PGCR: Pose–Geometry Coupled Reasoning for Image-to-Point Cloud Registration

Xinjun Li[1], Wenfei Yang[1], Yihan Chen[1], Zhixin Cheng[1], Shifeng Zhang[2], Xu Zhou[2], and Tianzhu Zhang[1,⋆]

[1] School of Information Science and Technology / National Key Laboratory of Deep Space Exploration, University of Science and Technology of China
[2] Sangfor Technologies

Abstract. Image-to-point cloud registration aims to estimate the 6-DoF camera pose of a query image with respect to a 3D point cloud, which remains challenging due to the heterogeneous nature of dense visual appearance and sparse geometric structure. Most existing learning-based methods follow an overlap–then–registration paradigm, where pose estimation is restricted to a predicted overlapping region. Such a sequential decomposition is inherently fragile, as inaccurate overlap prediction may discard geometrically consistent regions and irreversibly constrain pose optimization. In this work, we reformulate image-to-point cloud registration as a pose–geometry coupled reasoning problem, where camera pose and scene geometry are treated as interdependent variables. Based on this formulation, we propose PGCR, a unified framework that estimates camera pose by explicitly reasoning over dense cross-modal geometry. PGCR jointly predicts camera pose together with dense geometric representations and grounds pose inference in global geometric consistency. To enable effective pose–geometry coupling, we introduce a Geometry-aware Pose Refinement mechanism to enforce prediction-level consistency between pose and geometry, and a Pose-guided Progressive Refinement strategy to adapt cross-modal interaction according to intermediate pose estimates. Extensive experiments on two widely used outdoor and two indoor benchmarks demonstrate that PGCR consistently outperforms prior state-of-the-art methods across all evaluation metrics while remaining computationally efficient.

Keywords: Image-to-point cloud registration · cross-modal geometry

1 Introduction

Image-to-point cloud registration aims to estimate the 6-degree-of-freedom (6-DoF) camera pose of a query 2D image with respect to a pre-built 3D point cloud. This task is fundamental to numerous computer vision applications, including visual localization [41], 3D reconstruction [16], and SLAM [18]. Compared to single-modality registration, image-to-point cloud registration is inherently more

⋆ Corresponding author

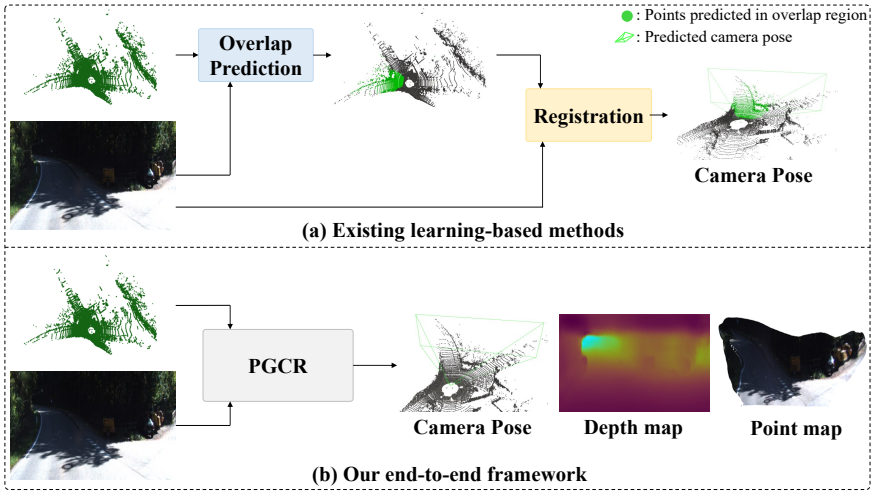

Fig. 1: Comparison between existing learning-based methods and our end-to-end framework. (a) Conventional methods decouple overlap prediction and pose estimation, making them vulnerable to early errors under sparse or ambiguous overlap. (b) Our proposed method, **PGCR**, jointly reasons about pose and geometry, allowing visibility and correspondence to emerge implicitly through geometric consistency. This coupled formulation enables global reasoning and avoids irreversible error propagation.

challenging due to the heterogeneous nature of the two modalities: images provide dense appearance cues on regular 2D grids, whereas point clouds capture sparse geometric structures in irregular 3D space. Bridging this modality gap to achieve accurate and robust registration remains a long-standing problem.

Existing learning-based methods [20,24,27,30,36,50] typically follow an overlap–then–registration paradigm, as shown in Figure 1(a). In these methods, features are extracted independently from 2D images and 3D point clouds, followed by an explicit overlap prediction stage that identifies potentially visible regions between the two modalities. A registration stage is then performed by restricting optimization to the predicted overlapping region. While this decomposition reduces the search space, it introduces a critical vulnerability: errors in overlap prediction may discard geometrically consistent regions or introduce misleading correspondences. This issue becomes particularly severe in outdoor environments, where sparse and noisy LiDAR observations, together with repetitive structures, make overlap prediction highly ambiguous.

In this work, we revisit the formulation of image-to-point cloud registration from a geometric reasoning perspective [46]. Unlike single-modality registration, this task exhibits a fundamental bidirectional dependency: camera pose directly determines which subset of the 3D scene is visible, while observed scene geometry simultaneously constrains the set of feasible pose hypotheses. This mutual dependency implies that pose estimation and geometric reasoning are inherently

coupled. However, existing pipelines explicitly decouple these two aspects by first estimating overlap and then regressing pose within the predicted overlapping region. Motivated by this observation, we formulate image-to-point cloud registration as a tightly coupled pose–geometry reasoning problem. Instead of treating overlap prediction as an explicit prerequisite, we directly model camera pose and scene geometry as interdependent latent variables within a unified framework, where visibility and correspondence are not predefined but emerge implicitly through geometric consistency during joint optimization. As a result, the model can reason globally over the entire scene, leverage partially visible but geometrically informative structures, and avoid error propagation caused by inaccurate overlap prediction. Based on this formulation, we propose **PGCR**, an end-to-end **P**ose–**G**eometry **C**oupled **R**easoning framework that estimates camera pose by explicitly reasoning over dense cross-modal geometry. Specifically, PGCR jointly predicts camera pose together with a dense depth map in the image domain and a global point map in the 3D domain, as shown in Figure 1(b). These geometric predictions are not auxiliary targets; rather, they introduce structured geometric constraints that continuously regularize pose estimation. Concretely, the predicted dense geometry provides spatially distributed reprojection signals across the entire image and point cloud, enabling the model to evaluate pose hypotheses through global geometric consistency rather than relying on sparse or local correspondences. This dense constraint distribution improves optimization stability, enlarges the effective convergence basin, and reduces sensitivity to partial overlap or noisy observations. Moreover, as pose and geometry are optimized in a mutually conditioned manner, visibility is determined implicitly by geometric alignment instead of being explicitly filtered at an early stage. This global reasoning mechanism avoids discarding potentially useful regions too early and allows the model to retain informative structures throughout optimization.

