International Journal of Computer Vision
https://doi.org/10.1007/s11263-025-02575-4

Enhance Image-to-Point-Cloud Registration with Beltrami Flow

Pei An[1] · You Yang[1] · Jiaqi Yang[2] · Muyao Peng[1] · Qiong Liu[1] · Liangliang Nan[3]

Received: 18 October 2024 / Accepted: 23 August 2025

Abstract
Image-to-point-cloud (I2P) registration is a fundamental yet challenging problem in computer vision. Despite significant advances in deep learning, I2P registration struggles with correspondence accuracy when training samples are limited. To address this challenge, we propose a Beltrami flow based I2P registration method termed Flow-I2P. From the perspective of information geometry, I2P registration can be reframed as a manifold alignment problem. Our in-depth analysis shows that Beltrami flow enhances I2P registration by improving manifold alignment quality. Building on this analysis, we introduce a Beltrami flow based cross-modality feature interaction layer, B-flow, to progressively refine manifold alignment. To reduce memory and computation demands, B-flow is then optimized into C-flow through the incorporation of feature covariance-based attention. We further enhance I2P registration performance by developing Flow-I2P, which incorporates normal features, stacked C-flow layers, and a two-stage training strategy. To evaluate the registration performance of Flow-I2P, we conduct extensive experiments on five indoor and outdoor datasets, including RGB-D V2, 7-Scenes, ScanNet, KITTI, and a self-collected dataset. Our results indicate that Flow-I2P achieves higher inlier ratio (IR) and registration recall (RR) compared to state-of-the-art methods. We conclude that Flow-I2P significantly enhances I2P registration with superior capabilities. The source code of Flow-I2P is available

Keywords Image-to-point-cloud registration · Information geometry · Manifold alignment · Beltrami flow

1 Introduction

Image-to-point-cloud (I2P) registration is a fundamental task in computer vision. It aims to establish 2D-3D correspondences between images and point clouds (David et al., 2004). These correspondences are essential for estimating the 6 degree-of-freedom (DoF) camera pose using perspective-n-point (PnP) (Lepetit et al., 2009). I2P registration has broad applications, including multi-sensor calibration (An et al., 2024b), 6 DoF object pose estimation (Xu et al., 2024), visual localization (Kim et al., 2023; Miao et al., 2023), robot state estimation (Ye et al., 2020), point cloud colorization (Vechersky et al., 2018), and visual navigation (Liu et al., 2024). As a result, I2P registration has garnered significant attention in recent years.

Unlike related tasks such as image registration and point cloud registration, I2P registration remains under-explored, due to the inherent challenge of **cross-modality between images and point clouds**. These two data types differ significantly in structure, dimension, and features (An et al., 2022). This makes the design of handcrafted 2D-3D descriptors particularly difficult. As a result, progress in I2P registration has lagged behind that of its sister tasks (Yang et al., 2021). Early approaches to I2P registration framed it as an optimization problem, focusing on simultaneously estimating pose and correspondences (Moreno-Noguer et al., 2008) or on outlier-robust estimation (Yang & Carlone, 2023). However, these methods face significant computational burdens, making it difficult to achieve [1] real-time performance and robustness (An et al., 2024b). Recently, deep learning has emerged as a promising approach to I2P registration (Wang et al., 2021). With sufficient data, learning based approaches

Communicated by Gunhee Kim.

✉ You Yang
yangyou@hust.edu.cn

[1] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China

[2] School of Computer Science, Northwestern Polytechnical University, Xi'an, China

[3] Urban Data Science Section, Delft University of Technology, Amsterdam, Netherlands

[1] https://github.com/anpei96/Flow-I2P-demo

Published online: 19 September 2025

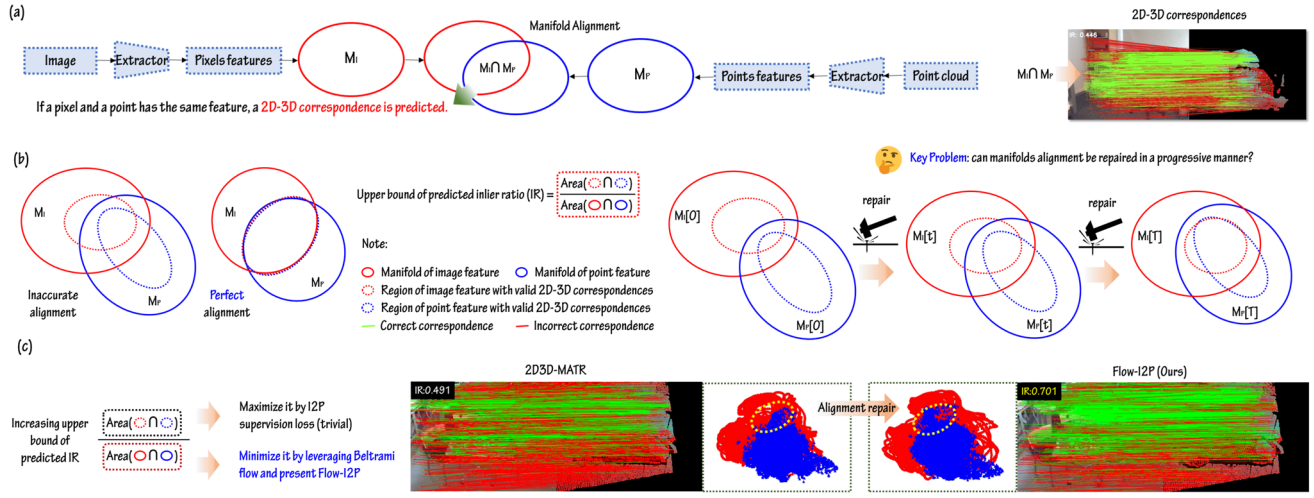

Fig. 1 Illustration of the core idea behind our approach in the ideal case of I2P registration. (a) I2P registration can be reframed as a manifold alignment problem in a latent feature space, where image features and point cloud features lie on manifolds $\mathbb{M}_I$ and $\mathbb{M}_P$, respectively. The intersection, $\mathbb{M}_I \cap \mathbb{M}_P$ **not only reflects the manifold alignment quality, but also determines the upper bound of the predicted inlier ratio**. More details are given in Appendix A. (b) This insight reveals that the key challenge in enhancing I2P registration lies in progressively improving manifold alignment. (c) Our proposed method, Flow-I2P, leverages Beltrami flow to enhance manifold alignment quality, thereby significantly increasing the inlier ratio. The manifolds generated from pixel and point features are visualized in 2D space using t-SNE

can overcome cross-modality differences and achieve strong performance on the public datasets (Wang et al., 2024). However, these methods often struggle with generalization when training data is limited, which is a common issue in applications such as mobile robot visual localization (Kim et al., 2023) . To improve generalization in I2P registration, it is crucial to **overcome the cross-modality challenge with limited training data**.

As deep learning based I2P registration is still in its early stages, few researchers emphasize enhancing generalization. For instance, Wang et al. designed a zero-shot I2P registration architecture, FreeReg (Wang et al., 2024), but it suffers from limited accuracy and relies on a large language model (LLM) based diffusion network, which is impractical for robots with limited computing resources. Therefore, improving generalization in I2P registration under limited data conditions remains an open challenge.

In this paper, we explore I2P registration generalization from the information geometry perspective. We conceptualize I2P registration as a manifold alignment problem, where learning-based methods aim to align the manifolds generated from image and point cloud features. As shown in Fig. 1, accurate manifold alignment directly correlates with accurate registration results. However, **existing approaches lack an effective mechanism to adaptively align cross-modal manifolds**. This explains the weak generalization performance when training samples are scarce. To address this, we present a Beltrami flow based I2P registration network.

Our in-depth analysis reveals that Beltrami flow (Chamberlain et al., 2021a) is well-suited for correcting manifold misalignment. By treating manifold alignment as a single manifold smoothing problem with the manifold alignment prior condition, Beltrami flow, a well-known technique for manifold smoothing, has the potential to significantly enhance I2P registration generalization.

Building on this analysis, we propose a Beltrami flow based feature interaction layer, B-flow, designed to progressively correct manifold alignment. In the Beltrami flow model, self- and cross-attention are used to represent the flow process. To reduce memory and computation demands of B-flow, we introduce C-flow, which leverages feature covariance-based attention. To further improve registration performance, we design an I2P registration network, Flow-I2P, with stacked C-flow layers, where its input features are enriched with surface normals and their gradients. A two-stage training scheme ensures the efficient supervision of Flow-I2P.

Finally, we evaluate the I2P registration performance of current approaches through extensive experiments on the RGB-D V2 (Lai et al., 2014), 7-Scenes (Glocker et al., 2013), ScanNet V2 (Dai et al., 2017), self-collected, and KITTI (Geiger et al., 2012) datasets. The results indicate that Flow-I2P achieves superior inlier ratio (IR) and registration recall (RR) compared to state-of-the-art methods. Thanks to the lightweight nature of C-flow, Flow-I2P achieves a favorable trade-off between I2P registration performance and runtime efficiency, as shown in Fig. 2. Thus, we believe that Flow-I2P significantly improves I2P registration accuracy. In summary, our main contributions are as follows:

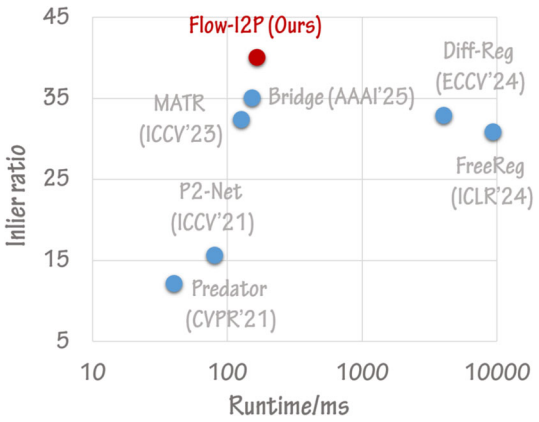

Fig. 2 Comparison of I2P registration performance among state-of-the-art methods. Results are evaluated on the RGBD-V2 dataset (Lai et al., 2014) in terms of inlier ratio and runtime. The proposed Flow-I2P achieves a superior trade-off between registration accuracy and efficiency, owing to its lightweight and effective Beltrami flow based feature interaction layers

- We recast I2P registration as a manifold alignment problem from the perspective of information geometry and establish the link between Beltrami flow and manifold alignment, demonstrating its utility in I2P registration.
- We propose Beltrami flow-based feature interaction layers, B-flow and C-flow, to progressively correct manifold misalignment. C-flow refines B-flow by incorporating feature covariance-based attention, reducing memory and computational requirements.
- We introduce Flow-I2P, an I2P registration network that leverages surface normals and their gradients, along with a two-stage training scheme. Flow-I2P outperforms state-of-the-art methods across five public indoor and outdoor datasets.

2 Related Work

As the proposed method focuses on I2P registration, we briefly review the background and representative approaches below.

2.1 Background of I2P Registration

We begin by explaining why I2P registration is important, as many existing works often fail to explore this issue in depth despite advancing the field.

