Yang, Qiong Liu, Jie Ma, and Liangliang Nan. Top-i2p: Explore open-domain image-to-point cloud registration using topology relationship. In *IJCAI*, pp. 674–683, 2025.

Pei An, Junfeng Ding, Jiaqi Yang, Yulong Wang, Jie Ma, and Liangliang Nan. Hg-i2p: Bridging modalities for generalizable image-to-point-cloud registration via heterogeneous graphs. *arXiv preprint arXiv:2603.27969*, 2026.

Lin Bie, Shouan Pan, Siqi Li, Yining Zhao, and Yue Gao. Graphi2p: Image-to-point cloud registration with exploring pattern of correspondence via graph learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22161–22171, 2025.

Mark Billinghurst, Adrian Clark, Gun Lee, et al. A survey of augmented reality. *Foundations and Trends® in Human–Computer Interaction*, 8(2-3):73–272, 2015.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

Zhixin Cheng, Jiacheng Deng, Xinjun Li, Baoqun Yin, and Tianzhu Zhang. Bridge 2d-3d: Uncertainty-aware hierarchical registration network with domain alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2491–2499, 2025a.

Zhixin Cheng, Jiacheng Deng, Xinjun Li, Xiaotian Yin, Bohao Liao, Baoqun Yin, Wenfei Yang, and Tianzhu Zhang. Ca-i2p: Channel-adaptive registration network with global optimal selection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27739–27749, 2025b.

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

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *IEEE robotics & automation magazine*, 13(2):99–110, 2006.

Mengdan Feng, Sixing Hu, Marcelo H Ang, and Gim Hee Lee. 2d3d-matchnet: Learning to match keypoints across 2d image and 3d point cloud. In *2019 International Conference on Robotics and Automation*, pp. 4790–4796. IEEE, 2019.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. Real-time rgb-d camera relocalization. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 173–179. IEEE, 2013.

Prajnan Goswami, Tianye Ding, Feng Liu, and Huaizu Jiang. Unicorrn: Unified correspondence transformer across 2d and 3d. *arXiv preprint arXiv:2605.04044*, 2026.

Shuhao Kang, Youqi Liao, Jianping Li, Fuxun Liang, Yuhao Li, Xianghong Zou, Fangning Li, Xieyuanli Chen, Zhen Dong, and Bisheng Yang. Cofii2p: Coarse-to-fine correspondences-based image to point cloud registration. *IEEE Robotics and Automation Letters*, 2024.

Kevin Lai, Liefeng Bo, and Dieter Fox. Unsupervised feature learning for 3d scene labeling. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3050–3057. IEEE, 2014.

Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep*n*p: An accurate $O(n)$ solution to the p*n*p problem. *International journal of computer vision*, 81:155–166, 2009.

Jiaxin Li and Gim Hee Lee. Deepi2p: Image-to-point cloud registration via deep classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15960–15969, 2021.

Minhao Li, Zheng Qin, Zhirui Gao, Renjiao Yi, Chenyang Zhu, Yulan Guo, and Kai Xu. 2d3d-matr: 2d-3d matching transformer for detection-free registration between images and point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14128–14138, 2023.

Xinjun Li, Wenfei Yang, Jiacheng Deng, Zhixin Cheng, Xu Zhou, and Tianzhu Zhang. Implicit correspondence learning for image-to-point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16922–16931, 2025.

Xinjun Li, Wenfei Yang, Zhixin Cheng, Jiacheng Deng, Fei Wang, Chen Qian, and Tianzhu Zhang. Rayi2p: Learning rays for image-to-point cloud registration. In *International Conference on Learning Representations*, volume 2026, pp. 44559–44576, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=a68SUt6zFt`. Featured Certification.

Julius Plücker. *Analytisch-geometrische Entwicklungen*, volume 2. GD Baedeker, 1828.

Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, Slobodan Ilic, Dewen Hu, and Kai Xu. Geotransformer: Fast and robust point cloud registration with geometric transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9806–9821, 2023.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12179–12188, October 2021.