While coupling pose estimation with dense geometry improves robustness, it exposes an important optimization gap. In typical architectures, camera pose is predicted from a compact global token. Although this design is computationally efficient, it introduces two limitations. First, spatially distributed geometric evidence must be compressed into a global representation, which inevitably discards fine-grained structural constraints. Second, pose and dense geometry are predicted through separate output heads, and nothing explicitly enforces their geometric consistency. To address these issues, we introduce a **Geometry-aware Pose Refinement (GPR)** mechanism that establishes a feedback loop between pose and geometry: inconsistencies between the current pose estimate and dense geometric predictions generate corrective signals that refine the pose. By grounding pose estimation in spatially distributed geometric evidence and aligning it with predicted scene structure, GPR improves both geometric fidelity and optimization stability, leading to more accurate and robust registration.

Beyond prediction level consistency, robust cross-modal alignment still faces a complementary challenge at the interaction level. Blindly performing dense 2D–3D interaction over the entire scene not only introduces substantial geo-

metric noise but also leads to unnecessary computational overhead. Therefore, effective pose–geometry coupling requires the interaction mechanism itself to evolve according to the current pose estimate. To this end, PGCR incorporates a **Pose-guided Progressive Refinement (PPR)** strategy that conditions cross-modal interaction on intermediate pose hypotheses. Starting from a coarse pose estimate, the model progressively focuses attention on increasingly plausible and visible 3D regions. By reducing irrelevant interactions and concentrating computation on geometrically consistent areas, this progressive alignment improves both optimization stability and computational efficiency.

Together, PGCR provides a unified pose–geometry coupled reasoning framework for image-to-point cloud registration. Extensive experiments on two widely used outdoor and two indoor datasets demonstrate that PGCR consistently outperforms prior state-of-the-art methods across all evaluation metrics, validating the effectiveness, robustness, and generality of the proposed formulation.

Our main contributions are summarized as follows:

- We propose PGCR, a unified framework for image-to-point cloud registration that formulates camera pose estimation as a pose–geometry coupled reasoning problem, enabling joint inference of camera pose and scene geometry through dense 2D–3D interaction without explicit overlap prediction.
- We introduce a Geometry-aware Pose Refinement mechanism and a Pose-guided Progressive Refinement strategy, which together make pose–geometry coupling accurate and stable by enforcing pose–geometry consistency at the prediction level and adapting cross-modal interaction at the interaction level.
- Extensive experiments on two widely used outdoor and two indoor datasets show that PGCR consistently outperforms prior state-of-the-art methods across all evaluation metrics, while remaining computationally efficient.

2 Related Works

In this section, We briefly overview methods that are related to image registration, point cloud registration and image-to-point cloud registration.

Image registration. Image registration under varying illumination, viewpoint, or scene appearance has been a long-standing topic in computer vision. Classical methods rely on sparse keypoint detection and matching, either using handcrafted descriptors such as SIFT [31] and ORB [38], or learned features like SuperPoint [15], D2-Net [19], R2D2 [37], and SuperGlue [40]. While keypoint-based pipelines remain efficient, their performance degrades in textureless or repetitive regions. Recent detector-free approaches [42, 43, 47, 51] directly learn dense correspondence fields through coarse-to-fine attention mechanisms, achieving more reliable matching across complex conditions. These methods have inspired subsequent registration frameworks beyond the single-modality setting.

Point cloud registration. Rigid alignment between 3D point sets is another extensively studied problem. Early works focus on feature-based matching using geometric descriptors such as FPFH [39] or PPF [17], followed by iterative optimization or RANSAC [21]. Learning-based approaches [1, 14, 32, 34, 49] have

since demonstrated improved robustness by encoding local and global geometric structures through neural networks. CoFiNet [49] introduces a coarse-to-fine refinement pipeline without explicit keypoint detection, while GeoTransformer [34] jointly learns correspondence and pose estimation within a transformer framework. Despite their success, these techniques operate purely within 3D geometry and are not directly applicable to heterogeneous 2D–3D settings.

Image-to-point cloud registration. Image-to-point cloud registration poses additional challenges due to the significant modality gap between dense 2D visual signals and sparse 3D geometry. A prevalent paradigm constructs 2D–3D correspondences and estimates the camera pose through PnP-style solvers [26]. Early cross-modal pipelines such as 2D3D-MatchNet [20] extract independent keypoints and features for each modality, but their reliance on hand-crafted detectors often results in poor geometric consistency. Subsequent learning-based approaches [2,4–8,10,12,13,24,27,29,30,36,50] focus on bridging modality discrepancies through feature-level fusion. CorrI2P [36] predicts overlapping regions before matching dense features, while voxel-based designs such as VP2P-Match [50] integrate differentiable PnP solvers for end-to-end optimization. CoFiI2P [24] adapts the LoFTR [43] architecture for cross-modal correspondence learning, but still relies on explicit overlap prediction that is unreliable under LiDAR sparsity. Implicit regression methods [30] bypass explicit correspondence estimation, yet their geometric reasoning remains limited by weak spatial priors. More recent frameworks like GraphI2P [2] incorporate auxiliary depth estimation to reduce domain gaps, but at the cost of heavy computation.

3 Method

3.1 Overview

Given a query RGB image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and a pre-built 3D point cloud $\mathcal{P} = \{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^{N}$, our goal is to estimate the 6-DoF camera pose $\mathbf{T} = [\mathbf{R} \mid \mathbf{t}] \in SE(3)$ that maps points from the world frame to the camera frame.

