2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Angle-I2P: Angle-Consistent-Aware Hierarchical Attention for Cross-Modality Outlier Rejection

Muyao Peng*, Shun Zou*, Pei An, You Yang and Qiong Liu

***Abstract*—Image-to-point-cloud registration (I2P) is a fundamental task in robotic applications such as manipulation, grasping, and localization. Existing deep learning-based I2P methods seek to align image and point cloud features in a learned representation space to establish correspondences, and have achieved promising results. However, when the inlier ratio of the initial matching pairs is low, conventional Perspective-n-Points (PnP) methods may struggle to achieve accurate results. To address this limitation, we propose Angle-I2P, an outlier rejection network that leverages angle-consistent geometric constraints and hierarchical attention. First, we design a scale-invariant, cross-modality geometric constraint based on angular consistency. This explicit geometric constraint guides the model in distinguishing inliers from outliers. Furthermore, we propose a global-to-local hierarchical attention mechanism that effectively filters out geometrically inconsistent matches under rigid transformation, thereby improving the *Inlier Ratio* (IR) and *Registration Recall* (RR). Experimental results demonstrate that our method achieves state-of-the-art performance on the 7Scenes, RGBD Scenes V2, and a self-collected dataset, with consistent improvements across all benchmarks.**

***Index Terms*—Image-to-Point Cloud Registration, outliers rejection, spatial consistency**

I. Introduction

Image-to-Point Cloud Registration is a basic task in numerous downstream tasks like robot manipulation, simultaneous localization and mapping (SLAM) and autonomous driving [1]–[4]. It aims to calculate the alignment transformation $\mathbf{T} \in SE(3)$, which contains a rotation matrix $\mathbf{R} \in SO(3)$ and a three-dimension translation vector $\mathbf{t} \in \mathbb{R}^3$. Existing methods have achieved promising results in image-to-point cloud registration tasks [3], [5]–[7]. However, when training with limited data or in the presence of significant point cloud noise, numerous incorrect pixel-to-point cloud correspondences may arise. It will significantly affect the performance of Perspective-n-Points (PnP) [8].

In order to reject outliers from initial correspondences, Gomatch [9] proposes a linear-layer-based classifier module. It takes concatenated geometric features from putative 2D-3D keypoint pairs to filter out unreliable matches (via a configurable threshold parameter). GraphI2P [7] and its preceding work [10], [11] employ monocular depth estimation to predict depth from an image and subsequently back-project the image into 3D space. This converts the image-to-point cloud registration problem into a point cloud registration (PCR) problem. Then they construct an adjacent matrix and apply a graph neural network (GNN) to enhance pattern consistency among correct correspondences. A key drawback is that the inherent scale ambiguity in depth estimation compromises performance. Consequently, the key challenges in eliminating incorrect image-to-point cloud matches are: 1) how to minimize the modality gap between images and point clouds as much as possible, and 2) how to design effective geometric constraints to filter out non-conforming outliers and enhance registration performance.

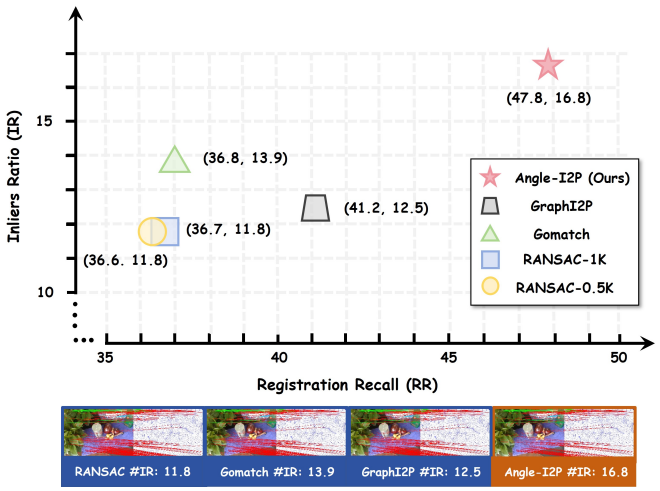

Fig. 1. Registration performance of existing image-to-point cloud outliers rejection methods on *Self-collected Datasets*. Existing works [7], [9], [12] have made efforts to reject outliers in I2P tasks. In this paper, we propose Angle-I2P. Through back-project, scale alignment and an angle-consistent-aware hierarchical attention, our method (★) demonstrates enhanced potential for practical deployment in real-world scenarios.

To solve the above problems, we propose Angle-I2P. It has better potential for practical deployment, shown as Fig. 1. To address the problem of mismatches in cross-modality registration, we propose a robust inlier criterion by transforming the image-point cloud matching problem into 3D space via monocular depth estimation and aligning scales globally. This method effectively leverages noise-robust 3D geometric properties to achieve efficient outlier rejection. To mitigate the influence of scale noise on inlier discrimination, we propose

Muyao Peng, Shun Zou, Pei An, You Yang, and Qiong Liu are with School of Electronic Information and Communications, Huazhong University of Science and Technology. (Email: {muyao99, zoushun, anpei96, yangyou, q.liu}@hust.edu.cn)
* The first two authors contributed equally.
Corresponding to yangyou@hust.edu.cn

a novel module named *Angle-Consistent-Aware Hierarchical Attention*. It effectively integrates angle-consistent features from global to local contexts. This integration enhances the capability to discriminate outliers, which in turn improves the overall registration performance. Extensive experiments on three challenging benchmarks demonstrate clear superiority of our method. Our main contributions are summarized as follows:

II. Related Works

A. Image-to-Point Cloud Registration

Existing works can be divided into two categories: detect-then-match based methods and no-detector based methods. The first methods mostly detect the keypoints in 2D and 3D space seperately and match them [5], [6], [9], [13]. MinCD-PnP [14] simplifies blind PnP into a Chamfer distance minimization task between 2D and 3D keypoints. It proposes MinCD-Net, a lightweight multi-task module that improves robustness to noise and outliers in image-to-point-cloud registration. This kind of methods lack interaction between two modality, result in low registration performance. No-detector based methods aim to establish cross-modality correspondences to support pose estimation by rationally fusing modalities, thereby enabling the expression and alignment of cross-modality features within a unified latent space. 2D3D-MATR [15] proposed a attention based methods to fuse the feature of different modality. FreeReg [16] and DiffReg [17] leverage diffusion models by formulating correspondence estimation as a denoising diffusion process within the doubly stochastic matrix space, where cross-modality features (image and point cloud) are iteratively refined through a lightweight transformer during reverse sampling to robustly establish geometrically consistent matches. GraphI2P [7] and its predecessor works [10], [11] transfer the modality gap to the distribution gap by introducing a monocular depth estimator to convert the images to point cloud.