Siyu Ren, Yiming Zeng, Junhui Hou, and Xiaodong Chen. Corri2p: Deep image-to-point cloud registration via dense correspondence. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1198–1208, 2022.

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*, pp. 4938–4947, 2020.

Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulo, Richard Newcombe, Peter Kontschieder, and Vasileios Balntas. Orienternet: Visual localization in 2d public maps with neural matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21632–21642, 2023.

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*, pp. 8922–8931, 2021.

Haiping Wang, Yuan Liu, Bing Wang, Yujing Sun, Zhen Dong, Wenping Wang, and Bisheng Yang. Freereg: Image-to-point cloud registration leveraging pretrained diffusion models and monocular depth estimators. In *International Conference on Learning Representations*, volume 2024, pp. 28484–28507, 2024.

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, June 2025.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4840–4851. IEEE, 2024.

Gongxin Yao, Yixin Xuan, Yiwei Chen, and Yu Pan. Quantity-aware coarse-to-fine correspondence for image-to-point cloud registration. *IEEE Sensors Journal*, 24(20):33826–33837, 2024.

Hao Yu, Fu Li, Mahdi Saleh, Benjamin Busam, and Slobodan Ilic. Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *Advances in Neural Information Processing Systems*, 34:23872–23884, 2021.

Kaiwen Yuan, Zhenyu Guo, and Z Jane Wang. Rggnet: Tolerance aware lidar-camera online calibration with geometric deep learning and generative model. *IEEE Robotics and Automation Letters*, 5(4):6956–6963, 2020.

Junsheng Zhou, Baorui Ma, Wenyuan Zhang, Yi Fang, Yu-Shen Liu, and Zhizhong Han. Differentiable registration of images and lidar point clouds with voxelpoint-to-pixel matching. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51166–51177, 2023.

A Appendix

A.1 Dataset Protocols

This appendix details the construction and splits of the six sources used by UniReg. Unless otherwise noted, all sources share the same feed-forward model and the same evaluation metrics as in the main paper. Reported translation errors are always converted back to meters after the network-internal scale normalization.

A.1.1 KITTI Odometry

We use the KITTI Odometry benchmark (Geiger et al., 2013) with calibrated cameras and a 64-beam Velodyne LiDAR. Sequences 0–8 are used for training and sequences 9–10 for testing. Both the left and right color cameras (P2 and P3) are included. For each frame, the reference geometry is the corresponding LiDAR scan. During training and evaluation, a random rigid perturbation is applied to the point cloud, and the ground-truth pose is the inverse of that perturbation. Outdoor LiDAR clouds are voxel-downsampled at 0.1 m before being fed to the network.

A.1.2 nuScenes

We follow the official nuScenes split (Caesar et al., 2020) with 850 training scenes and 150 test scenes. For each camera frame, neighboring LiDAR sweeps are aggregated into a denser reference cloud and paired with the current RGB image. The same outdoor rigid perturbation protocol as on KITTI is applied. This source differs from KITTI not only in scene layout, but also in the sensing process: the cloud is an accumulated sweep rather than a single 64-beam scan.

A.1.3 CARLA

Standard outdoor benchmarks fix both the camera and the LiDAR. To introduce controlled variation within the LiDAR regime, we collect a CARLA (Dosovitskiy et al., 2017) corpus of 3,072 runs over six towns: Town01, Town02, Town03, Town04, Town05, and Town10HD, with 512 runs in each town. Each run freezes one sensor configuration and records 100 synchronized frames containing the RGB image, camera pose, LiDAR pose, metric depth, and LiDAR cloud. We index frames with a stride of three. Town05 is held out entirely for evaluation; the other five towns are used for training.

The sensor configurations cross:

- horizontal fields of view: $60°$, $75°$, $90°$, $120°$;
- image resolutions: 640×480, 1280×720, 1280×960, 1920×1080;
- LiDAR beam counts: 32 and 64;
- weather conditions: ClearNoon, ClearSunrise/ClearSunset, CloudyNoon, and SoftRainNoon.