PGCR formulates image-to-point cloud registration as a *pose–geometry coupled reasoning* problem. Instead of predicting camera pose from a compact latent representation or relying on explicit correspondence construction, PGCR explicitly maintains a pose hypothesis throughout the network and uses it to drive dense geometric interaction. Crucially, geometric structures predicted during inference are not treated as auxiliary outputs; they serve as explicit carriers of geometric evidence that continuously inform pose estimation.

As illustrated in Figure 2, PGCR operates as a closed-loop reasoning framework. Image and point cloud features are first encoded and fused through iterative cross-modal attention, conditioned on a learnable camera token representing the current pose hypothesis. Intermediate pose estimates are periodically extracted and used to progressively restrict cross-modal interaction to geometrically plausible and visible 3D regions, enabling robust reasoning under sparse observations and partial overlap. From the fused representations, the framework

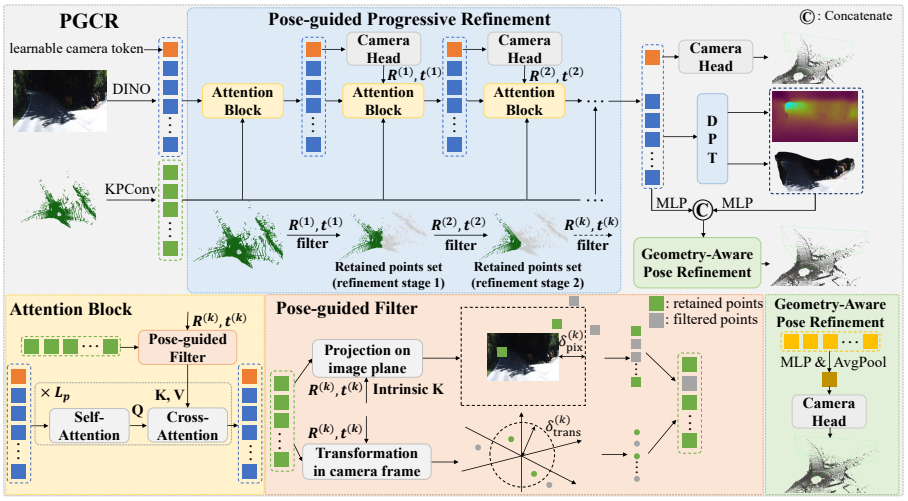

Fig. 2: Overview of the proposed PGCR framework. PGCR formulates image-to-point cloud registration as a pose–geometry coupled reasoning process. Given an input image and a 3D point cloud, image and point features are encoded and iteratively fused through cross-modal attention, conditioned on a learnable camera token representing the current pose hypothesis. Pose-guided Progressive Refinement leverages intermediate pose estimates to progressively restrict 2D–3D interaction to geometrically plausible regions. The framework jointly predicts camera pose and dense geometric structures, including a depth map and a point map. A Geometry-aware Pose Refinement module explicitly enforces consistency between the estimated pose and predicted geometry, closing the loop between pose inference and geometric reasoning.

jointly predicts camera pose together with dense geometric structures, including a depth map in the image domain and a point map in the 3D domain. A geometry-aware pose refinement module then explicitly enforces consistency between the estimated pose and the predicted geometry, closing the loop between pose inference and geometric reasoning.

3.2 2D–3D Feature Encoding

Image Encoding. The input image $\mathbf{I}$ is encoded by a DINO transformer [33] to produce a set of image patch tokens:

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

where M denotes the number of image patches and d is the feature dimension. A learnable camera token $\mathbf{f}_{\text{cam}}^{(0)}$ is prepended to the image tokens to represent a global pose hypothesis and to aggregate scene-level geometric evidence.

Point Cloud Encoding. The 3D point cloud $\mathcal{P}$ is processed using KPConv [44], generating geometry-aware point features:

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

These features capture local shape and spatial context, and serve as 3D keys and values in subsequent cross-modal attention layers.

3.3 Transformer-Based Pose–Geometry Fusion

PGCR employs stacked transformer layers to perform dense pose–geometry interaction between image and point cloud features. At layer l, the image tokens (including the camera token) are updated through intra-modal self-attention:

$$\mathbf{F}_{\text{img}}^{(l)} = \text{SelfAttn}(\mathbf{F}_{\text{img}}^{(l-1)}), \tag{3}$$

followed by cross-modal attention that aggregates geometric information:

$$\mathbf{F}_{\text{img}}^{(l)} = \text{CrossAttn}(\mathbf{F}_{\text{img}}^{(l)}, \mathbf{F}_{\text{pc}}). \tag{4}$$

Through iterative fusion, image tokens progressively acquire 3D geometric context, while the camera token evolves to encode the current pose hypothesis conditioned on global scene structure. This iterative attention mechanism enables dense, bidirectional information exchange between modalities without relying on explicit correspondence construction.

3.4 Pose-guided Progressive Refinement

Dense attention over entire point cloud is computationally inefficient and dominated by irrelevant regions outside the field of view. To address this, we introduce a **Pose-guided Progressive Refinement** strategy that dynamically narrows attention toward geometrically consistent 3D regions.

Intermediate Pose Prediction. Every L_p layers, the current camera token is passed through a camera head (MLP) to predict an intermediate pose:

$$(\mathbf{R}^{(k)}, \mathbf{t}^{(k)}) = \text{MLP}_{\text{cam}}(\mathbf{f}_{\text{cam}}^{(L_p \cdot k)}), \tag{5}$$

where $k \in \{1, \ldots, L_c\}$ indexes the refinement stage. These intermediate poses provide auxiliary supervision that stabilizes training and generate visibility cues to guide attention in subsequent layers.