In mainstream robotic applications, **I2P registration often occupies a marginal role**. The core application of I2P registration is localization. However, for general robot localization tasks, point cloud registration is typically preferred. This is because point clouds generated by LiDAR or depth camera share the same modality as the pre-built 3D map, thus enhancing the localization accuracy (Yin et al., 2024). Even for specialized localization tasks (visual localization), researchers tend to exploit image registration (i.e., matching image queries to an image database) instead of I2P registration, as single modality registration is more stable. Some researchers have used I2P registration in a simplified case (Yu et al., 2020) where each point in the 3D point cloud has a visual descriptor (i.e., SIFT (Lowe, 2004)). However, this does not represent the general I2P registration problem. Others designed end-to-end networks to predict poses directly (Chang et al., 2021), or even without 3D pre-built maps (Brachmann et al., 2023). Based on the above discussion, I2P registration appears to be at a disadvantage, with other methods potentially offering better performance.

Is I2P registration a redundant technique? The answer is no. We address this question from both application and theoretical perspectives. First, from the application viewpoint, the demand for visual localization in a given pre-built 3D map is rapidly growing (Kim et al., 2023), because of three reasons: (i) large-scale high-precision 3D maps are now easy to obtain, thanks to the development of multi-sensor based simultaneous localization and mapping (SLAM) (Lv et al., 2023); (ii) most robotic applications are conducted in a fixed scene (Thomas et al., 2023). In this context, pre-built 3D maps can be stored in the robot systems before usage; (iii) to reduce costs, monocular camera-based localization and navigation is a promising technique for robots (Matsumoto et al., 2024; Kim et al., 2023). In this context, I2P registration is crucial for robot perception because it can map images captured by a robot to a pre-built 3D map. Second, from the theoretical viewpoint, I2P registration holds academic value in two aspects: (i) unlike many end-to-end visual localization methods, I2P registration is not a black box. Instead, it is grounded in projective geometry, making it beneficial for researchers seeking to design safe, certifiable, and accurate localization algorithms; (ii) establishing accurate 2D-3D correspondences between images and point clouds remains a challenge in computer vision. Moreover, studying I2P registration helps uncover the mechanism behind general cross-modality feature learning. Thus, I2P registration is a crucial technique in robotics.

2.2 Representative Methods of I2P Registration

After analyzing the significance of I2P registration, we illustrate the mainstream approaches to this task. In general, feature interaction denotes a kind of operation to strengthen source and target data representation via feature fusion. Feature interaction is a key module in general registration tasks (Wu et al., 2021). Based on the usage of feature interaction, existing methods can be broadly categorized into non-interaction and interaction-based approaches.

Non-interaction-based approaches. This kind of approach estimates 2D-3D correspondences without interacting with cross-modality features. It typically employs two independent backbone networks to extract 2D pixel and 3D point features, respectively. After that, 2D-3D correspondences are determined by comparing feature distances. Feng et al. were the pioneers in this task (Feng et al., 2019). They designed two independent extractors to learn descriptors from image and point cloud patches. In the training stage, as false-positive correspondences are far greater than true-positives, the triplet loss (Schroff et al., 2015) is used for supervision. Following the thought of previous work (Feng et al., 2019), Wang et al. designed a registration network P2-Net (Wang et al., 2021). Instead of splitting images or point clouds into patches, they directly extracted pixel and point features using established 2D and 3D backbone networks, more convenient than previous work (Feng et al., 2019). Extended from image matching work D2-Net (Dusmanu et al., 2019), they extracted 2D and 3D keypoints from 2D, 3D feature maps via channel-wise and spatial-wise non-maximum suppression (NMS). To face I2P registration with a low inlier ratio, they leveraged circle loss (Sun et al., 2020) to enhance the supervision efficiency. Kim et al. designed a cross-modality large-scale visual localization method EP2P-Loc (Kim et al., 2023). They focused on the registration of images and a large-scale point cloud map. In this task, they split the point cloud map into cubic submaps and then learnt global cross-modality descriptors to align the image with the corresponding submaps. The task is then converted into a regular I2P registration between images and submaps. They designed a network similar to P2-Net (Wang et al., 2021). In the training stage, a differentiable PnP solver (Chen et al., 2022) is used to supervise the registration quality. Recently, Wang et al. leveraged two pretrained diffusion models to map cross-modality features into the same latent space (Wang et al., 2024). Although it achieves certain accuracy in unseen scenes, it requires a diffusion model based on a large language model (LLM) and cannot be used on memory-limited devices.

From 2019 to 2024, backbone networks in non-interaction-based I2P registration have involved a transition from classical networks designed for images and point clouds to diffusion models, while the whole registration architecture remains nearly the same. As backbone networks are independent of each other, it is tough to ensure that they can map cross-modality features into the same latent space. However, the architectures and loss functions used in these approaches have inspired the development of interaction-based I2P registration.

Interaction-based approaches. This kind of approach aims to learn 2D-3D correspondences with cross-modality feature interaction. In practice, cross-modality feature interaction schemes are highly flexible and varied.

In the early stage, Li and Lee were the first to consider cross-modality interaction in I2P registration (Li & Lee, 2021). Their interaction method is named **global feature interaction** (GFI). They first abstracted 1D global features from 2D images and point clouds. Then, cross-attention is utilized to interact with these global features. After interaction, the global features are replicated with the same shapes of the 2D image and the 3D point cloud. Finally, the replicated global features are concatenated with the original image and point cloud features, respectively. The advantage of GFI is easy to implement. However, the drawbacks are evident: (i) 1D global features cannot represent local structures, which are very crucial for 2D-3D correspondence learning; (ii) feature concatenation with replicated global features reduces the discriminative ability of cross-modality features.

Several researchers have attempted to refine this work (Li & Lee, 2021). Ren et al. designed a complex GFI that can not only predict 2D-3D overlap regions but also estimate 2D-3D correspondences (Ren et al., 2023). Zhou et al. leveraged GFI only in the 2D-3D overlap region prediction (Zhou et al., 2023). The core of their network is similar to P2-Net (Wang et al., 2021). Although the performance of these works (Ren et al., 2023; Zhou et al., 2023) is higher than previous work (Li & Lee, 2021), the network or loss revision cannot overcome the disadvantage of GFI.

In the meantime, other researchers attempted to explore cross-modality feature interaction in a different way. We roughly categorize these approaches under the term **local feature interaction** (LFI). A common geometrical feature is a cue of LFI. Some researchers modeled the projection of object edges from 3D space to 2D image by nearest searching (Yuan et al., 2021; An et al., 2020, 2024b). Although these approaches are not learning-based, they provide valuable insights for designing registration networks. Common semantic feature is also a critical cue of LFI. Liu et al. established 2D-3D correspondences via the nearest searching pixel and point with the same semantic label (Liu et al., 2021). However, these nearest searching based methods require a reliable initial pose. Recently, An et al. conducted a survey and summarized the geometrical and semantic LFI methods in the background of LiDAR-camera extrinsic calibration (a special case of I2P registration) (An et al., 2024a).

Recently, some researchers used cross-attention to construct LFI. Inspired by SuperGlue (Sarlin et al., 2020), Zhou et al. first leveraged graph neural network (GNN) based self- and cross-attention to interact 2D and 3D keypoints features (Zhou et al., 2022). However, strictly speaking, their method involves 2D-3D keypoints matching rather than I2P registration, as it assumes the 2D and 3D keypoints are provided beforehand. Building on previous works (Sarlin et al., 2020), they employed a differentiable Sinkhorn layer (Campbell et al., 2020) for correspondence learning. Inspired by the 3D point cloud registration work GeoTransformer (Qin et al.,

2022), Li et al. proposed an I2P registration network called 2D3D-MATR (Li et al., 2023). In their work, LFI is constructed by a series of transformer-based stacked attention layers. However, due to the large computational complexity of standard transformer layers, LFI in 2D3D-MATR is only used to extract the fine-grained 2D and 3D patch features. It increases the accuracy of 2D-3D patches matching, and indirectly improves the accuracy of 2D-3D correspondences. Recently, some researchers have improved the transformer-based LFI by leveraging the denoising diffusion probabilistic model (DDPM) (Wu et al., 2024) and uncertainty estimation module (Cheng et al., 2025).

From the above discussion, LFI-based methods have attracted more and more attention in I2P registration, because they improve the granularity of cross-modality feature interaction and are more effective than non-interaction-based or GFI-based methods.

2.3 Summary

Based on the preceding discussion, the key aspects of I2P registration can be summarized in three main points:

- **Significance**. With advancements in 3D scene reconstruction, effectively reusing pre-built point cloud maps is essential for enhancing visual localization. As a result, there has been increasing interest in learning based I2P registration.
- **Tendency**. In line with trends in learning-based image and point cloud registration, designing feature interaction modules has become a key focus in I2P registration. Compared to GFI, LFI offers greater advantages by enhancing the discriminative ability of cross-modality features.
- **Open problem**. Although current transformer-based LFI schemes (Li et al., 2023; Wu et al., 2024; Cheng et al., 2025) have made progress in I2P registration, they suffer from the large computation burden and fail to learn the fine-grained pixel-level features. This issue limits the I2P registration performance and generalization ability. Thus, the construction of a lightweight and effective LFI scheme to learn the representative pixel-level features is still an open problem.

3 Proposed Method

To address the open problem mentioned in Sec. 2.3, we aim to rethink transformer-based LFI and learn the fine-grained pixel-level features from the perspective of information geometry. To achieve this goal, we establish a connection between Beltrami flow and LFI, and propose I2P registration based on Beltrami flow.

3.1 Problem Statement

The goal of I2P registration is to identify 2D-3D correspondences between an image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ and a point cloud $\mathcal{P} \in \mathbb{R}^{N\times 3}$. Here, H and W represent the image height and width, respectively, while N is the number of points. The 2D-3D correspondence $\langle I_i, P_j\rangle$ must satisfy the pinhole camera projection constraint (Zhang, 2000):

$$d_i I_i = \mathbf{K}(\mathbf{R}P_j + \mathbf{t}) \tag{1}$$

where $I_i \in \mathbb{R}^3$, $P_j \in \mathbb{R}^3$ represent a pixel in $\mathcal{I}$ and a point in $\mathcal{P}$, respectively. $\mathbf{K} \in \mathbb{R}^{3\times 3}$ is the intrinsic matrix of the camera (Zhang, 2000). $\mathbf{R}$ and $\mathbf{t}$ are the rotation matrix and translation vector from the world coordinate system to the camera coordinate system. d_i is depth of I_i. Since $\mathbf{K}$, $\mathbf{R}$, $\mathbf{t}$, and d_i are unknown in I2P registration, predicting the correspondence $\langle I_i, P_j\rangle$ from $\mathcal{I}$ and $\mathcal{P}$ becomes a challenging task. Current approaches (Feng et al., 2019; Wang et al., 2021; Li et al., 2023) tend to determine $\langle I_i, P_j\rangle$ via the feature distance. In I2P registration, an ideal criterion of 2D-3D correspondence is:

Definition 1 ***(Ideal correspondence)***. 2D-3D correspondence $\langle I_i, P_j\rangle$ is established if and only if the features of I_i and P_j are equal. According to this definition, $\langle I_i, P_j\rangle$ forms a 2D-3D correspondence if and only if (see Fig. 1(a))

$$\mathbf{x}_i = \mathbf{y}_j \Leftrightarrow \langle I_i, P_j\rangle \textit{ is a 2D-3D correspondence.} \tag{2}$$

$$\mathbf{x}_i = \mathbf{F}(I_i|\mathcal{I}),\ \ \mathbf{y}_j = \mathbf{G}(P_j|\mathcal{P}) \tag{3}$$

where $\mathbf{x}_i \in \mathbb{R}^c$ and $\mathbf{y}_j \in \mathbb{R}^c$ are the feature vectors learned from I_i and P_j, respectively. $\mathbf{F}(\cdot)$ and $\mathbf{G}(\cdot)$ are the learnable feature extractors. In the practical I2P registration, $\langle I_i, P_j\rangle$ is a correspondence if $\|\mathbf{x}_i - \mathbf{y}_j\| \le \delta_f$ where δ_f is a feature distance threshold (Wang et al., 2021). In Sec. 3.2, we consider the ideal I2P registration mainly for simplicity of discussion.