Camera intrinsics are determined by the field of view and the image resolution. This corpus keeps the registration objective fixed while changing projective geometry and cloud density.

A.1.4 RGB-D Scenes V2

RGB-D Scenes V2 (Lai et al., 2014) contains 11,427 RGB-D frames from 14 indoor scenes. For each scene, a point-cloud fragment is fused from every 25 consecutive depth frames, and one RGB image is sampled on the same stride. Image–fragment pairs with overlap below 30% are removed. Scenes 1–8 are used for training, scenes 9–10 for validation, and scenes 11–14 for testing. The reference geometry is therefore a local fragment: density is high, but the map is incomplete outside the fused window.

A.1.5 7-Scenes

7-Scenes (Glocker et al., 2013) contains 46 RGB-D sequences from 7 indoor rooms. We build fragment pairs in the same way as for RGB-D Scenes V2 and keep pairs whose overlap is at least 30%. We follow the official sequence split for training and testing. Relative to RGB-D Scenes V2, viewpoint changes are larger and the retained overlap is often smaller, which makes orientation recovery more demanding.

A.1.6 ScanNet, scene level

Fragment benchmarks never ask the model to register an image to a complete room. We therefore construct a scene-level ScanNet (Dai et al., 2017) I2P setting. Each color frame is paired with the cleaned scene mesh released with the sequence (`_vh_clean_2.ply`), rather than with a short-window depth fragment. The mesh is voxel-downsampled at 0.05 m and is *not* cropped to the camera frustum. Frames with invalid poses are discarded, and the remaining frames are taken with a stride of 10. Scenes are sorted by name and split at a ratio of 85/15 into training and evaluation, with no scene shared between the two sides. The visible overlap is then whatever portion of the full room projects into the image. This source keeps the RGB-D sensing family, but changes the reference from a local fragment to scene-scale geometry.

A.1.7 Shared preprocessing

All images are resized to a width of 518 pixels with aspect ratio preserved; the height is rounded to a multiple of 14, and the intrinsics are rescaled accordingly. Point coordinates, depth, and translation are divided by a fixed scale before entering the network: 5 for KITTI, nuScenes, and CARLA, and 1 for RGB-D Scenes V2, 7-Scenes, and ScanNet. Reported RTE values are converted back to meters after inference.

A.2 Additional Implementation Details

A.2.1 Network configuration

The image encoder is a pretrained DINOv2 ViT-L/14 (Oquab et al., 2024). The point encoder is Point Transformer V3 (Wu et al., 2024) with embedding dimension 1024 and base voxel size 0.03 m. Image and point tokens are fused by 8 transformer layers that alternate self-attention over the token sequence and cross-attention to the point features, with 16 heads in each layer. A camera token and 16 scene tokens are prepended to the image tokens. Indoor and outdoor samples index separate scene-token banks, so the two acquisition regimes do not share the same context parameters. Dense depth and ray heads are DPT decoders (Ranftl et al., 2021) reading fused features from layers $\{1, 3, 5, 7\}$. The camera head is an MLP on the camera token; after confidence-aware fusion, the same head type decodes the updated token into the final pose. We do not use an explicit point-map head.

A.2.2 Pose encoding and ray representation

Camera poses are encoded as absolute translation and quaternion rotation. The predicted transformation maps a point from the point-cloud frame to the camera frame. Ray heads predict Plücker coordinates $(\mathbf{d}, \mathbf{m})$ in the point-cloud frame, where $\mathbf{d}$ is ℓ_2-normalized and $\mathbf{m}$ is unconstrained. Ground-truth rays are obtained by placing the camera center in the point-cloud frame and taking the normalized direction of each pixel ray, with $\mathbf{m} = \mathbf{C} \times \mathbf{d}$. Ground-truth depth is obtained by projecting the reference geometry with the ground-truth pose and camera intrinsics.