B. Graph-Based Outliers Rejection Methods

Outliers Rejection has been widely studied in point cloud registration (PCR) [12], [18]–[23]. RANSAC [12] is the most popular method in this field, subsequent work achieves more robust results by incorporating graph structures [19]–[21], [24]. However, traditional methods are often constrained by limitations such as computational efficiency. In contrast, learning-based approaches can effectively approximate the optimal solution [22], [25], [26]. As our baseline, GraphSCNet [27] proposes the first learning-based outlier rejection method for non-rigid point cloud registration. It presents a graph-based local spatial consistency measure that leverages the local rigidity of deformations, combined with an attention-based correspondence embedding module to enhance feature representation and accurately filter outliers. HyperGCT [23] proposes a dynamic hypergraph neural network-based method that learns high-order geometric constraints for robust 3D point cloud registration, achieving state-of-the-art performance and improved generalization across diverse datasets. To the best of our knowledge, existing works don't focus on cross-modality outliers rejection.

III. Method

A. Overview

Given an image $\mathcal{I} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times 3}$ and a point cloud $\mathcal{P} \in \mathbb{R}^{\mathrm{N}\times 3}$, existing image-to-point cloud registration networks [5], [15], [16] aim to get the correspondences $\mathcal{C} = \{\mathbf{c}_i = (\mathbf{p}_i, \mathbf{q}_i) \in \mathbb{R}^5 |\ \mathbf{p}_i \in \mathcal{I}, \mathbf{q}_i \in \mathcal{P}\}_{i=1}^{N}$ first. Then $\mathcal{C}$ is used to calculate the alignment transformation $\mathbf{T} \in SE(3)$, which contains a rotation matrix $\mathbf{R} \in SO(3)$ and a three-dimension translation vector $\mathbf{t} \in \mathbb{R}^3$.

Despite achieving satisfactory results, existing image-point cloud registration methods [15], [28] still underperform when trained with limited data and validated across different scenes. Therefore, it is necessary to filter outliers from the initial set of correspondences $\mathcal{C}$ to obtain a more robust correspondence set $\mathcal{C}'$.

In this paper, an outliers rejection network is present for I2P tasks. First, we convert images into point clouds using a monocular depth estimator, and then enforce global-to-local spatial consistency to constrain the cross-modality data. Subsequently, to mitigate the impact of scale errors on spatial consistency computation, we designed an angular metric-based approach for calculating spatial consistency (Sec. III-B). Finally, we estimated and aligned the scale of the estimated point cloud with the ground truth. After rescale the estimated points cloud, we tailor a global-to-local cross-attention to constrained the geometric structure of correspondences from the global to local level. Thereby we believe it can effectively filter out image-point cloud correspondences under rigid transformations (Section III-C). The pipeline of our method is shown as Fig. 2

B. Angle-Based Spatial Consistency

Estimated Point Cloud Generation. In order to solve the modality gap between images and point clouds, we convert the images to point clouds using depth estimation follow GraphI2P [7]. First, we use Depth Anything V2 [29] to estimate the depth $\mathcal{D}$ of the image $\mathcal{I}$, expressed as (1). Then we use the intrinsics of the camera to convert image to point cloud, expressed as (2).

$$\mathcal{D} = \mathrm{DepthAnythingV2}(\mathcal{I}) \tag{1}$$

$$\mathcal{O}^{\mathrm{gt}} = \pi^{-1}(\mathcal{D}^{\mathrm{gt}}, \mathcal{K}) = \pi^{-1}(s\mathcal{D} + t, \mathcal{K}) = s\mathcal{O} + bias \tag{2}$$

in which $\mathcal{O} = \{\mathbf{o}_i\}_{i=1}^{N} \in \mathbb{R}^{\mathrm{N}\times 3}$ is the estimated point cloud and $\mathcal{O}^{\mathrm{gt}} = \{\mathbf{o}_i^{\mathrm{gt}}\}_{i=1}^{N} \in \mathbb{R}^{\mathrm{N}\times 3}$ is the point cloud computed using the ground-truth depth. $\pi^{-1}(\cdot)$ is the back projection method and $\mathcal{K}$ is the intrinsics of the camera. s, t are the scale and bias between the ground-truth depth and the estimated depth. $bias$ is the bias between the estimated point cloud and the ground-truth point cloud. Then we could get the estimated correspondences $\mathcal{C}^{\mathrm{est}} = \{(\mathbf{o}_i, \mathbf{q}_i) \in \mathbb{R}^6 |\ \mathbf{o}_i \in \mathcal{O}, \mathbf{q}_i \in \mathcal{P}\}_{i=1}^{N}$.