3.2 Relation Between Beltrami Flow and I2P Registration

Let $\mathbb{M}_I$ and $\mathbb{M}_P$ represent a manifold in c-dimensional Euclidean space, which contains $\{\mathbf{x}_i\}_{i=1}^M$ and $\{\mathbf{y}_j\}_{j=1}^N$, where $M = HW$. As illustrated in Fig. 1, the key to improving I2P registration capacity is to **correct manifold alignment** on $\mathbb{M}_I$ and $\mathbb{M}_P$. We analyze how the Beltrami flow can help refine this alignment. The analysis is summarized in Fig. 3.

First, we examine the alignment of feature manifolds under ideal correspondence conditions. For a given correspondence $\langle I_i, P_j\rangle$, according to Eq. (2), we have $\mathbf{x}_i = \mathbf{y}_j \in \mathbb{M}_I \cap \mathbb{M}_P$. This condition is satisfied for every valid correspondence. This implies that $\mathbb{M}_I \cap \mathbb{M}_P$ contains all the

Fig. 3 Overview of the motivation. To improve the generalization ability of I2P registration, we establish a connection between I2P registration and Beltrami flow by reformulating the problem as a manifold smoothing task. Specifically, **we introduce a manifold alignment prior condition**, which allows the manifold alignment objective to be transformed into a manifold smoothing process. This condition is generally easy to satisfy in practice, making the reformulated approach broadly applicable

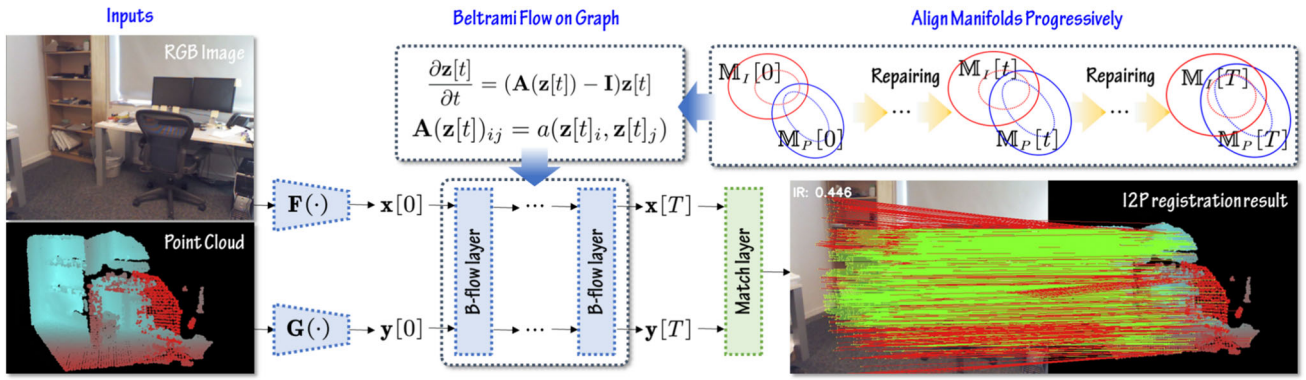

Fig. 4 Overview of our prototype registration architecture. To achieve I2P registration with high generalization ability, we develop a prototype registration architecture that integrates the proposed B-flow layers. These layers enhance feature interaction and alignment across modalities, contributing to improved registration performance

information about 2D-3D correspondences. An in-depth discussion of $\mathbb{M}_I \cap \mathbb{M}_P$ is provided in Appendix A. In the ideal case, $\mathbb{M}_I$ and $\mathbb{M}_P$ are perfectly aligned.

Next, we discuss manifold alignment in the actual I2P registration case. Actually, $\mathbf{x}_i = \mathbf{y}_j$ does not necessarily indicate that $\langle I_i, P_j \rangle$ is a correct 2D-3D correspondence. This implies that $\mathbb{M}_I \cap \mathbb{M}_P$ contains incorrect correspondences, indicating that $\mathbb{M}_I$ and $\mathbb{M}_P$ are misaligned, as shown in Fig. 1. Thus, it is essential to correct the structure of $\mathbb{M}_I \cap \mathbb{M}_P$. However, since $\mathbb{M}_I$ and $\mathbb{M}_P$ are complex surfaces in the latent space, it is difficult to directly correct the alignment of $\mathbb{M}_I$ and $\mathbb{M}_P$.

To address this problem, we propose a solution that aligns the manifolds progressively. We aim to construct a learnable mechanism $\mathcal{F}(\cdot)$ such that $(\mathbb{M}_I[t], \mathbb{M}_P[t]) = \mathcal{F}(\mathbb{M}_I[t-1], \mathbb{M}_P[t-1])$. As t increases, the alignment quality between $\mathbb{M}_I[t]$ and $\mathbb{M}_P[t]$ is progressively improved. To explore $\mathcal{F}(\cdot)$, we introduce the **manifold alignment prior condition**, that $\mathbb{M}_I[0]$ and $\mathbb{M}_P[0]$ should be coarsely aligned. This implies that the upper bound of feature distance (i.e., $\|\mathbf{x}_i - \mathbf{y}_j\|$ for the correspondence $\langle I_i, P_j \rangle$) of the whole correct 2D-3D correspondences in $\mathbb{M}_I[0] \cap \mathbb{M}_P[0]$ is $\delta[0]$. Under this condition, the function of $\mathcal{F}(\cdot)$ is to ensure that the distance upper bound is decreasing in the next iteration, i.e., $\delta[t] < \delta[t-1]$, meaning the stitched manifold $\mathbb{M}_H = \mathbb{M}_I \cup \mathbb{M}_P$ is smoothing at the regions with the correct correspondences. Based on this, the manifold alignment on $\mathbb{M}_I$ and $\mathbb{M}_P$ can be reformulated as manifold smoothing on $\mathbb{M}_H$, where $\mathcal{F}(\cdot)$ acts as a manifold smoothing operator.

Then, we explore a scheme to construct $\mathcal{F}(\cdot)$. In practical I2P registration, $\mathbb{M}_I$ and $\mathbb{M}_P$ are discretized as graphs in a latent space. To approximate $\mathbb{M}_H$, we construct a graph $\mathcal{G}_H$ from $\{\mathbf{x}_i\}_{i=1}^M$ and $\{\mathbf{y}_j\}_{j=1}^N$. Thus, smoothing the manifold $\mathbb{M}_H$ can be reformulated as a graph smoothing problem on $\mathcal{G}_H$. Beltrami flow is a well-known technique for graph smoothing (Chamberlain et al., 2021a). Hence, based on the above analysis, we conclude that Beltrami flow can significantly enhance the capacity of I2P registration.

3.3 Beltrami Flow based Feature Interaction Layer

From Sec. 3.2, we have discussed that Beltrami flow is helpful for I2P registration. To leverage Beltrami flow for progressively repairing manifold alignment, we propose B-flow layers (i.e., a Beltrami flow based LFI module), and develop a prototype network for I2P registration, as shown in Fig. 4. According to the literature (Chamberlain et al., 2021b), a Beltrami flow on $\mathcal{G}_H$ is:

$$\frac{\partial \mathbf{z}[t]}{\partial t} = (\mathbf{A}(\mathbf{z}[t]) - \mathbf{I})\mathbf{z}[t] \tag{4}$$

$$\mathbf{z}[t] = (\mathbf{x}_1^T[t], ..., \mathbf{x}_M^T[t], \mathbf{y}_1^T[t], ..., \mathbf{y}_N^T[t])^T \tag{5}$$

$$\mathbf{A}(\mathbf{z}[t])_{ij} = a(\mathbf{z}[t]_i, \mathbf{z}[t]_j) \tag{6}$$

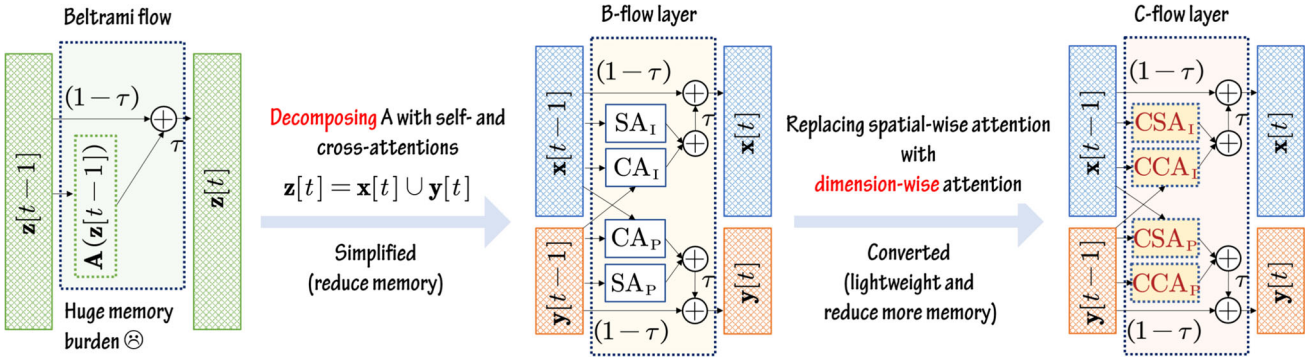

Fig. 5 Motivation and architecture of the proposed B-flow and C-flow layers. To reduce memory overhead associated with Beltrami flow, we decompose the matrix $\mathbf{A}$ using self- and cross-attention, resulting in the B-flow layer. To further reduce memory and parameter costs, we propose a covariance-based attention mechanism to replace the standard attention, leading to the more efficient C-flow layer

where $\mathbf{A}(\mathbf{z}[t])$ is an $(M+N)\times(M+N)$ attention symmetric matrix, which reflects the edge connections of vertices. $a(\cdot,\cdot)$ is a learnable function. $\mathbf{I}$ is an identity matrix. By forward Euler method, $\mathbf{z}[t]$ is solved as (Chamberlain et al., 2021b):

$$\begin{aligned}\mathbf{z}[t] &= (\mathbf{I}+\tau(\mathbf{A}(\mathbf{z}[t-1])-\mathbf{I}))\cdot\mathbf{z}[t-1] \\ &= (1-\tau)\cdot\mathbf{z}[t-1]+\tau\cdot\mathbf{A}(\mathbf{z}[t-1])\mathbf{z}[t-1]\end{aligned} \tag{7}$$

$$\mathbf{z}[0] = (\mathbf{x}_1^T,...,\mathbf{x}_M^T,\mathbf{y}_1^T,...,\mathbf{y}_N^T)^T \tag{8}$$

where τ is the time step (i.e., the discretisation parameter). Literature (Chamberlain et al., 2021b) shows that Eq. (7 is stable only if $\tau\in(0,1)$. Thus, the manifold smoothing operator $\mathcal{F}(\cdot)$ can be regarded as Eq. (7).