A.2.3 Data augmentation

For outdoor sources (KITTI, nuScenes, CARLA), the reference cloud is perturbed by a random rigid transform with translation ranges $t_x, t_z \in [-10, 10]$ m, $t_y \in [-1, 1]$ m, and rotation ranges yaw $\in [-180°, 180°]$, pitch $\in [-60°, 60°]$, roll $\in [-15°, 15°]$. For indoor sources (RGB-D Scenes V2, 7-Scenes, ScanNet), the same rotation ranges are used, while translation amplitudes are scaled by 0.2. The same perturbation protocol is applied at validation time, so a fixed dataset index does not correspond to a fixed relative pose across runs. Training images are additionally augmented by color jitter; no color jitter is used at test time.

A.3 Additional Qualitative Results

Figure 3 in the main paper shows one example from each of the six sources. Here we provide additional examples under the same jointly trained weights and the same visualization protocol. For each example we show the input image, the predicted depth, the predicted equal-length rays on the reference cloud, and the camera frustums before and after confidence-aware fusion. Ground-truth frustums are drawn in green and predicted frustums in red. Outdoor examples use a top-down view of the LiDAR cloud; indoor examples use a slightly tilted camera-aligned view so that the frustum remains readable as a 3D volume rather than a flat triangle.

The additional examples support three observations already suggested by the main figure. First, predicted depth follows image layout on both sparse outdoor LiDAR and dense indoor RGB-D inputs. Second, rays remain a coherent fan from the camera center even when the reference cloud is sparse, incomplete, or much larger than the visible frustum. No dataset-specific post-processing or manual alignment is used.

A.4 Limitations

UniReg improves cross-sensor I2P registration with one feed-forward model, but several limitations remain.

Dependence on calibrated intrinsics. The method assumes known camera intrinsics after image resizing. We do not address joint estimation of intrinsics and extrinsics, nor strong distortion beyond the pinhole model used in our datasets.

Sensor coverage of training. Joint training covers sparse outdoor LiDAR, RGB-D fragments, scene-level meshes, and CARLA configurations with varied FoV, resolution, and beam count. This is

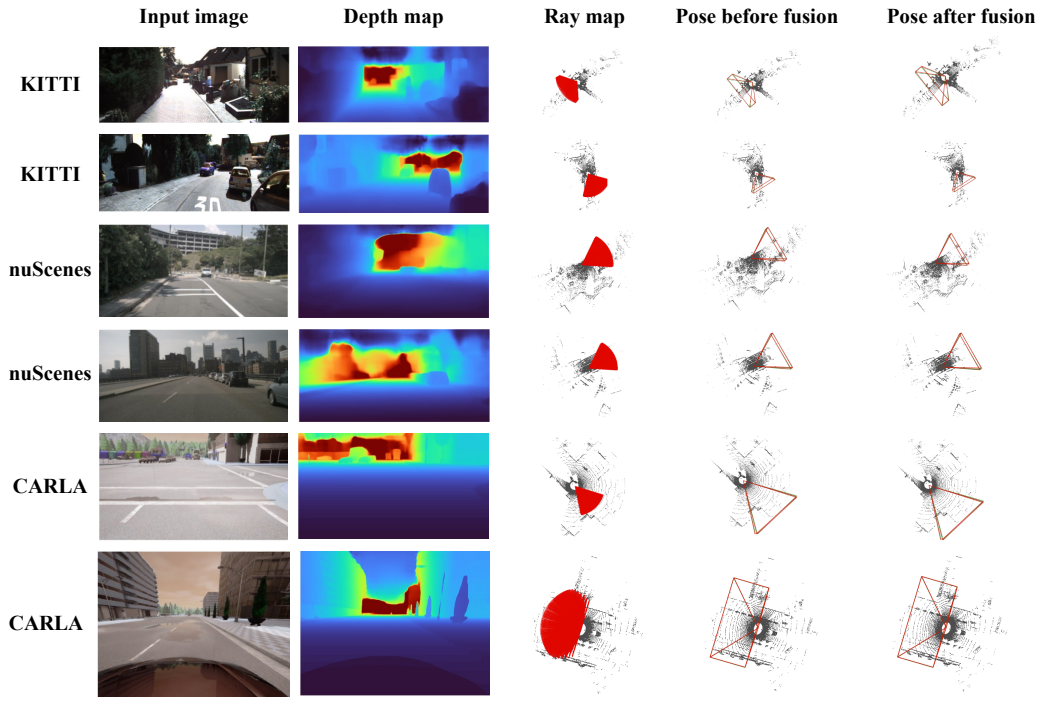

Figure 4: Additional outdoor qualitative results with the same UniReg weights. Rows are KITTI, nuScenes, and CARLA examples not shown in the main paper. Columns are input image, predicted depth, predicted rays, pose before fusion, and pose after fusion.