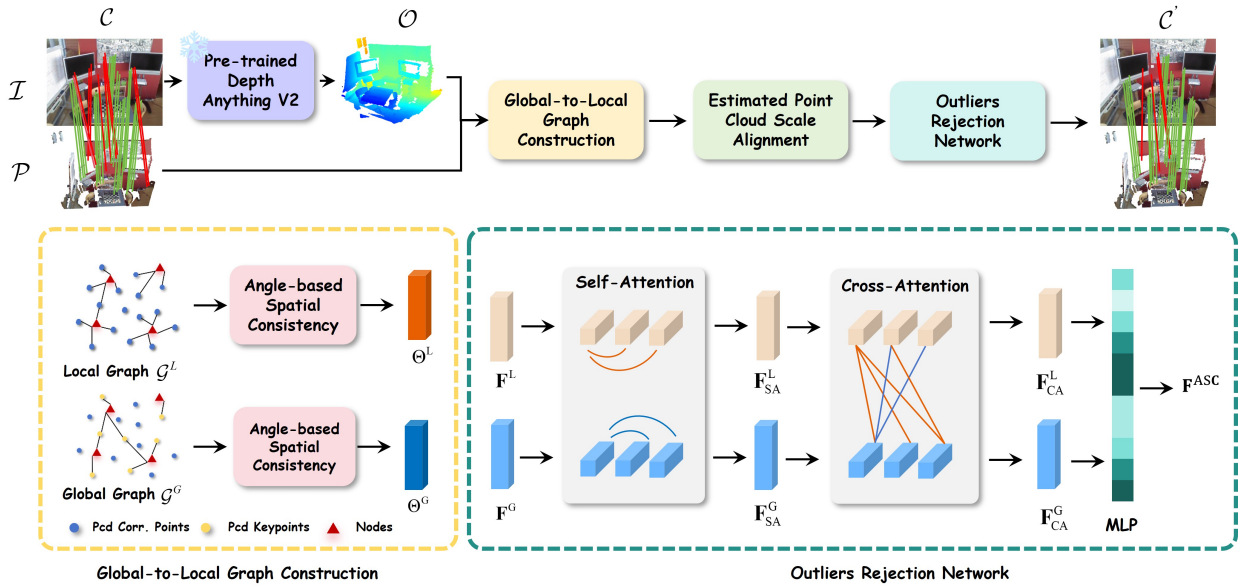

Fig. 2. Pipeline of the Angle-I2P. First, we back-project the image $\mathcal{I}$ to point cloud $\mathcal{O}$ using the depth estimated from pre-trained monocular depth estimator. Then we construct global-wise and local-wise graph $\mathcal{G}^{\mathrm{G}}$ and $\mathcal{G}^{\mathrm{L}}$. We use them to compute the global-wise and local-wise angle-based spatial consistency. Also, we use them to get the feature $\mathbf{F}^{\mathrm{G}}$ and $\mathbf{F}^{\mathrm{L}}$. Then $\mathbf{F}^{\mathrm{G}}$ and $\mathbf{F}^{\mathrm{L}}$ are put into the outliers rejection network to obtain a more robust correspondences $\mathcal{C}'$

Angle-Based Spatial Consistency. Any two correspondences in $\mathcal{C}$ should be satisfied with (3).

$$\|\mathbf{q}_i - \mathbf{q}_j\| \approx \|\pi^{-1}(\mathbf{p}_i, d_i^{\mathrm{gt}}, \mathcal{K}) - \pi^{-1}(\mathbf{p}_j, d_j^{\mathrm{gt}}, \mathcal{K})\| \tag{3}$$

in which d_i^{gt} is the ground-truth depth value of pixel $\mathbf{p}_i$. But due to the scale ambigious cased by the monocular depth estimate function, (3) is converted to (4):

$$\begin{aligned} &\|\mathbf{q}_i - \mathbf{q}_j\| \\ &\approx \|s * \pi^{-1}(\mathbf{p}_i, d_i, \mathcal{K}) - s * \pi^{-1}(\mathbf{p}_j, d_j, \mathcal{K})\| \\ &\approx s * \|\pi^{-1}(\mathbf{p}_i, d_i, \mathcal{K}) - \pi^{-1}(\mathbf{p}_j, d_j, \mathcal{K})\| \end{aligned} \tag{4}$$

in which d_i is the estimated depth value of pixel $\mathbf{p}_i$. Due to the presence of scale errors, the commonly used spatial consistency constraint in point cloud outliers rejection networks [20], [21], [27] is relaxed, leading to mismatches. Therefore, we propose a scale-invariant spatial consistency calculation method to constrain the geometric relationships between matching pairs. First, the centroid of each point cloud, $\mathbf{o}_c$ and $\mathbf{p}_c$, was computed, expressed as (5).

$$\mathbf{o}_c = \frac{1}{N}\sum_{i=1}^{N}\mathbf{o}_i, \quad \mathbf{p}_c = \frac{1}{N}\sum_{i=1}^{N}\mathbf{p}_i \tag{5}$$

where $N = |\mathcal{C}|$ is the number of the correspondences.

Next, all points were translated to a coordinate system centered at their respective centroid to eliminate translational degrees of freedom, denoted as $\hat{\mathbf{o}}_i$ and $\hat{\mathbf{p}}_i$. It can also be interpreted as a vector pointing from the origin toward the point cloud. For each pair of matching points, we compute the cosine of the angle between their corresponding vectors and subtract these values, as detailed in (6).

$$\cos(\alpha_{ij}^{\mathcal{I}}) = \frac{\hat{\mathbf{o}}_i \cdot \hat{\mathbf{o}}_j}{\|\hat{\mathbf{o}}_i\|\|\hat{\mathbf{o}}_j\|} = \frac{s^2\ (\hat{\mathbf{o}}_i^{\mathrm{gt}} \cdot \hat{\mathbf{o}}_j^{\mathrm{gt}})}{s^2\ \|\hat{\mathbf{o}}_i^{\mathrm{gt}}\|\|\hat{\mathbf{o}}_j^{\mathrm{gt}}\|} = \frac{\hat{\mathbf{o}}_i^{\mathrm{gt}} \cdot \hat{\mathbf{o}}_j^{\mathrm{gt}}}{\|\hat{\mathbf{o}}_i^{\mathrm{gt}}\|\|\hat{\mathbf{o}}_j^{\mathrm{gt}}\|} \tag{6}$$

in which $\alpha_{ij}^{\mathcal{I}}$ is the angle between the two vectors. Through derivation, we have proven that (6) is independent of scale s. Therefore, we can compute spatial consistency using the estimated point cloud, expressed as (7):

$$\theta_{ij} = [1 - \delta_{ij}^2/\sigma_d^2]_+ \tag{7}$$

Here, $[\cdot]_+ = \max(0, \cdot)$, $\delta_{ij} = \big|\,|\cos(\alpha_{ij}^{\mathcal{I}})| - |\cos(\alpha_{ij}^{\mathcal{P}})|\,\big|$ represents the difference in the cosine values of the vector angles between the two point clouds, and σ_d is a hyper-parameter controlling sensitivity to distance variation. If both points are inliers, δ_{ij} should be small, causing θ_{ij} to approach 1. Conversely, if at least one point is an outlier, δ_{ij} tends to be large and therefore θ_{ij} should be 0. This provides strong geometric justification for rejecting outliers in rigid scenarios, and we argue that it imposes equivalent constraints to traditional spatial consistency.