Although Eq. (7) provides an iterative solution to manifold smoothing, it has a huge computation burden in calculating $\mathbf{A}(\mathbf{z}[t])\mathbf{z}[t]$. $\mathcal{I}$ has the large number of pixels such that $M\approx 10^6$. $\mathbf{A}(\mathbf{z}[t])$ contains nearly 10^{12} elements ($\approx$ 100 GB for storage). To reduce the memory demand, we propose a B-flow layer to decompose $\mathbf{A}(\mathbf{z}[t])\mathbf{z}[t]$ using self-attention (SA) and cross-attention (CA):

$$\begin{aligned}\mathbf{A}(\mathbf{z}[t])\mathbf{z}[t] &= \begin{pmatrix}\mathbf{A}_{xx}(\mathbf{z}[t]) & \mathbf{A}_{xy}(\mathbf{z}[t])\\ \mathbf{A}_{xy}^T(\mathbf{z}[t]) & \mathbf{A}_{yy}(\mathbf{z}[t])\end{pmatrix}\begin{pmatrix}\mathbf{x}[t]\\ \mathbf{y}[t]\end{pmatrix} \\ &\approx \begin{pmatrix}\mathbf{A}_{xx}(\mathbf{x}[t]) & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_{yy}(\mathbf{y}[t])\end{pmatrix}\begin{pmatrix}\mathbf{x}[t]\\ \mathbf{y}[t]\end{pmatrix}+ \\ &\quad \begin{pmatrix}\mathbf{0} & \mathbf{A}_{xy}(\mathbf{z}[t])\\ \mathbf{A}_{xy}^T(\mathbf{z}[t]) & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{x}[t]\\ \mathbf{y}[t]\end{pmatrix} \\ &\approx \begin{pmatrix}\mathrm{SA}_I(\mathbf{x}[t])\\ \mathrm{SA}_P(\mathbf{y}[t])\end{pmatrix}+\begin{pmatrix}\mathrm{CA}_I(\mathbf{x}[t],\mathbf{y}[t])\\ \mathrm{CA}_P(\mathbf{y}[t],\mathbf{x}[t])\end{pmatrix}\end{aligned} \tag{9}$$

$$\begin{aligned}\mathrm{SA}_I(\mathbf{x}[t]) &= \mathrm{Softmax}\left(\frac{(\mathbf{x}[t]\mathbf{w}_x)(\mathbf{x}[t]\mathbf{w}_x)^T}{\sqrt{c}}\right)\cdot\mathbf{x}[t] \\ \mathrm{SA}_P(\mathbf{y}[t]) &= \mathrm{Softmax}\left(\frac{(\mathbf{y}[t]\mathbf{w}_y)(\mathbf{y}[t]\mathbf{w}_y)^T}{\sqrt{c}}\right)\cdot\mathbf{y}[t]\end{aligned} \tag{10}$$

$$\begin{aligned}\mathrm{CA}_I(\mathbf{x}[t],\mathbf{y}[t]) &= \mathrm{Softmax}\left(\frac{(\mathbf{x}[t]\mathbf{w}_x)(\mathbf{y}[t]\mathbf{w}_y)^T}{\sqrt{c}}\right)\cdot\mathbf{y}[t] \\ \mathrm{CA}_P(\mathbf{y}[t],\mathbf{x}[t]) &= \mathrm{Softmax}\left(\frac{(\mathbf{y}[t]\mathbf{w}_y)(\mathbf{x}[t]\mathbf{w}_x)^T}{\sqrt{c}}\right)\cdot\mathbf{x}[t]\end{aligned} \tag{11}$$

where $\mathbf{A}(\mathbf{z}[t])$ is decomposed into four parts. In Eq. (9), we simplify $\mathbf{A}_{xx}(\mathbf{z}[t])\mathbf{x}[t]$ as $\mathbf{A}_{xx}(\mathbf{x}[t])\mathbf{x}[t]$ and $\mathbf{A}_{yy}(\mathbf{z}[t])\mathbf{y}[t]$ as $\mathbf{A}_{yy}(\mathbf{y}[t])\mathbf{y}[t]$ by assuming that $\mathbf{y}[t]$ is independent with $\mathbf{x}[t]$. As $\mathbf{A}_{xx}(\mathbf{x}[t])\mathbf{x}[t]$ and $\mathbf{A}_{yy}(\mathbf{y}[t])\mathbf{y}[t]$ are feature computation with the inter-feature relations, we use SA layers to compute them via Eq. (10). As $\mathbf{A}_{xy}(\mathbf{z}[t])\mathbf{y}[t]$ and $\mathbf{A}_{xy}^T(\mathbf{z}[t])\mathbf{x}[t]$ are feature computation with the intra-feature relations, we use CA layers to compute them via Eq. (11). $\mathbf{w}_x\in\mathbb{R}^{c\times c}$ and $\mathbf{w}_y\in\mathbb{R}^{c\times c}$ in Eq. (10) are learnable matrices. The computation detail of the B-flow layer is summarized in Fig. 5. As the assumption that $\mathbf{x}[t]$ and $\mathbf{y}[t]$ are independent is just an approximation, Eq. (9) is an approximation to Eq. (7). In practical applications, to make sure the effect of Eq. (9) is close to Eq. (7), feature computation in Eq. (9) should run iteratively.

We further discuss computation advantage of Eq. (9). In fact, from Eq. (7) to Eq. (9), the attention matrix size is converted from $(M+N)^2$ to M^2+N^2 in Eq. (10) and $2MN$ in Eq. (11). Although the attention matrix is not reduced, Eq. (9) reformulates Beltrami flow as a parallel and interactive attention layers by approximating $\mathbf{A}_{xx}(\mathbf{z}[t])$ as $\mathbf{A}_{xx}(\mathbf{x}[t])$ and $\mathbf{A}_{yy}(\mathbf{z}[t])$ as $\mathbf{A}_{yy}(\mathbf{y}[t])$. It brings new possibilities for the further refinement of the Beltrami flow.

With the proposed B-flow layer, we establish a prototype Beltrami flow based I2P registration architecture, as shown in Fig. 4. It contains an image feature extractor $\mathbf{F}(\cdot)$, a point cloud feature extractor $\mathbf{G}(\cdot)$, a series of B-flow layers, and a non-learnable matching layer. The matching layer is used to establish 2D-3D correspondences with the normalized fea-

ture distances (Li et al., 2023). The detail of $\mathbf{F}(\cdot)$ and $\mathbf{G}(\cdot)$ is provided in Sec. 4.1.

3.4 Covariance-based Feature Interaction Layer

Although the B-flow layer presented in Fig. 5 seems like an ideal solution for I2P registration, it still needs much GPU memory, because the matrices, such as $(\mathbf{x}[t]\mathbf{w}_x)(\mathbf{x}[t]\mathbf{w}_x)^T \in \mathbb{R}^{M\times M}$ and $(\mathbf{x}[t]\mathbf{w}_x)(\mathbf{y}[t]\mathbf{w}_y)^T \in \mathbb{R}^{M\times N}$ in Eqs. (10) and (11), still have a large number of elements (i.e., $M \approx 10^6$ and $N \approx 10^5$). As most applications of I2P registration are based on robots or platforms with limited computational resources, it is necessary to design a lightweight LFI module from the B-flow layer. To address this issue, we introduce a covariance-based feature attention mechanism, resulting in the design of the C-flow layer, as the refinement of B-flow.

We explore an alternative to the attention matrices in Eqs. (10) and (11). As $(\mathbf{x}[t]\mathbf{w}_x)(\mathbf{x}[t]\mathbf{w}_x)^T$, $(\mathbf{y}[t]\mathbf{w}_y)(\mathbf{y}[t]\mathbf{w}_y)^T$, and $(\mathbf{x}[t]\mathbf{w}_x)(\mathbf{y}[t]\mathbf{w}_y)^T$ are the low-rank matrices (i.e. ranks are not greater than c), we may find another learnable matrices $\mathbf{v}_x$ and $\mathbf{v}_y$ so that Eq. (10) is equivalent to the following equation (detailed discussion is provided in Appendix. B):

$$\begin{aligned}&\text{Softmax}\left(\frac{(\mathbf{x}[t]\mathbf{w}_x)(\mathbf{x}[t]\mathbf{w}_x)^T}{\sqrt{c}}\right)\cdot\mathbf{x}[t]\\&=\mathbf{x}[t]\cdot\text{Softmax}\left(\frac{(\mathbf{x}[t]\mathbf{v}_x)^T(\mathbf{x}[t]\mathbf{v}_x)}{\sqrt{c}}\right)\end{aligned}$$

where attention type shifts from spatial-wise to dimensional-wise. Its advantage is to significantly reduce the attention size from $N\times N$ to $c\times c$ (N is 10^4 times greater than c). To avoid the negative impact of feature basis, we revise $\mathbf{x}[t]\mathbf{v}_x$ with zero-mean normalization, so that $(\mathbf{x}[t]\mathbf{v}_x)^T(\mathbf{x}[t]\mathbf{v}_x)$ is converted as $\text{Cov}(\mathbf{x}[t]\mathbf{v}_x)$ where $\text{Cov}(\cdot)$ is an operator to compute the covariance matrix. Based on the above thought, we propose the computation procedure of C-flow that consists of covariance-based SA (CSA) and CA (CCA) (for the best readability, we still use $\mathbf{w}_x$ and $\mathbf{w}_y$ instead of $\mathbf{v}_x$ and $\mathbf{v}_y$):

$$\begin{aligned}&\text{CSA}_I(\mathbf{x}[t])=\mathbf{x}[t]\cdot\text{Softmax}\left(\frac{\text{Cov}(\mathbf{x}[t]\mathbf{w}_x)}{\sqrt{c}}\right)\\&\text{CSA}_P(\mathbf{y}[t])=\mathbf{y}[t]\cdot\text{Softmax}\left(\frac{\text{Cov}(\mathbf{y}[t]\mathbf{w}_y)}{\sqrt{c}}\right)\end{aligned}\tag{12}$$

$$\begin{aligned}&\text{CCA}_I(\mathbf{x}[t],\mathbf{y}[t])=\\&\mathbf{y}[t]\cdot\text{Softmax}\left(\frac{\text{Cov}(\mathbf{x}[t]\mathbf{w}_x)\cdot(\text{Cov}(\mathbf{y}[t]\mathbf{w}_y))^\dagger}{\sqrt{c}}\right)\\&\text{CCA}_P(\mathbf{y}[t],\mathbf{x}[t])=\\&\mathbf{x}[t]\cdot\text{Softmax}\left(\frac{\text{Cov}(\mathbf{y}[t]\mathbf{w}_y)\cdot(\text{Cov}(\mathbf{x}[t]\mathbf{w}_x))^\dagger}{\sqrt{c}}\right)\end{aligned}\tag{13}$$

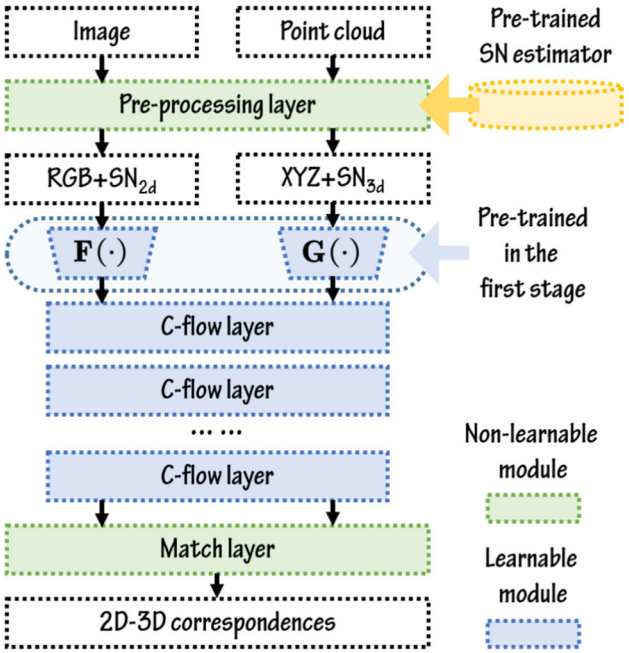

Fig. 6 Architecture of the proposed Flow-I2P. SN denotes surface normals. With the pretrained SN estimator, we can predict SN from a 2D image

where † denotes the general matrix inverse. The covariance matrix reflects feature dimensional correlation in the different dimensions. If $\mathbf{x}[t]$ and $\mathbf{y}[t]$ coincide in the same manifold with large overlap, $\text{Cov}(\mathbf{x}[t])$ and $\text{Cov}(\mathbf{y}[t])$ have the high similarity. This implies that the difference of $\text{Cov}(\mathbf{x}[t])$ and $\text{Cov}(\mathbf{y}[t])$ is a cue for manifold stitching of $\mathbb{M}_I$ and $\mathbb{M}_P$. That is the reason why the CSA layer is designed according to Eq. (13).