Pose-guided Filter. At each refinement stage k, a subset of geometrically plausible points $\mathcal{M}^{(k)}$ is selected based on the intermediate pose. A 3D point $\mathbf{p}_i$ is retained if its projection lies within a relaxed image boundary:

$$[u_i^{(k)}, v_i^{(k)}, 1]^\top \propto \mathbf{K}(\mathbf{R}^{(k)}\mathbf{p}_i + \mathbf{t}^{(k)}), \tag{6}$$

$$-\delta_{\text{pix}}^{(k)} \leq u_i^{(k)} \leq W + \delta_{\text{pix}}^{(k)}, \quad -\delta_{\text{pix}}^{(k)} \leq v_i^{(k)} \leq H + \delta_{\text{pix}}^{(k)}, \tag{7}$$

or if its transformed distance to the camera center satisfies:

$$\|\mathbf{R}^{(k)}\mathbf{p}_i + \mathbf{t}^{(k)}\|_2 \leq \delta_{\text{trans}}^{(k)}. \tag{8}$$

The tolerance thresholds $\delta_{\text{pix}}^{(k)}$ and $\delta_{\text{trans}}^{(k)}$ decay across refinement stages:

$$\delta^{(k)} = \delta^{(1)} \cdot \eta^{k-1}, \quad \eta \in (0, 1), \tag{9}$$

allowing coarse global reasoning at early stages and fine-grained geometric consistency at later stages.

Refined Cross-Attention. Given $\mathcal{M}^{(k)}$, the subsequent cross-attention layers attend only to filtered point subset:

$$\mathbf{F}_{\text{img}}^{(l)} = \text{CrossAttn}\big(\mathbf{F}_{\text{img}}^{(l)}, \mathbf{F}_{\text{pc}}[\mathcal{M}^{(k)}]\big), \tag{10}$$

allowing transformer to focus computation on visible, informative regions. This stage-wise refinement forms a feedback loop between pose estimation and attention modulation, leading to efficient and geometrically consistent convergence.

3.5 Multi-Task Geometric Prediction

From the fused representations, PGCR jointly predicts camera pose and dense geometric structures:

- **Camera Head:** An MLP regresses the camera pose $(\mathbf{R}, \mathbf{t})$ from the refined camera token, with rotation represented as a normalized quaternion.
- **Depth Head:** A Dense Prediction Transformer (DPT) [35] outputs a per-pixel depth map $\mathbf{D} \in \mathbb{R}^{H \times W}$.
- **Point Map Head:** A parallel DPT head predicts a per-pixel 3D point map $\mathbf{M} \in \mathbb{R}^{H \times W \times 3}$ in the world frame.

These predictions are supervised jointly to enforce geometric consistency across pose, depth, and point estimation.

3.6 Geometry-Aware Pose Refinement

While transformer-based fusion enables effective pose–geometry interaction, compressing spatially distributed geometric evidence into a single camera token inevitably discards fine-grained constraints. Moreover, pose and dense geometry are predicted through separate output heads, without any mechanism explicitly enforcing their geometric consistency. To address these issues and keep pose estimation continuously grounded in dense geometry, we introduce a **Geometry-aware Pose Refinement** module. First, all predicted modalities are projected into a shared embedding space through MLPs:

$$\mathbf{E}_{\text{depth}} = \text{MLP}_{\text{depth}}(\mathbf{D}), \quad \mathbf{E}_{\text{point}} = \text{MLP}_{\text{point}}(\mathbf{M}), \tag{11}$$

while the final image tokens $\mathbf{F}_{\text{img}}^{(L_p L_c)}$ are projected and upsampled to obtain $\mathbf{E}_{\text{img}}$. The final camera token $\mathbf{f}_{\text{cam}}^{(L_p L_c)}$ is projected and repeated to match spatial dimensions, forming $\mathbf{E}_{\text{cam}}$. All embeddings are concatenated and fused through another MLP followed by average pooling:

$$\mathbf{z}_{\text{fuse}} = \text{AvgPool}\big(\text{MLP}_{\text{fuse}}([\mathbf{E}_{\text{depth}}, \mathbf{E}_{\text{point}}, \mathbf{E}_{\text{img}}, \mathbf{E}_{\text{cam}}])\big), \tag{12}$$

and the refined pose is obtained as:

$$(\mathbf{R}, \mathbf{t}) = \text{MLP}_{\text{pose}}(\mathbf{z}_{\text{fuse}}). \tag{13}$$

This module explicitly enforces consistency between the estimated pose and predicted geometry, forming a closed geometric feedback loop that improves robustness under sparse and ambiguous observations.

3.7 Loss Functions

The overall training objective is:

$$\mathcal{L} = \lambda_p \mathcal{L}_{\text{pose}} + \lambda_d \mathcal{L}_{\text{depth}} + \lambda_m \mathcal{L}_{\text{point}},$$

where $\lambda_p = 5$, $\lambda_d = 1$, and $\lambda_m = 1$. The pose loss supervises both intermediate and final predictions:

$$\mathcal{L}_{\text{pose}} = \sum_{k=1}^{L_c} (\|\mathbf{R}_{gt} - \mathbf{R}^{(k)}\|_2 + \|\mathbf{t}_{gt} - \mathbf{t}^{(k)}\|_2) + \|\mathbf{R}_{gt} - \mathbf{R}\|_2 + \|\mathbf{t}_{gt} - \mathbf{t}\|_2,$$

where we adopt the continuous 6D representation for rotation learning on $\mathbf{SO}(3)$. The depth/point map losses are:

$$\mathcal{L}_{\text{depth}} = \|\mathbf{D}_{gt} - \mathbf{D}\|_2, \quad \mathcal{L}_{\text{point}} = \|\mathbf{M}_{gt} - \mathbf{M}\|_2.$$

Since ground-truth depth/point maps are obtained by LiDAR projection using ground-truth poses, losses are computed only on valid projected pixels using a validity mask.

4 Experiments

4.1 Implementation Details

The image encoder follows a DINO-style transformer [33]. For outdoor datasets, input images are resized while preserving aspect ratio, resulting in resolutions of 154×518 for KITTI and 154×308 for nuScenes. For indoor datasets, images are resized to 476×630 to preserve fine-grained geometric details. Each image is divided into non-overlapping 14×14 patches to generate image tokens. The point cloud encoder adopts KPConv [44]. For outdoor scenes, a voxel size of 0.15 m is used, while for indoor scenes, we employ a finer voxel size of 0.025 m to better capture small-scale structures. Both image and point features are projected into

Table 1: Image-to-point cloud registration accuracy on the KITTI and nuScenes benchmarks. † denotes methods that rely on external depth estimation models. PGCR consistently achieves the best performance across all metrics on both datasets.