C. Outliers Rejection Network

We follow GraphSCNet [27] to use coordinates of points as initial features. For scale discrepancies exist between the estimated and actual point clouds, it will introduces biases into the extracted features, which ultimately degrades the model's performance. So we roughly estimated a global average scale factor between the reconstructed point cloud $\mathcal{O}$ and the point cloud $\mathcal{P}$. Since rotation preserves the Euclidean distance between points, the distance from each point to its cloud's

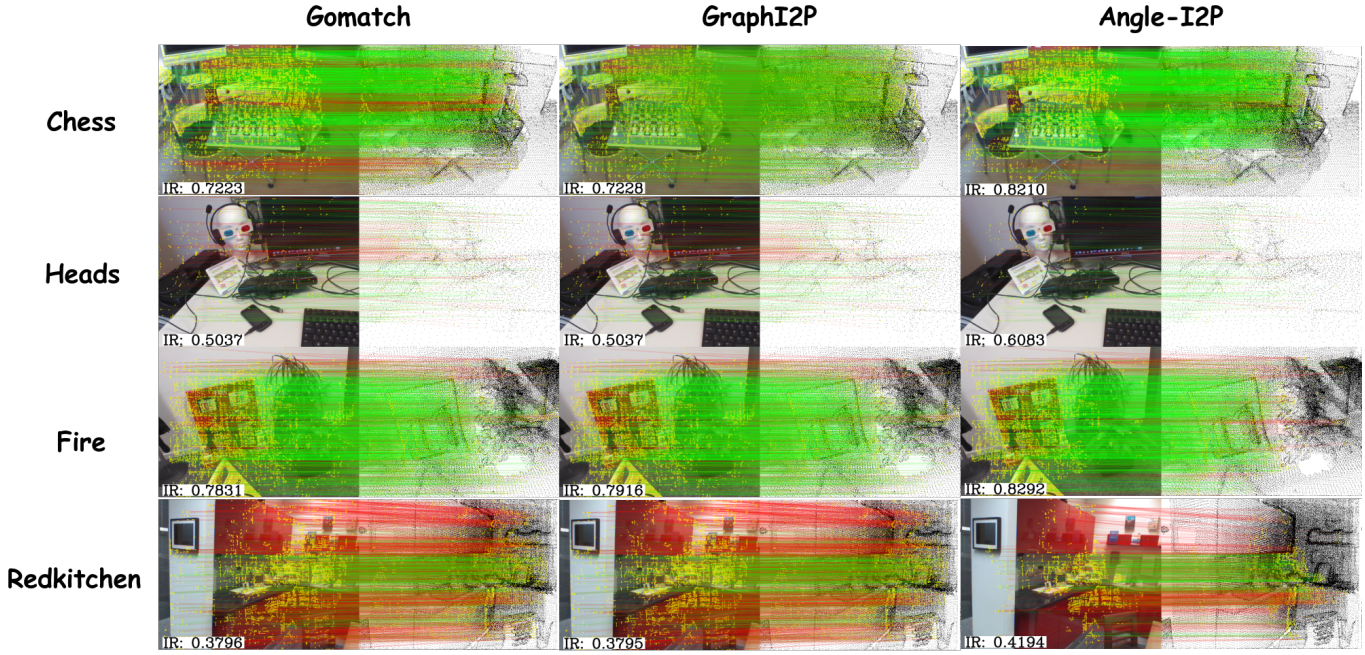

Fig. 3. Visualization of the outliers rejection results of each method. We show four selected scenes in 7Scenes datasets. Green lines represent inliers. Red lines represent outliers. Our method achieve the best performance of all methods (Compare the 3rd column to the 1st and 2nd column).

centroid is calculated, which are recorded as $\mathbf{l}_i^{\mathcal{I}}$ and $\mathbf{l}_i^{\mathcal{P}}$. Finally, the scale factor s between the two point clouds is estimated using the ratio of corresponding distances:

$$s^{\text{est}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{l}_i^{\mathcal{I}}}{\mathbf{l}_i^{\mathcal{P}}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\|\mathbf{o}_i - \mathbf{o}_c\|}{\|\mathbf{p}_i - \mathbf{p}_c\|} \tag{8}$$

We rescale the estimated point cloud $\hat{\mathbf{o}}_i$ to $\tilde{\mathbf{o}}_i = s^{\text{est}} \times \hat{\mathbf{o}}_i$ using the estimated scale s^{est}. Using the estimated point cloud coordinates as features may introduce initial errors. To address this, we further incorporate the point cloud normals as constraints. Then we could get the initial feature $\mathbf{f}_i$:

$$\mathbf{f}_i = [\tilde{\mathbf{c}}_i; \sin(2^{-1}\tilde{\mathbf{c}}_i); \cos(2^{-1}\tilde{\mathbf{c}}_i); \mathbf{n}_i^{\mathcal{I}}; \mathbf{n}_i^{\mathcal{P}}] \in \mathbb{R}^{24} \tag{9}$$

in which $\tilde{\mathbf{c}}_i$ is the conbination of $\tilde{\mathbf{o}}_i$ and $\hat{\mathbf{p}}_i$. $\mathbf{n}_i^{\mathcal{I}}$ and $\mathbf{n}_i^{\mathcal{P}}$ are the normals of the point cloud and estimated point cloud. The initial feature is then put into a linear layer block to get the $\mathbf{F}^{\text{init}}$.