By comparing Eqs. (12) and (13) to Eqs. (10) and (11), we observe a substantial reduction in memory usage from $O(MN)$ to $O(c^2)$. As c is nearly 10^2, $c^2 \approx MN\cdot 10^{-7}$ (i.e., $c^2 \ll MN$). CCA saves 10^7 times memory than the standard CA layer. It makes the proposed Beltrami flow based I2P model feasible for practical applications.

3.5 Practical Beltrami Flow based I2P registration

To achieve optimal registration performance with C-flow layers, we have developed the Flow-I2P network, shown in Fig. 6. It consists of the pre-processing layer, $\mathbf{F}(\cdot)$, $\mathbf{G}(\cdot)$, T stacked C-flow layers, and the matching layer. Compared with the prototype architecture in Fig. 4, the major differences of Flow-I2P lie in (i) feature pre-processing and (ii) training scheme.

We first illustrate the pre-processing step. It is recalled in Sec. 3.2 that the **manifold alignment prior condition** is a crucial support to convert manifold alignment into manifold smoothing. To satisfy this condition, we attempt to enrich the

input features from $(\mathcal{I}, \mathcal{P})$ to $(\mathcal{I}_{\text{aug}}, \mathcal{P}_{\text{aug}})$ via adding surface normals:

$$\mathcal{I}_{\text{aug}} = \mathcal{I} \oplus \text{SN}_{2d}, \ \ \mathcal{P}_{\text{aug}} = \mathcal{P} \oplus \text{SN}_{3d} \tag{14}$$

$$\text{SN}_{2d} = \text{SN\_Pred}(\mathcal{I}) \tag{15}$$

where $\oplus$ is the feature concatenation operation. $\text{SN}_{2d} \in \mathbb{R}^{H \times W \times 3}$ is surface normals predicted from $\mathcal{I}$ with a pre-trained model SN_Pred$(\cdot)$. $\text{SN}_{3d} \in \mathbb{R}^{N \times 3}$ denotes the surface normals of the point cloud. Surface normals of image and point cloud can constrain 2D-3D correspondences (reason is provided in Appendix D), so that we leverage surface normals as the extra features in 2D pixels and 3D points. Using $(\mathcal{I}_{\text{aug}}, \mathcal{P}_{\text{aug}})$, the discriminative ability of $\mathbf{F}(\cdot)$ and $\mathbf{G}(\cdot)$ is significantly increased (see ablation studies in Sec. 4.3), ensuring a high-quality manifold alignment of $\mathbb{M}_I[0] \cap \mathbb{M}_P[0]$. Eqs. (14) and (15) denote the pre-processing layer illustrated in Fig. 6.

Next, we illustrate the training scheme of Flow-I2P. It is recalled that $\delta[0]$ is the upper bound of feature distance of correct correspondences in $\mathbb{M}_I[0] \cap \mathbb{M}_P[0]$. A smaller $\delta[0]$ significantly enhances the effectiveness of manifold smoothing. In this thought, pre-training $\mathbf{F}(\cdot)$ and $\mathbf{G}(\cdot)$ to minimize $\delta[0]$ is another way to satisfy the **manifold alignment prior condition**. Thus, we design a two-stage training scheme for Flow-I2P. In the first stage, only $\mathbf{F}(\cdot)$ and $\mathbf{G}(\cdot)$ are pretrained by minimizing the following function:

$$\min_{\Theta_{\mathbf{F}}, \Theta_{\mathbf{G}}} \mathbb{E}_{(\mathcal{I}, \mathcal{P}) \sim \mathcal{D}} L_{\text{I2P}}(\mathbf{x}[0], \mathbf{y}[0] | \mathcal{C}_{\text{GT}}) \tag{16}$$

where L_{I2P} is a standard loss for I2P registration (Li et al., 2023). $(\mathcal{I}, \mathcal{P}) \sim \mathcal{D}$ denotes a pair of image and point cloud randomly sampled from the training dataset $\mathcal{D}_t$. $\mathbb{E}$ is the expectation operator. $\Theta_{\mathbf{F}}$ and $\Theta_{\mathbf{G}}$ are learnable parameters of $\mathbf{F}(\cdot)$ and $\mathbf{G}(\cdot)$. As $\mathbb{M}_I$ and $\mathbb{M}_P$ are generated from $\mathbf{x}[0]$ and $\mathbf{y}[0]$, Eq. (16) ensures that $\mathbb{M}_I$ and $\mathbb{M}_P$ are aligned with a certain degree of accuracy. In the second stage, Flow-I2P is fine-tuned by minimizing the following function:

$$\min_{\Theta} \mathbb{E}_{(\mathcal{I}, \mathcal{P}) \sim \mathcal{D}} L_{\text{I2P}}(\mathbf{x}[T], \mathbf{y}[T] | \mathcal{C}_{\text{GT}}) \tag{17}$$

where Θ denotes the whole parameters in Flow-I2P. This coarse-to-fine training scheme helps stabilize the learning of 2D-3D correspondences.

4 Experiments

This section provides implementation details and evaluation metrics of the proposed Flow-I2P method. We then conduct ablation studies to evaluate the effectiveness of Flow-I2P modules, followed by comprehensive comparisons on five indoor and outdoor datasets. Finally, the performance of Flow-I2P is analyzed.

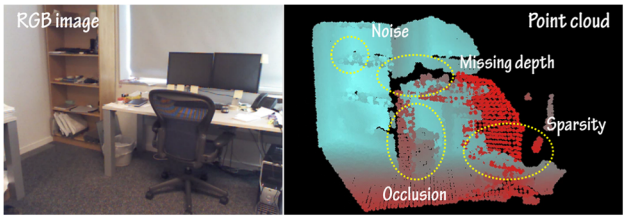

Fig. 7 An example of the images and point clouds in the datasets used in our experiments. All datasets are collected from real-world environments and inherently exhibit challenges such as occlusions, noise, and missing depth values

4.1 Implementation Details

Due to the limited number of published works on I2P registration, we implement Flow-I2P mainly based on 2D3D-MATR (Li et al., 2023)[2].

Network architecture. The proposed Flow-I2P requires a surface normal estimation model, for which we use the pre-trained DSINE model (Bae & Davison, 2024) to estimate surface normals from a monocular RGB image. Flow-I2P also requires $\mathbf{F}(\cdot)$ and $\mathbf{G}(\cdot)$ to extract features from images and point clouds, respectively. Following 2D3D-MATR (Li et al., 2023), $\mathbf{F}(\cdot)$ is a 4-stage feature pyramid network (FPN) based on ResNet (He et al., 2016), while $\mathbf{G}(\cdot)$ is a 4-stage FPN based on KPConv (Thomas et al., 2019). The training loss L_{I2P} employs circle loss (Sun et al., 2020), which has proven effective in I2P registration (Wang et al., 2021; Li et al., 2023). The hyperparameters for circle loss are consistent with those in P2-Net (Wang et al., 2021). In Flow-I2P, we do not refine $\mathbf{F}(\cdot)$, $\mathbf{G}(\cdot)$, or the circle loss, as they have already been optimized for I2P registration in previous work (Li et al., 2023). However, despite their optimization, there is still significant room for improvement in I2P registration performance. So, we focus on cross-modality feature interaction. In Flow-I2P, the channel number c is set to 256. The key hyper-parameters τ and the number of stacked layers are examined in the ablation studies. The training and testing configurations follow those of 2D3D-MATR (Li et al., 2023). The batch size is set to 1, and the model is trained with adaptive moments estimation (Adam) optimizer. Flow-I2P exploits the two-stage training strategy. The first stage has 20 epochs with a learning rate of 10^{-4}, and the second stage has 10 epochs with a learning rate of 0.5×10^{-4}. All experiments were performed on a single NVIDIA GeForce RTX 3080 GPU. In all datasets, the input point clouds are down-sampled

[2] https://github.com/minhaolee/2d3dmatr

Table 1 Results of learning-based I2P registration models on the RGB-D Scenes V2 testing dataset. ↑ denotes that a higher value indicates better performance. † † denotes that the I2P registration method requires depth information from RGB-D camera. The Bolditalic highlight indicates the best average of the I2P registration method without requiring physical scale depth

Model	Scene-11	Scene-12	Scene-13	Scene-14	Mean
Inlier Ratio ↑					
FCGF (Choy et al., 2019)	6.8%	8.5%	11.8%	5.4%	8.1%
Predator (Huang et al., 2021)	17.7%	19.4%	17.2%	8.4%	15.7%
P2-Net (Wang et al., 2021)	9.7%	12.8%	17.0%	9.3%	12.2%
2D3D-MATR (Li et al., 2023)	32.8%	34.4%	39.2%	23.3%	32.4%
FreeReg (Wang et al., 2024)	36.6%	34.5%	34.2%	18.2%	30.9%
Diff-Reg† (Wu et al., 2024)	47.2%	48.7%	32.9%	22.4%	37.8%
Bridge (Cheng et al., 2025)	36.4%	32.7%	43.8%	27.4%	35.1%
Flow-I2P (Ours)	***49.6%***	***44.0%***	***36.5%***	***30.4%***	***40.1%***
Feature Matching Recall ↑					
FCGF (Choy et al., 2019)	11.1%	30.4%	51.5%	15.5%	27.1%
Predator (Huang et al., 2021)	86.1%	89.2%	63.9%	24.3%	65.9%
P2-Net (Wang et al., 2021)	48.6%	65.7%	82.5%	41.6%	59.6%
2D3D-MATR (Li et al., 2023)	98.6%	98.0%	88.7%	77.9%	90.8%
FreeReg (Wang et al., 2024)	91.9%	93.4%	93.1%	49.6%	82.0%
Diff-Reg† (Wu et al., 2024)	100.0%	100.0%	88.7%	77.0%	91.4%
Bridge (Cheng et al., 2025)	***100.0%***	***99.0%***	***92.8%***	***85.8%***	***94.4%***
Flow-I2P (Ours)	100.0%	100.0%	94.5%	78.7%	93.3%
Registration Recall ↑					
FCGF (Choy et al., 2019)	26.4%	41.2%	37.1%	16.8%	30.4%
Predator (Huang et al., 2021)	44.4%	41.2%	21.6%	13.7%	30.2%
P2-Net (Wang et al., 2021)	40.3%	40.2%	41.2%	31.9%	38.4%
2D3D-MATR (Li et al., 2023)	63.9%	53.9%	58.8%	49.1%	56.4%
FreeReg (Wang et al., 2024)	74.2%	72.5%	54.5%	27.9%	57.3%
Diff-Reg† (Wu et al., 2024)	98.6%	96.1%	83.5%	63.7%	85.5%
Bridge (Cheng et al., 2025)	58.3%	60.8%	74.2%	60.2%	63.4%
Flow-I2P (Ours)	***90.0%***	***65.9%***	***54.8%***	***63.0%***	***68.4%***

with a voxel size of 1.5 cm, where the maximum number of points is 3×10^5.