Method	KITTI			nuScenes		
	RTE(m)↓	RRE(°)↓	Acc(%)↑	RTE(m)↓	RRE(°)↓	Acc(%)↑
DeepI2P(3D) [27]	4.06±3.54	24.73±31.69	3.77	2.88±2.12	20.65±12.24	2.26
DeepI2P(2D) [27]	3.59±3.21	11.66±18.16	25.95	2.78±1.99	4.80±6.21	38.10
CorrI2P [36]	3.78±65.16	5.89±20.34	72.42	3.04±60.76	3.73±9.03	49.00
VP2P-match [50]	0.75±1.13	3.29±7.99	83.04	0.89±1.44	2.15±7.03	88.33
CFI2P [48]	0.63±1.19	1.72±2.58	97.15	0.86±1.47	1.83±1.28	95.23
FreeReg† [45]	0.95±1.05	2.06±3.21	91.68	-	-	-
CoFiI2P [24]	0.31±0.20	1.24±0.84	-	1.21±9.55	2.54±8.93	-
ICL [30]	0.20±0.21	1.24±2.34	97.49	0.63±0.44	2.13±3.75	90.94
GraphI2P† [2]	0.32±0.81	1.65±1.32	99.61	0.49±1.22	1.73±1.63	99.48
Ours	**0.15±0.15**	**0.44±0.34**	**100.00**	**0.43±0.42**	**0.88±1.16**	**99.51**

Table 2: Performance on indoor image-to-point cloud benchmarks. PGCR significantly outperforms prior methods on both RGB-D Scenes V2 and 7-Scenes, despite being trained under the same unified formulation as outdoor datasets.

Method	RGB-D Scenes v2		7-Scenes	
	RTE(m)↓	RRE(°)↓	RTE(m)↓	RRE(°)↓
2D3D-MATR [28]	0.077±0.066	3.026±2.649	0.079±0.076	3.291±3.058
B2-3Dnet [9]	0.059±0.055	2.600±2.410	0.076±0.073	3.232±2.988
CA-I2P [11]	0.061±0.055	2.559±2.190	0.076±0.069	3.200±2.825
Ours	**0.035±0.023**	**2.215±1.207**	**0.049±0.029**	**2.622±1.603**

a shared latent space and fused through $L_p L_c = 8$ transformer layers alternating between self- and cross-attention. Each transformer block maintains a learnable camera token, which encodes the current pose hypothesis and is iteratively updated through pose–geometry interaction. Intermediate poses are predicted every $L_p = 2$ layers, resulting in $L_c = 4$ progressive refinement stages. For outdoor datasets, the initial pixel tolerance $\delta_{\text{pix}}^{(1)}$ and translation tolerance $\delta_{\text{trans}}^{(1)}$ are set to 80 pixels and 8 m, respectively. For indoor datasets, we use $\delta_{\text{pix}}^{(1)} = 160$ pixels and $\delta_{\text{trans}}^{(1)} = 1\,\text{m}$. All tolerances decay by a factor of $\eta = 0.5$ across refinement stages. The entire framework is trained end-to-end using AdamW with a batch size of 8 and an initial learning rate of 5×10^{-5}. Training is performed for 20 epochs on a single NVIDIA RTX 3090 GPU.

4.2 Datasets

We evaluate PGCR on two widely used outdoor benchmarks and two challenging indoor benchmarks, covering diverse sensing conditions and scene geometries. **KITTI Odometry [22].** The KITTI odometry dataset contains 22 driving sequences captured in urban environments, among which 11 sequences provide

ground-truth calibration. Following common practice, sequences 0–8 are used for training and 9–10 for testing. To simulate realistic sensor misalignment, we apply random 2D translations within $\pm 10\,\mathrm{m}$ on the ground plane and random yaw rotations. All images are resized to 154×518, and each point cloud is uniformly downsampled to 40,960 points.

nuScenes [3]. We construct image–LiDAR pairs using official SDK by aggregating LiDAR scans from neighboring frames and pairing them with images from the current frame. We follow the official train/test split with 850 training and 150 testing scenes. The same perturbation strategy as KITTI is applied. Images are resized to 154×308, and each point cloud is downsampled to 40,960 points.

RGB-D Scenes V2 [25]. RGB-D Scenes V2 contains 11,427 RGB-D frames from 14 indoor scenes. For each scene, we fuse a point cloud fragment from every 25 consecutive depth frames and sample one RGB image every 25 frames. Image–point cloud pairs with at least 30% overlap are retained. Scenes 1–8 are used for training, 9–10 for validation, and 11–14 for testing, resulting in 1,748 training, 236 validation, and 497 testing pairs.

7-Scenes [23]. The 7-Scenes dataset consists of 46 RGB-D sequences from 7 indoor environments. We follow the same pairing procedure as above and preserve pairs with at least 50% overlap. The official split is adopted, yielding 4,048 training, 1,011 validation, and 2,304 testing pairs. Compared to prior benchmarks, this setting includes richer viewpoint changes and smaller overlap ratios.

4.3 Evaluation Metrics

We follow the evaluation protocol of VP2P-Match [50] and report Relative Translation Error (RTE), Relative Rotation Error (RRE), and registration accuracy (Acc). Acc is defined as the percentage of test samples with RTE $< 2\,\mathrm{m}$ and RRE $< 5^\circ$ for outdoor datasets. All test pairs are evaluated without discarding high-error samples to provide a comprehensive assessment of robustness.

4.4 Comparison with State-of-the-art Methods

Outdoor Datasets. Table 1 reports quantitative results on KITTI and nuScenes. PGCR consistently achieves the best performance across all metrics on both datasets. On KITTI, PGCR attains an average RTE of $0.15\,\mathrm{m}$ and RRE of 0.44°, surpassing the previous best method ICL [30] by 25.0% and 64.5%, respectively. Notably, it achieves a perfect registration accuracy of 100%, indicating robust convergence across all test cases. These gains are significant given the sparse LiDAR observations and repetitive structures that challenge overlap-dependent pipelines. On nuScenes, which features diverse scenes and more complex motion patterns, PGCR maintains strong performance with an RTE of $0.43\,\mathrm{m}$ and RRE of 0.88°. Compared with the strongest baseline GraphI2P [2], further reduces translation and rotation errors by 12.2% and 49.1%, demonstrating robustness under large viewpoint changes and partial overlap. These results validate the effectiveness of pose–geometry coupled reasoning in outdoor environments.