Angle-Consistent-Aware Hierarchical Attention. For local spatial consistency, we sample a set of nodes $\mathcal{V} = \{\mathbf{v}_j \in \mathbb{R}^3 | \ j = 1, \cdots, V\}$ from the original point cloud $\mathcal{P}$. And then get a set of correspondences assigned to a node $\mathbf{v}_j$, denote as $\mathcal{C}_j = \{\mathbf{c}_i | \ \mathbf{c}_i \in \mathcal{N}_j\}_{i=1}^{K}$, in which $\mathcal{N}_j$ is the k-nearest neighbours of $\mathbf{v}_j$.

For global spatial consistency, we employ the method from P2-Net [5] to select the top M value-maximizing keypoints. These M keypoints effectively represent the global geometric information of the scene or object. Subsequently, we assign each of these M points to the nodes $\mathcal{V}$ using k-nearest neighbors approach, thereby constructing the global graph $\mathcal{G}^{\text{G}} \in \mathbb{R}^{V\times K\times 3}$.

Finally, global and local spatial consistency is computed by applying (7) to $\mathcal{G}^{\text{G}}$ and $\mathcal{G}^{\text{L}}$. We iteratively use self-attention and cross-attention to refine the features. In self-attention, $\mathbf{F}^{\text{Q}}$, $\mathbf{F}^{\text{K}}$ and $\mathbf{F}^{\text{V}}$ are local features or global features. In the cross-attention, we use either global or local features as the *Query*, and the other type of features as the *Key* and *Value*. This integrates both global and local geometric information.

$$\mathbf{Q} = \mathbf{W}^{\text{Q}}\mathbf{F}^{\text{Q}} \quad \mathbf{K} = \mathbf{W}^{\text{K}}\mathbf{F}^{\text{K}} \quad \mathbf{V} = \mathbf{W}^{\text{V}}\mathbf{F}^{\text{V}} \tag{10}$$

where $\mathbf{W}^{\text{Q}}$, $\mathbf{W}^{\text{K}}$ and $\mathbf{W}^{\text{V}}$ is the projection weight. Inspired by GraphSCNet [27], we reweight the attention score using the global angle-based spatial consistency and local angle-based spatial consistency.

$$\text{Attention} = \text{Softmax}(\Theta\frac{\mathbf{Q}\mathbf{K}^{\text{T}}}{\sqrt{d}})\mathbf{V} \tag{11}$$

We expect this strategy to effectively enable feature fusion leveraging both global and local geometric information, thereby filtering out geometrically inconsistent outliers.

Classification Head. We put the angle-based spatial consistency aware feature $\mathbf{F}^{\text{ASC}}$ into a MLP block to get the confidence score score of the inliers/outliers. At last, we could obtain a more robust correspondence set $\mathcal{C}' = \{\mathbf{c}_i \mid \text{score}_i \geq \tau\}_{i=1}^{L}$.

IV. Experiments

A. Experimental Setup

Implemental Details. We conduct our experiments on a single NVIDIA GeForce RTX 4090 GPU and remain most settings

TABLE I
EVALUATION RESULTS ON 7SCENES DATASETS. **BOLDFACED** NUMBERS HIGHLIGHT THE BEST AND THE SECOND BEST ARE UNDERLINED. ↑ MEANS HIGHER IS BETTER AND ↓ MEANS LOWER IS BETTER. † INDICATES THAT WE INDEPENDENTLY REVIEWED THE RELEVANT CODE USING THE METHODOLOGY DESCRIBED IN THE PAPER.

Model	IR↑	MRE↓	MTE↓	RR (@0.1m)↑
RANSAC-0.5k [12]	44.6	3.410	0.081	75.2
RANSAC-1k [12]	44.6	3.093	**0.075**	75.5
GoMatch [9]	45.0	3.119	0.076	76.2
GraphI2P† [7]	45.3	**3.090**	**0.075**	76.0
Angle-I2P (Ours)	**49.5**	3.237	**0.075**	**78.5**

TABLE II
EVALUATION RESULTS ON RGBD SCENES V2 DATASETS. **BOLDFACED** NUMBERS HIGHLIGHT THE BEST AND THE SECOND BEST ARE UNDERLINED. ↑ MEANS HIGHER IS BETTER AND ↓ MEANS LOWER IS BETTER. † INDICATES THAT WE INDEPENDENTLY REVIEWED THE RELEVANT CODE USING THE METHODOLOGY DESCRIBED IN THE PAPER.

Model	IR↑	MRE↓	MTE↓	RR (@0.1m)↑
RANSAC-0.5k [12]	33.0	2.052	0.046	43.1
RANSAC-1k [12]	33.0	1.967	0.045	43.7
GoMatch [9]	33.0	1.967	0.045	43.7
GraphI2P† [7]	33.2	1.967	0.045	43.8
Angle-I2P (Ours)	**39.1**	**1.887**	**0.044**	**44.7**

in GraphSCNet [27]. Learning rate is set to 1×10^{-4} and the weight decay is set to 1×10^{-6}. The M and the number of neighbors K in Section III-C is set to 100 and 32.

Baselines. As there are no existing end-to-end image-to-point cloud outliers rejection network, we have selected and reproduced networks with outlier removal modules. We mainly compare with three methods: (1) RANSAC: a method which commonly used to reject outliers in PCR and I2P tasks. (2) GoMatch [9], a method for inlier-outlier classification using a linear classification network. (3) GraphI2P [7], a method for inlier-outlier classification using a graph convolution network. Unfortunately, authors of GraphI2P don't public their code, we implement it based on the methodology described in their paper. It should be noted that we use Depth Anything V2 to get the virtual points. All initial correspondences $\mathcal{C}$ are obtained based on 2D3D-MATR [15].

Metrics. We evaluate our method with four metrics: (1) *Inlier Ratio* (IR), the ratio of pixel-point correspondences whose 3D distance is below a certain threshold (*i.e.*, 5cm). (2) *Mean Rotation Error* (MRE): The mean value of the rotation error. (3) *Mean Translation Error* (MTE): The mean value of the translation error. (4) *Registration Recall* (RR): the ratio of corresponding 3D points whose reprojection distances (using the ground-truth pose and the estimated pose) exceed a given threshold (*i.e.*, 0.1m).