Evaluation metrics. We used three primary metrics to evaluate the overall performance of I2P registration. (i) IR measures the proportion of 2D-3D correspondences with a 3D distance of less than 5 cm; (ii) feature matching recall (FMR) calculates the ratio of image-point-cloud pairs with IR values greater than 10%; (iii) RR assesses the ratio of image-point-cloud pairs with root mean square errors (RMSE) are below 10 cm. Additionally, we compute relative translational error (RTE, in cm) and relative rotational error (RRE, in deg). With the 2D-3D correspondences predicted by the I2P registration model, RR, RTE, and RRE are computed using a RANSAC-based PnP algorithm (Lepetit et al., 2009).

Datasets. Given the close relationship between I2P registration and indoor visual localization, we evaluate I2P registration performances on three indoor datasets: RGB-D V2 (Lai et al., 2014), 7-Scenes (Glocker et al., 2013), and ScanNet V2 (Dai et al., 2017). These datasets feature continuous RGB-D frames with accurate 6-DoF pose annotations. We adopt the same data split as in previous work (Li et al., 2023) for the RGB-D V2 and 7-Scenes datasets. On the RGB-D V2 dataset, scenes 1-8 are used for training, scenes 9-10 for validation, and scenes 11-14 for testing. Pairs of images and point clouds with an overlap ratio of at least 30% are used as samples in the RGB-D V2 dataset. On the 7-Scenes dataset, 18,228 training pairs, 478 validation pairs, and 9,554 testing pairs are selected, with an overlap ratio threshold of 50%. On the ScanNet V2 dataset, only 10% of the data is used for fine-tuning, while 90% of the data is used for testing. Besides, we record the real-world RGB-D data as a self-collected dataset using an Intel RealSense camera. It is used to evaluate performance in real-world environments. Similar to the ScanNet V2 dataset, 10% of the data is used for fine-tuning, while 90% is used for testing. Overall, all datasets are collected in the real world (seen in Fig. 7). Point cloud inevitably has sensor noise, occlusion, sparsity, and missing depths, which pose challenges to all I2P registration methods.

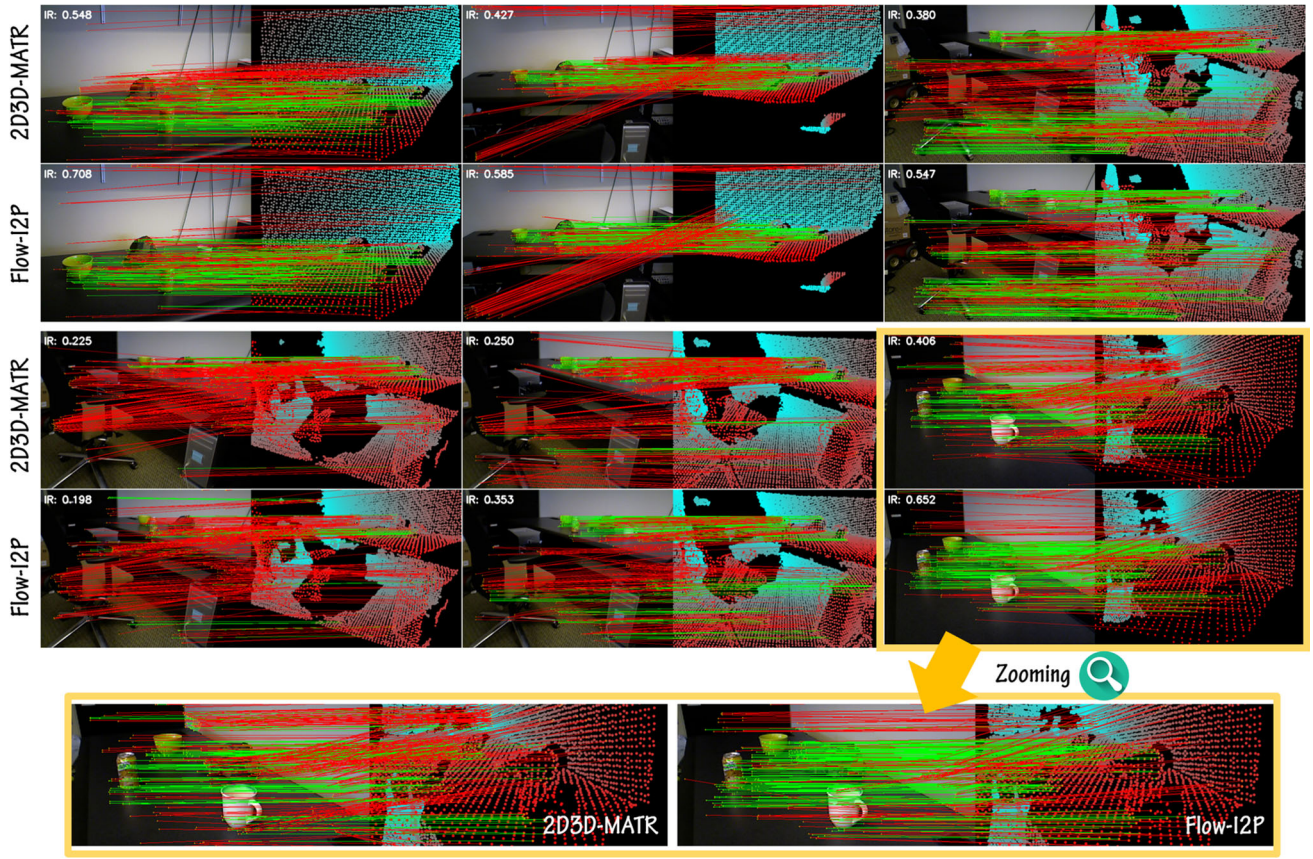

Fig. 8 Qualitative comparison between the proposed Flow-I2P and 2D3D-MATR (Li et al., 2023) on Scene-14 of the RGB-D V2 dataset. Scene-14 represents the greatest challenge due to low image texture and sparse point clouds. Flow-I2P demonstrates greater robustness and reliability under these difficult conditions

4.2 Comparisons

We compare the proposed Flow-I2P with state-of-the-art I2P registration methods across five different indoor and outdoor datasets. After that, we assess the generalization ability of these methods in cross-dataset scenarios.

RGB-D V2 dataset. This experiment evaluates the performance of Flow-I2P and state-of-the-art I2P methods on the RGB-D V2 test dataset (Lai et al., 2014). The methods compared to include **P2-Net** (Wang et al., 2021), **2D3D-MATR** (Li et al., 2023), **FreeReg** (Wang et al., 2024), **Diff-Reg** (Wu et al., 2024), and **Bridge** (Cheng et al., 2025). We also compare Flow-I2P with point cloud registration methods, **FCGF** (Choy et al., 2019) and **Predator** (Huang et al., 2021), both of which can be extended to the I2P registration task. Table 1 presents the experimental results showing that **Flow-I2P** outperforms all other methods in this dataset. **Flow-I2P** achieves a 7.7% higher IR than **2D3D-MATR** and an 11.1% higher RR than **FreeReg**. It also achieves a 5.0% higher IR and 5.0% higher RR than recently published work **Bridge** (Cheng et al., 2025). Compared to **2D3D-MATR**, **Flow-I2P** utilizes LFI modules to correct the manifold structures of $\mathbb{M}_I$ and $\mathbb{M}_P$, thus enhancing the generalization ability of cross-modality feature matching. Unlike **FreeReg**, **Flow-I2P** relies solely on surface normals to eliminate cross-modality feature differences, instead of a large-scale diffusion pretrained model. It achieves a more accurate result than **FreeReg**. Unlike other methods, **Diff-Reg** (Wu et al., 2024) requires the physical-scale depth from RGB-D data to filter incorrect 2D-3D correspondences in the diffusion procedure. In fact, the standard I2P registration task generally does not use physical-scale depth as input. In this context, **Diff-Reg** achieves a significantly higher RR than **Flow-I2P**. Compared to **Diff-Reg**, the proposed **Flow-I2P** offers three advantages: (i) it works without physical-scale depth; (ii) it achieves higher IR than **Diff-Reg**; (iii) its runtime is nearly 3× faster than **Diff-Reg** (seen in Fig. 2). Compared to **Bridge** (Cheng et al., 2025), the proposed C-flow layers outperform the uncertainty estimation layers in **Bridge**, as **Flow-I2P** significantly improves IR and RR than **Bridge**.

As shown in Table 1, Scene-14 is the most challenging; therefore, we visualize the correspondence quality of

Table 2 Results of learning-based I2P registration models trained with the complete training data on the 7-Scenes testing dataset. ↑ denotes that a higher value indicates better performance. † denotes that the I2P registration method requires depth information from RGB-D camera. The Bolditalic highlight indicates the best average of the I2P registration method without requiring physical scale depth

Model	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Mean
Inlier Ratio ↑								
FCGF (Choy et al., 2019)	34.2%	32.8%	14.8%	26.0%	23.3%	22.5%	6.0%	22.8%
Predator (Huang et al., 2021)	34.7%	33.8%	16.6%	25.9%	23.1%	22.2%	7.5%	23.4%
P2-Net (Wang et al., 2021)	55.2%	46.7%	13.0%	36.2%	32.0%	32.8%	5.8%	31.7%
2D3D-MATR (Li et al., 2023)	72.1%	66.0%	31.3%	60.7%	50.2%	52.5%	18.1%	50.1%
Diff-Reg† (Wu et al., 2024)	73.3%	60.8%	45.5%	63.1%	47.8%	53.3%	20.4%	52.0%
Bridge (Cheng et al., 2025)	73.8%	66.7%	33.1%	61.7%	50.8%	52.3%	18.1%	50.9%
Flow-I2P (Ours)	***76.7%***	***64.7%***	***37.1%***	***62.0%***	***52.3%***	***52.8%***	***18.5%***	***52.0%***
Feature Matching Recall ↑								
FCGF (Choy et al., 2019)	99.7%	98.2%	69.9%	97.1%	83.0%	87.7%	16.2%	78.8%
Predator (Huang et al., 2021)	91.3%	95.1%	76.7%	88.6%	79.2%	80.6%	31.1%	77.5%
P2-Net (Wang et al., 2021)	100.0%	99.3%	58.9%	99.1%	87.2%	92.2%	16.2%	79.0%
2D3D-MATR (Li et al., 2023)	100.0%	99.6%	98.6%	100.0%	92.4%	95.9%	58.1%	92.1%
Diff-Reg† (Wu et al., 2024)	100.0%	98.5%	97.3%	100.0%	87.8%	96.8%	60.8%	91.6%
Bridge (Cheng et al., 2025)	***100.0%***	***100.0%***	***98.6%***	***100.0%***	***92.7%***	***95.6%***	***64.9%***	***93.1%***
Flow-I2P (Ours)	100.0%	99.7%	95.1%	99.9%	93.1%	96.8%	56.7%	91.6%
Registration Recall ↑								
FCGF (Choy et al., 2019)	89.5%	79.7%	19.2%	85.9%	69.4%	79.0%	6.8%	61.4%
Predator (Huang et al., 2021)	69.6%	60.7%	17.8%	62.9%	56.2%	62.6%	9.5%	48.5%
P2-Net (Wang et al., 2021)	96.9%	86.5%	20.5%	91.7%	75.3%	85.2%	4.1%	65.7%
2D3D-MATR (Li et al., 2023)	98.8%	87.1%	46.7%	93.2%	77.1%	87.2%	38.2%	75.5%
Diff-Reg† (Wu et al., 2024)	99.3%	94.3%	91.8%	99.1%	79.9%	91.8%	25.7%	83.1%
Bridge (Cheng et al., 2025)	98.3%	90.5%	56.2%	96.4%	84.0%	86.1%	32.4%	77.7%
Flow-I2P (Ours)	***98.8%***	***90.0%***	***58.4%***	***93.9%***	***82.1%***	***88.6%***	***37.6%***	***78.4%***

Flow-I2P and its baseline method in Fig. 8. We observe that **Flow-I2P** demonstrates greater robustness under the extreme input condition, which is attributed to its effective feature interaction module. Thus, **Flow-I2P** demonstrates a significant improvement over existing methods on the RGB-D V2 test dataset.