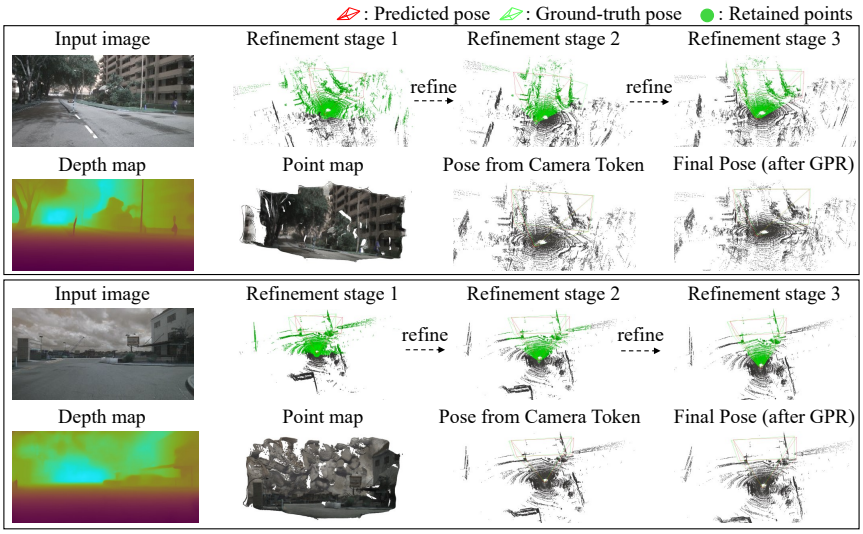

Fig. 3: Qualitative visualization of pose–geometry coupled reasoning. We show the input image, predicted depth and point maps, intermediate pose refinement stages, and the final refined pose. As refinement progresses, the estimated pose converges toward the ground truth, while the retained 3D points gradually concentrate within the true visible region. The final Geometry-aware Pose Refinement further improves alignment by enforcing consistency between pose and dense geometry.

Indoor Datasets. Table 2 presents results on RGB-D Scenes V2 and 7-Scenes. Despite being designed without indoor-specific heuristics, PGCR significantly outperforms prior methods on both datasets. This improvement is attributed to its unified formulation, which jointly reasons about pose and dense geometry without relying on explicit overlap estimation. The strong performance across both indoor and outdoor benchmarks highlights the generality of proposed framework.

4.5 Qualitative Results

Figure 3 visualizes the pose–geometry coupled reasoning process. We show the input image, predicted depth and point maps, intermediate pose refinement stages, and final refined pose. As refinement progresses, estimated pose gradually converges toward the ground truth, while the retained 3D points become increasingly concentrated within the true visible region. The final geometry-aware refinement further improves alignment by enforcing consistency between predicted geometry and pose. These visualizations confirm that PGCR performs pose estimation as a geometry-grounded reasoning process rather than a single-step regression.

Video Demonstration. We provide a supplementary video demonstrating full-sequence registration results over 1,587 consecutive image–point cloud pairs. The video visualizes complete camera trajectories and includes challenging scenarios

Table 3: Average inference time per image–point cloud pair (RTX 3090). Our method achieves the fastest inference speed. PPR improves both accuracy and efficiency.

Method	CorrI2P	VP2P-match	CoFiI2P	ICL	Ours (w/o PPR)	Ours
Inference Time (s)	8.96	0.19	0.18	0.24	0.26	**0.17**

such as sparse observations, severe viewpoint changes, partial overlap, and repetitive structures. This long-horizon evaluation further validates the robustness and temporal consistency of the proposed pose–geometry coupled formulation.

4.6 Inference Efficiency

Table 3 reports average inference time on a single RTX 3090 GPU. PGCR achieves the fastest runtime among recent transformer-based approaches while maintaining state-of-the-art accuracy. Removing Pose-guided Progressive Refinement increases runtime, confirming its role in reducing redundant computation by focusing attention on geometrically relevant regions. Overall, the unified formulation eliminates costly overlap prediction and external optimization steps, enabling both high accuracy and low latency.

Table 4: Ablation of the Geometry-aware Pose Refinement (GPR). GPR improves pose accuracy by enforcing pose–geometry consistency.

GPR	KITTI		nuScenes	
	RTE(m)↓	RRE(°)↓	RTE(m)↓	RRE(°)↓
w/o	0.17±0.18	0.49±0.36	0.50±0.66	1.06±1.33
w	**0.15±0.15**	**0.44±0.34**	**0.43±0.42**	**0.88±1.16**

4.7 Ablation Study

We conduct comprehensive ablation studies on KITTI and nuScenes to disentangle the contribution of each proposed component and to better understand their roles in improving registration accuracy and stability.

Effect of Geometry-aware Pose Refinement. We first evaluate the effectiveness of the proposed Geometry-aware Pose Refinement (GPR) module by removing it from the full model. As reported in Table 4, the absence of GPR leads to noticeable degradation in both translation and rotation accuracy on both datasets. By explicitly feeding back geometric consistency between the predicted dense structure and the estimated pose, GPR provides corrective signals that refine pose estimation in a geometry-aware manner. This explicit geometric feedback effectively reduces pose ambiguity and stabilizes optimization, resulting in

consistently lower RTE and RRE. These results confirm that incorporating geometric reasoning directly into pose refinement is crucial for accurate and robust image-to-point cloud registration.

Effect of Pose-guided Progressive Refinement. We further investigate the proposed Pose-guided Progressive Refinement (PPR) strategy, with quantitative results summarized in Table 5. Removing PPR ("w/o PPR") increases both translation and rotation errors, indicating that direct one-shot alignment is insufficient for precise cross-modal registration under large viewpoint and modality gaps. Introducing intermediate supervision already leads to steady performance improvements, as stage-wise optimization imposes progressive geometric constraints and alleviates optimization instability. As the refinement proceeds from Stage-1 to Stage-4, both RTE and RRE consistently decrease, demonstrating that progressive refinement effectively stabilizes pose estimation and prevents error accumulation. When pose-guided filter is further incorporated (Full PPR), the model achieves the best performance across all stages. The pose-guided filter suppresses attention to background, occluded, or geometrically inconsistent regions, allowing the network to focus on reliable 3D structures during refinement. Together with intermediate supervision, PPR enables robust stage-by-stage pose correction, yielding accurate and stable convergence. These quantitative improvements are further supported by qualitative visualizations in Figure 3.