B. Evaluation Results on 7Scenes Datasets

Datasets. 7Scenes [30] is a datasets which contains seven different indoor scenes. We trained 2D3D-MATR [15] using full training set and then evaluated on both the full training set and the test set to generate initial correspondences. Consequently, we obtain matching results for 4,048 training pairs and 2,034 test pairs. For our experiments, the test set remained consistent with the validation set.

Evaluation Results. Since the test set shares the same scenes as the training set, we obtain more image-point cloud correspondences. The results are shown in Tab. I. Our method outperforms the second best (GraphI2P [7]) by 4.2 percent on *Inlier Ratio* and outperforms the second best Gomatch (Gomatch [9]) 2.3 percent on *Registration Recall*. Furthermore, our model demonstrates significantly improved localization performance. When utilizing filtered matches for pose estimation, it achieves higher localization accuracy.

In order to better show our results, we visualize our results in Fig. 3 (The 3rd column is our method). It can be seen that our method can filter out more outliers than other methods.

C. Evaluation Results on RGBDScenesV2 Datasets

Datasets. RGBD Scenes V2 [31] contains 14 different indoor scenes. In this set of experiments, we evaluate models trained on Scenes 1-9. The trained model is then tested on Scenes 11-14 to get the initial correspondences, then we get 1748 training pairs, 497 testing pairs and 236 valid pairs. Note that the test scenes is unseen, we leverage this dataset to validate the performance of our outlier removal method in eliminating mismatches for cross-scene registration.

Evaluation Results. The results are shown in Table II. Due to the low number of initial correspondences $\mathcal{C}$ and low inlier ratio, methods like GoMatch [9] (which do not incorporate geometric constraints) failed to effectively filter outliers. Similarly, the performance of GraphI2P [7] in outlier removal was suboptimal, as it lacks hierarchical processing integrating global-to-local information. Our method outperforms the second best GraphI2P by 5.9 percent on *Inlier Ratio* and 0.9 percent on *Registration Recall*. We also visualize the results in Fig. 4.

D. Evaluation Results on Self-Collected Datasets

Datasets. To effectively validate the performance of our model in real-world scenarios, we captured data from eight distinct indoor scenes in a laboratory setting using an Intel Real Sense RGB-D camera. We evaluated the pre-trained 2D3D-MATR [15] model on the 7Scenes dataset across all eight scenes, obtaining corresponding results for both the training and testing sets.

Evaluation Results. The results are shown in Table III. On *Inlier Ratio*. our method outperforms the second best GoMatch

TABLE III
EVALUATION RESULTS ON SELF-COLLECTED DATASETS. **BOLDFACED** NUMBERS HIGHLIGHT THE BEST AND THE SECOND BEST ARE UNDERLINED. ↑ MEANS HIGHER IS BETTER AND ↓ MEANS LOWER IS BETTER. † INDICATES THAT WE INDEPENDENTLY REVIEWED THE RELEVANT CODE USING THE METHODOLOGY DESCRIBED IN THE PAPER.

Model	IR↑	MRE↓	MTE↓	RR (@0.2m)↑
RANSAC-0.5k [12]	11.8	21.283	0.330	36.6
RANSAC-1k [12]	11.8	18.316	0.298	36.7
GoMatch [9]	13.9	19.390	**0.268**	36.8
GraphI2P† [7]	12.5	17.394	0.304	41.2
Angle-I2P (Ours)	**16.8**	**16.035**	0.282	**47.8**

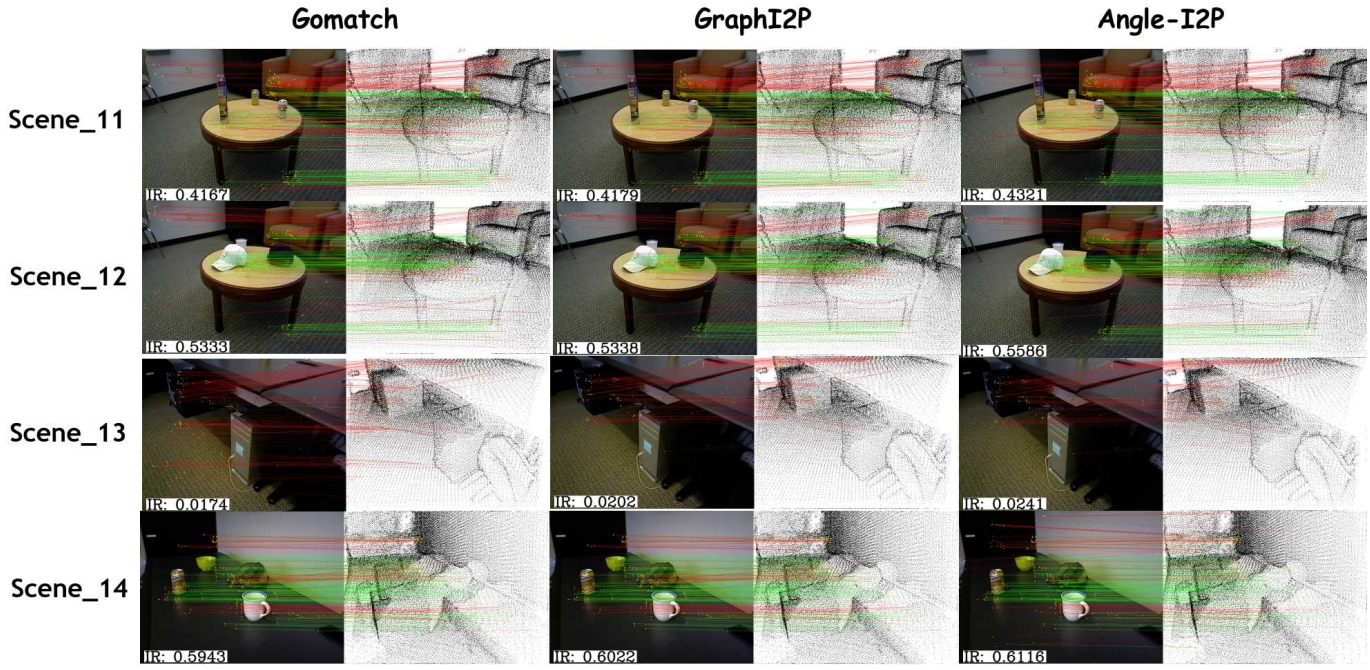

Fig. 4. Visualization of the outliers rejection results of each method. We show the results of RGBDScenesV2 datasets. Green lines represent inliers. Red lines represent outliers. Our method achieve the best performance of all methods (Compare the 3rd column to the 1st and 2nd column).