broader than a single specialist setup, but it does not exhaust all real sensors (e.g., solid-state LiDAR patterns, thermal cameras, or drastically different indoor scanning pipelines). Transfer to sensing processes far outside the training mixture may still require additional data or light adaptation.

Compute footprint. UniReg uses a large image encoder and a strong point backbone, together with dense DPT heads. The resulting accuracy comes with a higher memory and runtime cost than lightweight correspondence pipelines. We leave distillation and lighter backbones for future work.

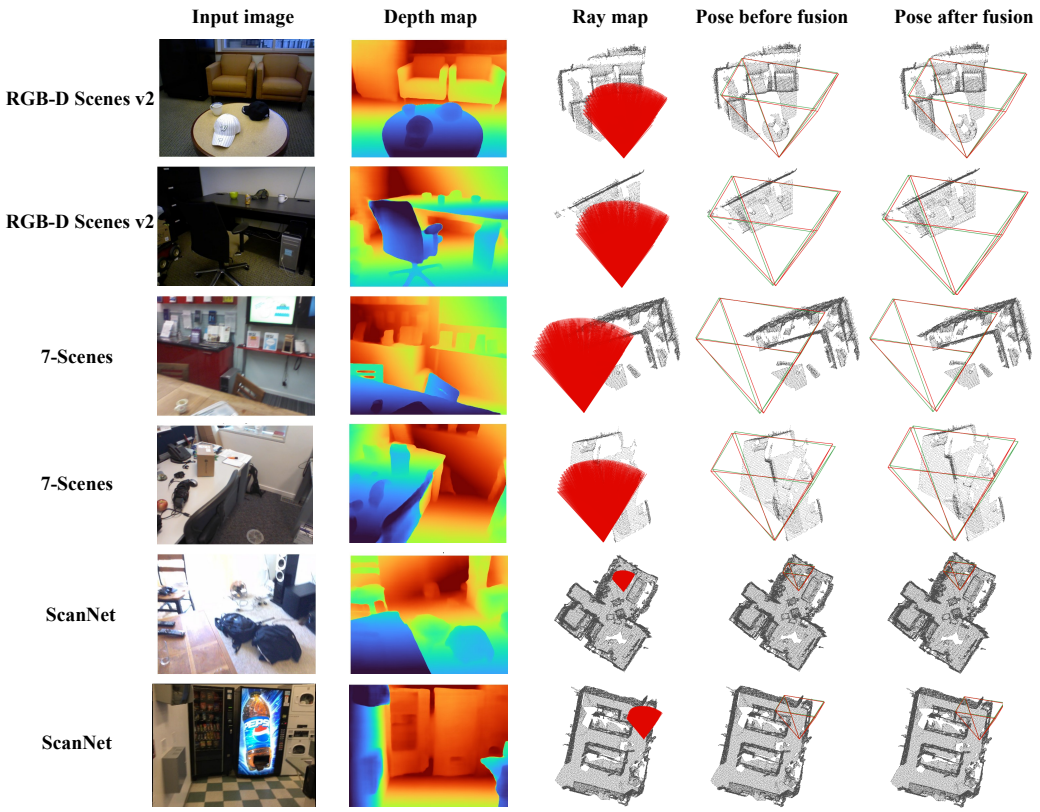

Figure 5: Additional indoor qualitative results with the same UniReg weights. Rows are RGB-D Scenes V2, 7-Scenes, and scene-level ScanNet examples not shown in the main paper. The fused pose remains close to the ground truth on both local fragments and full-room meshes.