Table 5: Ablation of the Pose-guided Progressive Refinement (PPR). Progressive refinement improves accuracy, and the pose-guided filter further enhances performance.

Method	KITTI		nuScenes	
	RTE(m)↓	RRE(°)↓	RTE(m)↓	RRE(°)↓
w/o PPR	0.33±0.26	0.94±1.19	0.74±0.73	1.66±2.55
Intermediate supervision (without pose-guided filter)				
Stage-1	0.35±0.31	0.96±0.79	0.77±0.76	1.34±1.68
Stage-2	0.29±0.18	0.88±0.70	0.71±0.73	1.29±1.57
Stage-3	0.26±0.17	0.85±0.66	0.66±0.73	1.26±1.55
Stage-4	0.25±0.16	0.82±0.64	0.67±0.73	1.24±1.55
Full PPR (intermediate supervision + pose-guided filter)				
Stage-1	0.35±0.47	0.98±0.80	0.81±0.86	1.70±1.99
Stage-2	0.27±0.35	0.74±0.55	0.70±0.73	1.54±1.70
Stage-3	0.20±0.21	0.50±0.40	0.58±0.70	1.20±1.47
Stage-4	**0.17±0.18**	**0.49±0.36**	**0.50±0.66**	**1.06±1.33**

Impact of Multi-task Supervision. We further analyze the contribution of depth and point map supervision, with results shown in Table 6. Removing both dense geometric heads ("w/o D & P") yields the weakest performance, indicating that pose supervision alone provides insufficient dense constraints for cross-modal registration. Enabling either branch individually already improves accuracy: depth supervision ("w/ D only") brings larger gains than point map supervision ("w/ P only") on both datasets, suggesting that metric-scale depth

cues are particularly effective for stabilizing relative pose estimation. Combining both heads ("w/ D & P") consistently achieves the best results, outperforming all single-branch variants on every metric. This indicates that depth and point map predictions provide complementary geometric signals (depth emphasizes continuous metric structure, while point maps offer explicit 3D correspondences) and their joint supervision encourages geometrically consistent dense representations that regularize pose estimation and reduce ambiguity.

Table 6: Ablation on depth (D) and point map (P) supervision. ✓/✗ indicate whether the corresponding prediction head is supervised during training.

D	P	KITTI		nuScenes	
		RTE(m)↓	RRE(°)↓	RTE(m)↓	RRE(°)↓
✗	✗	0.21±0.25	0.60±0.53	0.57±0.68	1.44±1.81
✗	✓	0.20±0.24	0.57±0.51	0.55±0.71	1.31±1.56
✓	✗	0.18±0.21	0.53±0.42	0.52±0.68	1.18±1.49
✓	✓	**0.17±0.18**	**0.49±0.36**	**0.50±0.66**	**1.06±1.33**

5 Conclusion

We presented a unified, pose–geometry coupled reasoning framework for image-to-point cloud registration that integrates dense geometric reasoning, explicit pose-level feedback, and progressive refinement within a single transformer architecture. By introducing Geometry-aware Pose Refinement and Pose-guided Progressive Refinement, the proposed method effectively addresses key challenges in cross-modal alignment, including pose ambiguity, visibility inconsistency, and optimization instability, without relying on explicit overlap prediction. Extensive experiments on both outdoor and indoor benchmarks demonstrate that our approach consistently outperforms previous state-of-the-art methods across all evaluation metrics, while maintaining high computational efficiency. Furthermore, the strong performance across diverse environments and extreme conditions highlights the robustness and generalization capability of the proposed framework. We believe this work provides a principled and extensible solution for accurate cross-modal registration, and offers valuable insights for future research in geometry-driven multi-modal perception.

6 Acknowledgements

This work was supported by the Scientific and Application Research Project (First Batch) of the Tianwen-2 Mission under China's Planetary Exploration Program (No. TW2-01-001).

References

1. Bai, X., Luo, Z., Zhou, L., Fu, H., Quan, L., Tai, C.L.: D3feat: Joint learning of dense detection and description of 3d local features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6359–6367 (2020)
2. Bie, L., Pan, S., Li, S., Zhao, Y., Gao, Y.: 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)
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11631 (2020)
4. Cheng, Z.: Mask 2d 3d adaptive dual masked autoencoder network for image to point cloud registration (March 2026). https://doi.org/10.13140/RG.2.2.29988.33923
5. Cheng, Z., Chen, Y., Tao, X., Zhang, T., et al.: Fs i2p: A hierarchical focus-sweep registration network with dynamically allocated depth (March 2026). https://doi.org/10.13140/RG.2.2.19922.00968
6. Cheng, Z., Deng, J., Li, X., Liao, B., Liu, L., Yin, X., Yin, B., Zhang, T.: Glass: Geometry-aware local alignment and structure synchronization network for 2d-3d registration (2026)
7. Cheng, Z., Deng, J., Li, X., Liu, L., Yin, X., Yin, B., Zhang, T.: B2-3d++: Uncertainty-aware hierarchical registration network with domain alignment. Authorea Preprints (2025)
8. Cheng, Z., Deng, J., Li, X., Yin, B., Zhang, T.: Bridge 2d-3d: Uncertainty-aware hierarchical registration network with domain alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2491–2499 (2025)
9. Cheng, Z., Deng, J., Li, X., Yin, B., Zhang, T.: Bridge 2d-3d: Uncertainty-aware hierarchical registration network with domain alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2491–2499 (2025)
10. Cheng, Z., Deng, J., Li, X., Yin, X., Liao, B., Yin, B., Yang, W., Zhang, T.: Ca-i2p: Channel-adaptive registration network with global optimal selection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 27739–27749 (October 2025)
11. Cheng, Z., Deng, J., Li, X., Yin, X., Liao, B., Yin, B., Yang, W., Zhang, T.: Ca-i2p: Channel-adaptive registration network with global optimal selection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27739–27749 (2025)
12. Cheng, Z., Liao, B., Deng, J., Yin, X., Li, X., Chen, Y., Yin, B., Zhang, T.: Rethinking 2d-3d registration: A novel network for high-value zone selection and representation consistency alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 39052–39063 (June 2026)
13. Cheng, Z., Yin, X., Deng, J., Liao, B., Chen, Y., Zhou, X., Yin, B., Zhang, T.: Adaptive agent selection and interaction network for image-to-point cloud registration. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 3335–3343 (3 2026). https://doi.org/10.1609/aaai.v40i5.37329, https://doi.org/10.1609/aaai.v40i5.37329