7-Scenes dataset. This experiment examines the performance of Flow-I2P and state-of-the-art I2P methods on the 7-Scenes test dataset (Glocker et al., 2013). The comparison results are provided in Table 2. The proposed **Flow-I2P** still achieves the best registration performance, because it leverages Beltrami flow and thus effectively reduces the modality gap between images and point clouds. In the degraded scene **Stairs**, **Flow-I2P** achieves an RR that is 5.2% higher than **Bridge** (Cheng et al., 2025).

In practical I2P registration applications, most inference scenes are either unseen or significantly different from the training dataset. To assess the generalization ability of these methods, we conduct comparisons using only 5% of the training samples. The results are presented in Table 3. **2D3D-MATR+SN** represents the original method using surface normals as input. We find that **Flow-I2P** significantly outperforms **2D3D-MATR**. With limited training samples, $\mathbb{M}_I[0]$ and $\mathbb{M}_P[0]$ align with large errors, and the proposed C-flow layer effectively corrects these errors. When training samples are sufficient and closely match the testing samples, $\mathbb{M}_I[0]$ and $\mathbb{M}_P[0]$ align with small errors, resulting in limited performance gains from the C-flow layer. This explains why **Flow-I2P** shows significant improvement in Table 3 but only minor gains in Table 2. From Fig. 9, the proposed **Flow-I2P** generates more inliers than other methods, even in challenging scenes.

ScanNet dataset. This experiment evaluates the performance of Flow-I2P on the ScanNet dataset (Dai et al., 2017). In this dataset, 10% of the data is used for training, and 90% for testing. All approaches are pretrained on the 7-Scenes dataset and fine-tuned on ScanNet. The comparison results are presented in Table 4. State-of-the-art methods, such as **FreeReg+PnP** (Wang et al., 2024), **LCD+PnP** (Pham et al., 2020), and **SuperGlue+PnP** (Sarlin et al., 2020), are included for comparison. The performance of these methods is taken from the literature (Wang et al., 2024) and it

Table 3 Results of learning-based I2P registration models trained with only 5% of the training data on the 7-Scenes testing dataset. The Bolditalic highlight indicates the best average. ↑ denotes that a higher value indicates better performance, while ↓ indicates that a smaller value is better

Model	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Mean
Inlier Ratio ↑								
2D3D-MATR (Li et al., 2023)	42.1%	23.9%	9.0%	24.4%	22.2%	22.3%	5.1%	21.3%
2D3D-MATR+SN	47.5%	32.0%	10.2%	33.9%	22.6%	24.3%	8.5%	25.6%
Flow-I2P (Ours)	***52.8%***	***37.6%***	***12.4%***	***31.4%***	***26.2%***	***29.1%***	***7.9%***	***28.2%*** (↑***6.9%***)
Feature Matching Recall ↑								
2D3D-MATR (Li et al., 2023)	99.5%	82.6%	34.5%	88.0%	78.7%	82.3%	14.5%	68.6%
2D3D-MATR+SN	99.6%	90.2%	44.4%	94.7%	84.0%	89.9%	36.4%	77.0%
Flow-I2P (Ours)	***99.7%***	***93.4%***	***42.7%***	***94.4%***	***91.2%***	***91.8%***	***35.2%***	***78.4%*** (↑***9.8%***)
Registration Recall ↑								
2D3D-MATR (Li et al., 2023)	71.9%	35.8%	2.2%	60.1%	46.4%	54.3%	10.9%	40.2%
2D3D-MATR+SN	87.9%	43.9%	2.7%	63.1%	46.8%	71.9%	10.4%	46.6%
Flow-I2P (Ours)	***89.0%***	***41.3%***	***0.7%***	***64.5%***	***51.4%***	***70.5%***	***7.2%***	***47.3%*** (↑***7.1%***)
Mean RRE ↓								
2D3D-MATR (Li et al., 2023)	2.791	4.027	9.717	2.910	3.194	3.124	3.019	4.112
2D3D-MATR+SN	2.232	3.767	6.680	2.529	2.893	2.961	3.154	3.459
Flow-I2P (Ours)	***2.616***	***3.979***	***4.735***	***3.067***	***2.884***	***2.873***	***3.183***	***3.334*** (↓***18.9%***)
Mean RTE ↓								
2D3D-MATR (Li et al., 2023)	0.074	0.112	0.138	0.101	0.120	0.103	0.100	0.107
2D3D-MATR+SN	0.058	0.103	0.105	0.084	0.106	0.091	0.122	0.096
Flow-I2P (Ours)	***0.070***	***0.117***	***0.052***	***0.106***	***0.100***	***0.088***	***0.095***	***0.090*** (↓***15.9%***)

is included in Table 4. It should be noted that the IR values from the literature (Wang et al., 2024) use a different threshold (i.e., 0.3m instead of the standard 0.05m), and these results are marked with a †. With the proposed C-flow layer, **Flow-I2P** outperforms existing approaches. Visualization of I2P registration is presented in Fig. 10. In most indoor scenarios, **Flow-I2P** achieves 2D-3D correspondences with more inliers than **2D3D-MATR**. This indicates that **Flow-I2P** delivers the best performance on the ScanNet dataset with only 10% training samples for fine-tuning.

Self-collected dataset. This experiment evaluates the performance of Flow-I2P on a self-collected dataset. Point clouds and RGB images were collected in 8 indoor office scenarios using the Intel RealSense D435i sensor, as shown in Fig. 11. Due to measurement errors, the 3D point clouds contain some noise and incomplete regions but have a higher density than other datasets. On this dataset, 10% of the samples are utilized for training, and 90% for testing. All methods are pretrained on the 7-Scenes dataset and fine-tuned on this dataset. The visualization of I2P registration is provided in Fig. 12. We observe that the 2D-3D correspondence distribution is sparse in scenes with nearby objects but becomes denser in scenes with distant objects. This phenomenon is likely influenced by the pretrained dataset 7-Scenes, which consists mainly of scenes with greater distances between objects. This suggests significant room for improving I2P registration. The comparison results are presented in Table 5. **Flow-I2P** achieves the highest IR and RR. To verify the robustness of I2P registration, we compare these methods with varying overlap ratios between images and point clouds. This is achieved by randomly masking a portion of the region in images. The results are provided in Fig. 13. **Flow-I2P** outperforms other methods with different overlap ratios. These results indicate that **Flow-I2P** has a higher generalization ability than **2D3D-MATR** (Li et al., 2023).

KITTI dataset. This experiment evaluates the registration performance of Flow-I2P and current I2P registration methods on the outdoor KITTI dataset (Geiger et al., 2012). **DeepI2P** (Li & Lee, 2021), **CorrI2P** (Ren et al., 2023), **VP2P-Match** (Zhou et al., 2023), **CoFiI2P** (Kang et al., 2024), **CMR-Agent** (Yao et al., 2024), and **OL-Reg** (An et al., 2024b) are used for comparison. Unlike the indoor datasets, KITTI has two challenges: (i) the sparsity and missing values of LiDAR point cloud, and (ii) the less overlap region between LiDAR point clouds and RGB images. These challenges make it difficult to directly train both the original **2D3D-MATR** (Li et al., 2023) and **Flow-I2P** on this dataset. To address this problem, we utilize the pretrained CorrI2P to preprocess the LiDAR point clouds to gather the overlapped LiDAR point clouds. Then, the images and processed point clouds are used as input in **2D3D-MATR** and **Flow-I2P**. The average relative translational error (RTE) and average rela-

Table 4 Results of learning-based I2P registration models on the ScanNet testing dataset. The Bolditalic highlight indicates the best average. ↑ means that a higher value indicates better performance, while ↓ indicates that a smaller value is better. † denotes that the 3D distance threshold for 2D-3D correspondence is set to 0.3m instead of the default 0.05m used in the experiments

Model	Scene-1	Scene-2	Scene-3	Scene-4	Scene-5	Scene-6	Scene-7	Scene-8	Mean
Inlier Ratio ↑									
FreeReg+PnP (Wang et al., 2024)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	56.8%†
LCD+PnP (Pham et al., 2020)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	30.7%†
SuperGlue+PnP (Sarlin et al., 2020)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	13.4%†
2D3D-MATR (Li et al., 2023)	33.4%	26.2%	24.9%	13.2%	32.8%	24.5%	39.0%	32.9%	28.4%
2D3D-MATR+SN	42.2%	32.0%	33.9%	19.3%	40.7%	29.8%	49.7%	40.3%	35.9%
Flow-I2P (Ours)	***43.9%***	***31.9%***	***35.6%***	***20.4%***	***42.0%***	***30.7%***	***51.0%***	***42.0%***	***37.2%***
Flow-I2P (Ours)	***94.5%***	***92.7%***	***88.6%***	***85.2%***	***93.5%***	***88.2%***	***96.2%***	***92.1%***	***91.4%***†
Feature Matching Recall ↑									
FreeReg+PnP (Wang et al., 2024)	***N/A***	***N/A***	***N/A***	***N/A***	***N/A***	***N/A***	***N/A***	***N/A***	***98.5%***
LCD+PnP (Pham et al., 2020)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	55.1%
SuperGlue+PnP (Sarlin et al., 2020)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	53.2%
2D3D-MATR (Li et al., 2023)	100.0%	100.0%	100.0%	40.0%	95.8%	90.9%	100.0%	100.0%	90.8%
2D3D-MATR+SN	100.0%	100.0%	100.0%	70.0%	95.8%	100.0%	100.0%	100.0%	95.7%
Flow-I2P (Ours)	100.0%	100.0%	100.0%	80.0%	96.8%	100.0%	100.0%	100.0%	97.1%
Registration Recall ↑									
FreeReg+PnP (Wang et al., 2024)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	78.0%
LCD+PnP (Pham et al., 2020)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.0%
SuperGlue+PnP (Sarlin et al., 2020)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	1.2%
2D3D-MATR (Li et al., 2023)	88.9%	83.3%	80.0%	30.0%	87.5%	72.7%	100.0%	86.7%	78.6%
2D3D-MATR+SN	89.0%	84.2%	80.0%	40.0%	86.4%	90.9%	100.0%	94.3%	83.1%
Flow-I2P (Ours)	***100.0%***	***88.9%***	***90.0%***	***40.0%***	***87.5%***	***72.7%***	***100.0%***	***96.7%***	***84.5%***