TABLE IV
ABLATION STUDIES ON 7SCENES DATASETS. **BOLDFACED** NUMBERS THE BEST. ADDITIONALLY, WE PRESENT QUANTITATIVE PERFORMANCE CHANGES RESULTING FROM THE REMOVAL OF RELEVANT MODULES. ↑ MEANS HIGHER IS BETTER AND ↓ MEANS LOWER IS BETTER.

Model	IR↑	RR (@0.1m)↑	MRE↓	MTE↓
(a.1) Angle-I2P (*full*)	**49.5**	**78.5**	3.237	**0.075**
(a.2) Angle-I2P w/o scale alignment	48.7	77.5	**3.153**	**0.075**
(b.1) Angle-I2P (*full*)	**49.5**	**78.5**	3.237	**0.075**
(b.2) Angle-I2P w/ distance-based spatial consistency	48.5	76.8	**3.216**	0.077
(c.1) Angle-I2P (*full*)	**49.5**	**78.5**	**3.237**	**0.075**
(c.2) Angle-I2P w/o global-to-local cross-attention	49.0	77.8	3.261	0.078
(c.3) Angle-I2P w/o reweight weights	45.0	76.9	3.516	0.083
(d.1) Angle-I2P w/ $\tau = 0.2$ (ours)	49.5	**78.5**	**3.237**	**0.075**
(d.2) Angle-I2P w/ $\tau = 0.4$	54.7	76.7	3.265	0.081
(d.3) Angle-I2P w/ $\tau = 0.5$	**58.0**	75.8	3.305	0.078

by 2.9 percent. And on the most important metric, *Registration Recall*, our method surpass GraphI2P by 6.6 percent. These results demonstrate the enhanced generalization capability and potential for practical applicability of our model.

E. Ablation Studies

Effectiveness of scale alignment. We remove the scale alignment module and denoted it as "*Angle-I2P w/o scale alignment*". The results are shown in Table IV(a). Without scale alignment, the estimated point cloud coordinates derived from the initial features contain errors. These errors propagate into the extracted features, consequently hindering the precise elimination of outliers within the image-point cloud correspondences. This ultimately leads to a decline in both the *Inlier Ratio* and the *Registration Recall*.

Effectiveness of angle-based spatial consistency. To effectively validate the role of angle-based spatial consistency, we use the original distance-based spatial consistency and denoted it as "*Angle-I2P w/ distance-based spatial consistency*". The results are shown in Table IV(b). Angle-based spatial consistency inherently eliminates the influence of scale errors, whereas distance-based spatial consistency imposes stricter requirements on coordinate accuracy. Removing the angle-based component allows scale errors to compromise the precision of this geometric constraint, consequently adversely affecting metrics such as *Inlier Ratio*, *Mean Rotation Error*, and *Mean Translation Error*.

Effectiveness of global-to-local cross-attention. To effectively validate the role of global-to-local cross-attention, we removed this mechanism from our model and denoted it as "*Angle-I2P w/o global-to-local cross-attention*". The results are shown in Table IV(c). In rigid point cloud registration, global information should also be considered. Neglecting global context would result in locally similar confidence scores for correspondences, causing filtered inliers to cluster excessively. This concentration ultimately leads to reduced

accuracy in the estimated pose.

To validate the role of spatial consistency weighting in the transformer architecture, we remove the consistency weights and denote this variant as *"Angle-I2P w/o reweight weights"*. The results indicate that without the guidance of consistency weighting, the model lacks a robust basis for distinguishing between inliers and outliers. The lackness of the weights leads to a 4.5 pp drop in the inlier ratio of the filtered matches compared to the full Angle-I2P model.

Effectiveness of inliers/outliers threshold. Finally, we study the effectiveness of threshold τ in Table IV(d). As the threshold increases, the inlier ratio rises while the registration accuracy decreases. This occurs because a higher threshold filters out more points. Additionally, neighboring points often exhibit similar spatial features, leading to comparable inlier scores. Consequently, the filtered points become more clustered, which can cause pose estimation algorithms to converge to incorrect solutions. Therefore, we adopt a balanced approach by setting the threshold to 0.2, which jointly improves both the *Inlier Ratio* and *Registration Recall*.

F. Limitation and Future Work

In this paper, the analysis of scale errors introduced by the monocular depth estimator remains relatively coarse, and the global scale as well as bias have not been adequately corrected to align the estimated point cloud with the actual one. Additionally, due to the coarse-to-fine matching strategy employed by 2D3D-MATR [15], the resulting matching pairs tend to form clusters. After filtering, these matches exhibit an uneven distribution, which leads to an improvement in IR (*Inlier Ratio*) but not a significant gain in RR (*Registration Recall*). These issues also require further investigation in future work.

V. Conclusion

In this paper, to effectively address the issue that existing image–point cloud registration methods still yield many outliers in the initial image–point cloud matches, we propose an end-to-end trainable framework for image–point cloud mismatch removal. This method can effectively eliminate mismatches in cross-modal matching and improve registration performance. We observe that in existing feature learning-based image-to-point cloud correspondences, many correspondences violate geometric consistency due to issues such as scale ambiguity. To address this, we design a scale-invariant cross-modal geometric constraint, combined with a global-to-local attention mechanism to effectively filter outliers. Experimental results across few-shot and cross-scenario settings demonstrate the superiority of our approach, showcasing enhanced outlier rejection performance and generalization capability. So we believe it has the potential to contribute to advancements in image–point cloud mismatch removal.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFC3015302, in part by the National Science Foundation of China under Grant 62502171, in part by the Key Research and Development Program of Hubei Province under Grant 2024BAB021 and Grant 2024BAB036.