14. Deng, H., Birdal, T., Ilic, S.: Ppfnet: Global context aware local features for robust 3d point matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 195–205 (2018)
15. DeTone, D., Malisiewicz, T., Rabinovich, A.: 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)
16. Dong, Z., Liang, F., Yang, B., Xu, Y., Zang, Y., Li, J., Wang, Y., Dai, W., Fan, H., Hyyppä, J., et al.: Registration of large-scale terrestrial laser scanner point clouds: A review and benchmark. ISPRS Journal of Photogrammetry and Remote Sensing **163**, 327–342 (2020)
17. Drost, B., Ulrich, M., Navab, N., Ilic, S.: Model globally, match locally: Efficient and robust 3d object recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 998–1005. Ieee (2010)
18. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping: part i. IEEE robotics & automation magazine **13**(2), 99–110 (2006)
19. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-net: A trainable cnn for joint description and detection of local features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8092–8101 (2019)
20. Feng, M., Hu, S., Ang, M.H., Lee, G.H.: 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)
21. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
22. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
23. Glocker, B., Izadi, S., Shotton, J., Criminisi, A.: Real-time rgb-d camera relocalization. In: 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 173–179. IEEE (2013)
24. Kang, S., Liao, Y., Li, J., Liang, F., Li, Y., Zou, X., Li, F., Chen, X., Dong, Z., Yang, B.: Cofii2p: Coarse-to-fine correspondences-based image to point cloud registration. IEEE Robotics and Automation Letters (2024)
25. Lai, K., Bo, L., Fox, D.: Unsupervised feature learning for 3d scene labeling. In: 2014 IEEE International Conference on Robotics and Automation (ICRA). pp. 3050–3057. IEEE (2014)
26. Lepetit, V., Moreno-Noguer, F., Fua, P.: Epnp: An accurate $O(n)$ solution to the pnp problem. International journal of computer vision **81**, 155–166 (2009)
27. Li, J., Lee, G.H.: 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)
28. Li, M., Qin, Z., Gao, Z., Yi, R., Zhu, C., Guo, Y., Xu, K.: 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)
29. Li, X., Yang, W., Cheng, Z., Deng, J., Wang, F., Qian, C., Zhang, T.: Rayi2p: Learning rays for image-to-point cloud registration. In: The Fourteenth International Conference on Learning Representations (2026), `https://openreview.net/forum?id=arfeGsDWoq`

30. Li, X., Yang, W., Deng, J., Cheng, Z., Zhou, X., Zhang, T.: 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)
31. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**, 91–110 (2004)
32. Lu, F., Chen, G., Liu, Y., Zhan, Y., Li, Z., Tao, D., Jiang, C.: Sparse-to-dense matching network for large-scale lidar point cloud registration. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 11270–11282 (2023). https://doi.org/10.1109/TPAMI.2023.3265531
33. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=a68SUt6zFt, featured Certification
34. Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Ilic, S., Hu, D., Xu, K.: Geotransformer: Fast and robust point cloud registration with geometric transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(8), 9806–9821 (2023)
35. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12179–12188 (October 2021)
36. Ren, S., Zeng, Y., Hou, J., Chen, X.: Corri2p: Deep image-to-point cloud registration via dense correspondence. IEEE Transactions on Circuits and Systems for Video Technology **33**(3), 1198–1208 (2022)
37. Revaud, J., De Souza, C., Humenberger, M., Weinzaepfel, P.: R2d2: Reliable and repeatable detector and descriptor. Advances in Neural Information Processing Systems **32** (2019)
38. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: International Conference on Computer Vision. pp. 2564–2571. Ieee (2011)
39. Rusu, R.B., Blodow, N., Beetz, M.: Fast point feature histograms (fpfh) for 3d registration. In: IEEE International Conference on Robotics and Automation. pp. 3212–3217. IEEE (2009)
40. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: 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)
41. Sarlin, P.E., DeTone, D., Yang, T.Y., Avetisyan, A., Straub, J., Malisiewicz, T., Bulo, S.R., Newcombe, R., Kontschieder, P., Balntas, V.: 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)
42. Shen, Z., Sun, J., Wang, Y., He, X., Bao, H., Zhou, X.: Semi-dense feature matching with transformers and its applications in multiple-view geometry. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 7726–7738 (2023)
43. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: 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)
44. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings

of the IEEE/CVF International Conference on Computer Vision. pp. 6411–6420 (2019)
45. Wang, H., Liu, Y., WANG, B., SUN, Y., Dong, Z., Wang, W., Yang, B.: Freereg: Image-to-point cloud registration leveraging pretrained diffusion models and monocular depth estimators. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=BPb5AhT2Vf
46. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5294–5306 (June 2025)
47. Wang, Y., He, X., Peng, S., Tan, D., Zhou, X.: Efficient loftr: Semi-dense local feature matching with sparse-like speed. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21666–21675 (2024)
48. Yao, G., Xuan, Y., Chen, Y., Pan, Y.: Quantity-aware coarse-to-fine correspondence for image-to-point cloud registration. IEEE Sensors Journal **24**(20), 33826–33837 (2024)
49. Yu, H., Li, F., Saleh, M., Busam, B., Ilic, S.: Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. Advances in Neural Information Processing Systems **34**, 23872–23884 (2021)
50. Zhou, J., Ma, B., Zhang, W., Fang, Y., Liu, Y.S., Han, Z.: Differentiable registration of images and lidar point clouds with voxelpoint-to-pixel matching. In: Advances in Neural Information Processing Systems. vol. 36, pp. 51166–51177 (2023)
51. Zhou, Q., Sattler, T., Leal-Taixe, L.: Patch2pix: Epipolar-guided pixel-level correspondences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4669–4678 (2021)