References

[1] X.-M. Wu, J.-F. Cai, J.-J. Jiang, D. Zheng, Y.-L. Wei, and W.-S. Zheng, "An economic framework for 6-dof grasp detection," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2024, pp. 357–375.

[2] R. Murai, E. Dexheimer, and A. J. Davison, "Mast3r-slam: Real-time dense slam with 3d reconstruction priors," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025, pp. 16 695–16 705.

[3] P. An, X. Hu, J. Ding, J. Zhang, J. Ma, Y. Yang, and Q. Liu, "Olreg: Registration of image and sparse lidar point cloud with object-level dense correspondences," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 7523–7536, 2024.

[4] P. An, Y. Yang, J. Yang, M. Peng, Q. Liu, and L. Nan, "Enhance image-to-point-cloud registration with beltrami flow: P. an et al." *International Journal of Computer Vision*, vol. 133, no. 12, pp. 8589–8616, 2025.

[5] B. Wang, C. Chen, Z. Cui, J. Qin, C. X. Lu, Z. Yu, P. Zhao, Z. Dong, F. Zhu, N. Trigoni *et al.*, "P2-net: Joint description and detection of local features for pixel and point matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 16 004–16 013.

[6] M. Feng, S. Hu, M. H. Ang, and G. H. Lee, "2d3d-matchnet: Learning to match keypoints across 2d image and 3d point cloud," in *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2019, pp. 4790–4796.

[7] L. Bie, S. Pan, S. Li, Y. Zhao, and Y. 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 (CVPR)*, 2025, pp. 22 161–22 171.

[8] M. Peng, P. An, Y. Yang, and Q. Liu, "Ldf-i2p: Learning discriminative cross-modality features for image-to-point cloud registration," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–12, 2025.

[9] Q. Zhou, S. Agostinho, A. Ošep, and L. Leal-Taixé, "Is geometry enough for matching in visual localization?" in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 407–425.

[10] L. Bie, S. Pan, K. Cheng, and L. Han, "Build a cross-modality bridge for image-to-point cloud registration," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2024, pp. 1–6.

[11] L. Bie, S. Li, and K. Cheng, "Image-to-point registration via cross-modality correspondence retrieval," in *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR)*, 2024, pp. 266–274.

[12] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[13] Q.-H. Pham, M. A. Uy, B.-S. Hua, D. T. Nguyen, G. Roig, and S.-K. Yeung, "Lcd: Learned cross-domain descriptors for 2d-3d matching," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020, pp. 11 856–11 864.

[14] P. An, J. Yang, M. Peng, Y. Yang, Q. Liu, X. Wu, and L. Nan, "Mincd-pnp: Learning 2d-3d correspondences with approximate blind pnp," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025, pp. 26 519–26 528.

[15] M. Li, Z. Qin, Z. Gao, R. Yi, C. Zhu, Y. Guo, and K. 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 (ICCV)*, 2023, pp. 14 128–14 138.

[16] H. Wang, Y. Liu, B. Wang, Y. Sun, Z. Dong, W. Wang, and B. Yang, "Freereg: Image-to-point cloud registration leveraging pretrained diffusion models and monocular depth estimators," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[17] Q. Wu, H. Jiang, L. Luo, J. Li, Y. Ding, J. Xie, and J. Yang, "Diff-reg: Diffusion model in doubly stochastic matrix space for registration problem," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024, pp. 160–178.

[18] H. Yang, J. Shi, and L. Carlone, "Teaser: Fast and certifiable point cloud registration," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 314–333, 2020.

[19] D. Barath and J. Matas, "Graph-cut ransac," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6733–6741.
[20] Y. Zhang, H. Zhao, H. Li, and S. Chen, "Fastmac: Stochastic spectral sampling of correspondence graph," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 17 857–17 867.
[21] X. Zhang, Y. Zhang, and J. Yang, "Mac++: Going further with maximal cliques for 3d registration," in *Proceedings of International Conference on 3D Vision (3DV)*, 2025, pp. 261–275.
[22] X. Zhang, J. Yang, S. Zhang, and Y. Zhang, "3d registration with maximal cliques," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 17 745–17 754.
[23] X. Zhang, J. Ma, J. Guo, W. Hu, Z. Qi, F. Hui, J. Yang, and Y. Zhang, "Hypergct: A dynamic hyper-gnn-learned geometric constraint for 3d registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025, pp. 24 750–24 759.
[24] D. Barath and J. Matas, "Graph-cut ransac: Local optimization on spatially coherent structures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4961–4974, 2021.
[25] Y. Zhang, J. Zhang, X. Qian, Y. Cen, B. Zhang, and J. Gong, "Muscle-reg: Multi-scale contextual embedding and local correspondence rectification for robust two-stage point cloud registration," *IEEE Robotics and Automation Letters*, 2025.
[26] J. Wang and Z. Li, "3dpcp-net: A lightweight progressive 3d correspondence pruning network for accurate and efficient point cloud registration," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1885–1894.
[27] Z. Qin, H. Yu, C. Wang, Y. Peng, and K. Xu, "Deep graph-based spatial consistency for robust non-rigid point cloud registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 5394–5403.
[28] J. Li and G. H. Lee, "Deepi2p: Image-to-point cloud registration via deep classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 960–15 969.
[29] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21 875–21 911, 2024.
[30] B. Glocker, S. Izadi, J. Shotton, and A. Criminisi, "Real-time rgb-d camera relocalization," in *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2013, pp. 173–179.
[31] K. Lai, L. Bo, and D. Fox, "Unsupervised feature learning for 3d scene labeling," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 3050–3057.