Table 5 Results of learning-based I2P registration models on the self-collected testing dataset. The Bolditalic highlight indicates the best average. ↑ means that a higher value indicates better performance

Model	Scene-1	Scene-2	Scene-3	Scene-4	Scene-5	Scene-6	Scene-7	Scene-8	Mean
Inlier Ratio ↑									
2D3D-MATR (Li et al., 2023)	50.9%	52.1%	66.2%	31.8%	52.4%	33.0%	59.0%	48.6%	49.3%
2D3D-MATR+SN	55.0%	58.2%	66.9%	33.1%	53.3%	35.9%	58.0%	53.5%	51.8%
Flow-I2P (Ours)	***57.9%***	***56.7%***	***68.6%***	***34.6%***	***54.4%***	***37.1%***	***60.3%***	***54.8%***	***53.0%***
Registration Recall ↑									
2D3D-MATR (Li et al., 2023)	77.8%	83.3%	94.4%	94.4%	100.0%	95.2%	94.4%	90.9%	91.3%
2D3D-MATR+SN	81.2%	90.8%	98.2%	95.4%	100.0%	95.2%	96.4%	90.9%	93.5%
Flow-I2P (Ours)	***88.9%***	***94.4%***	***100.0%***	***94.4%***	***100.0%***	***95.2%***	***100.0%***	***90.9%***	***95.5%***

Table 6 I2P registration performance on the KITTI dataset. The Bolditalic highlight indicates the best RTE and RRE

Methods	Publish information	RTE/m	RRE/deg
DeepI2P (Li & Lee, 2021)	CVPR 2021	1.460	4.270
CorrI2P (Ren et al., 2023)	IEEE T-CSVT 2022	0.740	2.070
VP2P-match (Zhou et al., 2023)	NeurIPS 2023	0.750	3.290
2D-3D MATR (Li et al., 2023)	ICCV 2023	0.020	0.275
CoFiI2P (Kang et al., 2024)	IEEE RAL 2024	0.290	1.140
CMR-Agent (Yao et al., 2024)	IROS 2024	0.195	0.589
Flow-I2P (Ours)		***0.012***	***0.251***

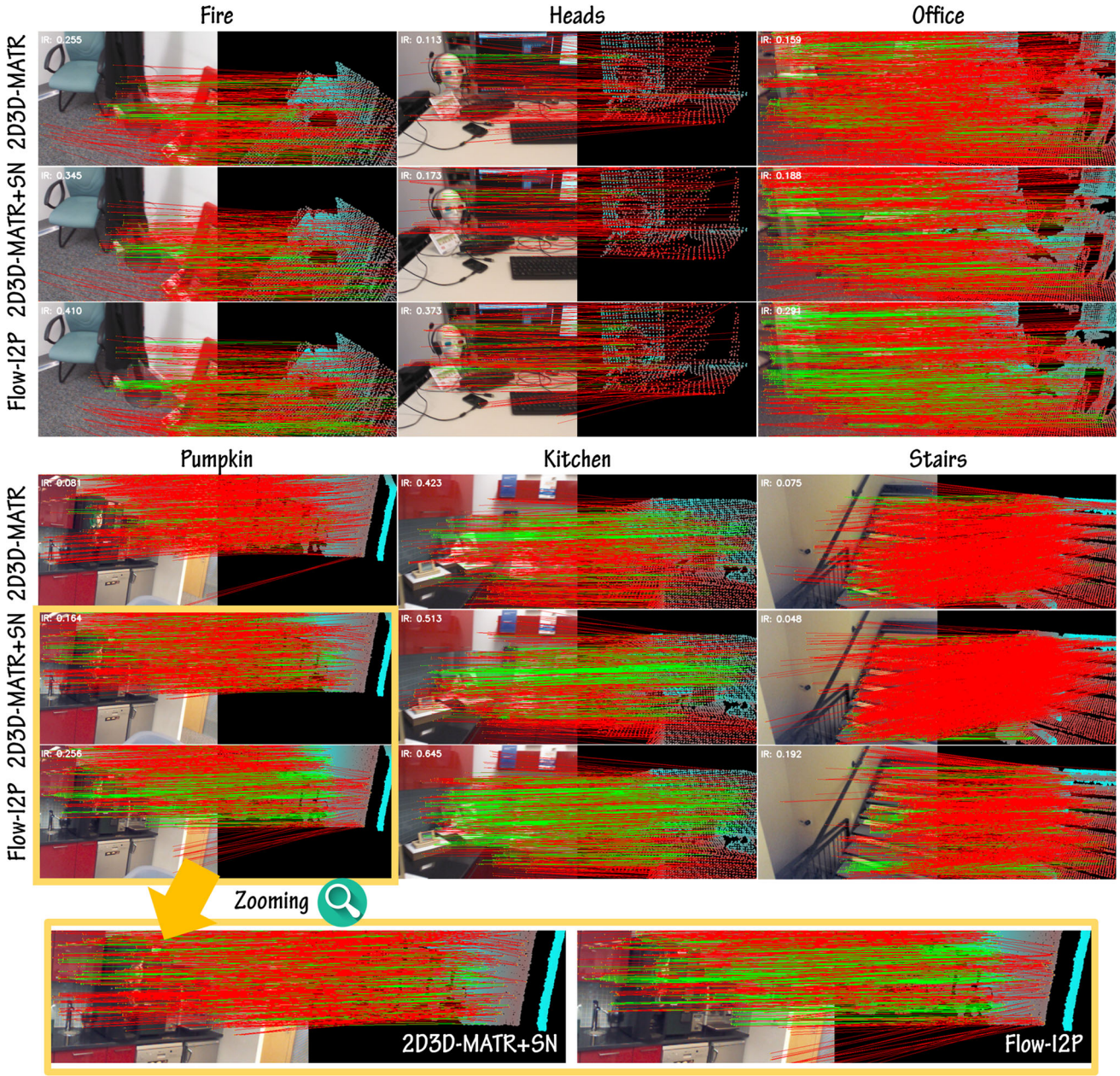

Fig. 9 Visualization of the proposed Flow-I2P, 2D3D-MATR (Li et al., 2023), and its refinement, 2D3D-MATR+SN in challenging scenes. All methods are trained using only 5% of the training samples. Flow-I2P generates more 2D-3D correct correspondences than the other methods

tive rotation error (RRE) are used as metrics. Results in Table 6 demonstrate that the proposed **Flow-I2P** achieves the best performance on the KITTI dataset, showing strong generalization by effectively bridging the modality gap between image and sparse LiDAR features via Beltrami flow.

Noise robustness test. In real-world applications, noise in point clouds can negatively affect I2P registration and its downstream tasks. Such noise often arises from depth sensing inaccuracies or errors in 3D reconstruction. Although existing datasets (e.g., 7-Scenes, RGBD-v2, ScanNet, and our self-collected dataset) already contain some measurement noise, we conduct a dedicated robustness test to evaluate model performance under extreme noise conditions. Specifically, uniformly distributed noise in the range $[-\delta_p, \delta_p]$ is added to the 3D point coordinates. The results, shown in Table 7, indicate that stacking C-flow layers significantly enhances manifold alignment quality, even in the presence of severe noise. This demonstrates that the proposed Flow-I2P is robust to point cloud noise in the I2P registration task.

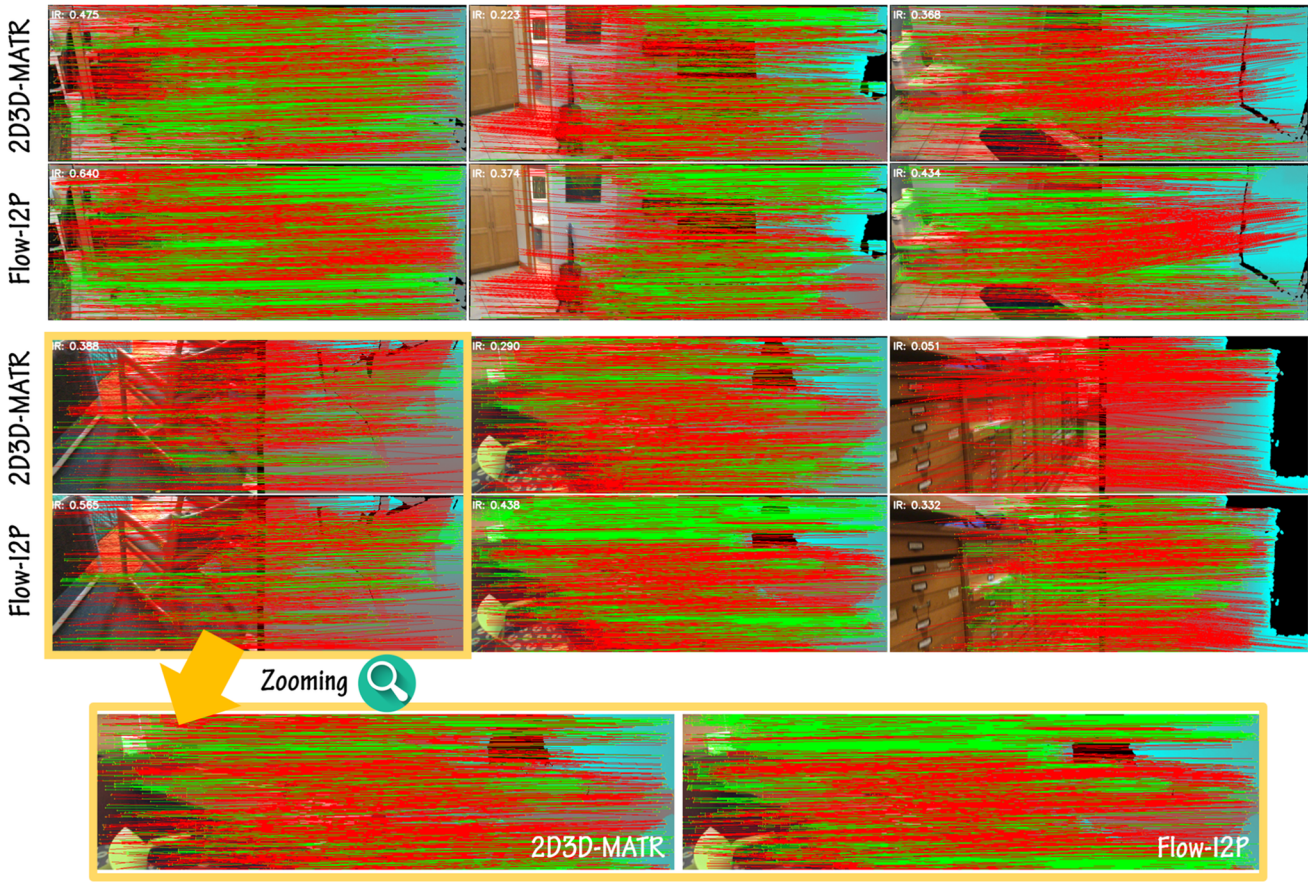

Fig. 10 Qualitative results of Flow-I2P on the ScanNet dataset. Across most scenes, Flow-I2P produces significantly higher-quality 2D-3D correspondences compared to 2D3D-MATR (Li et al., 2023)

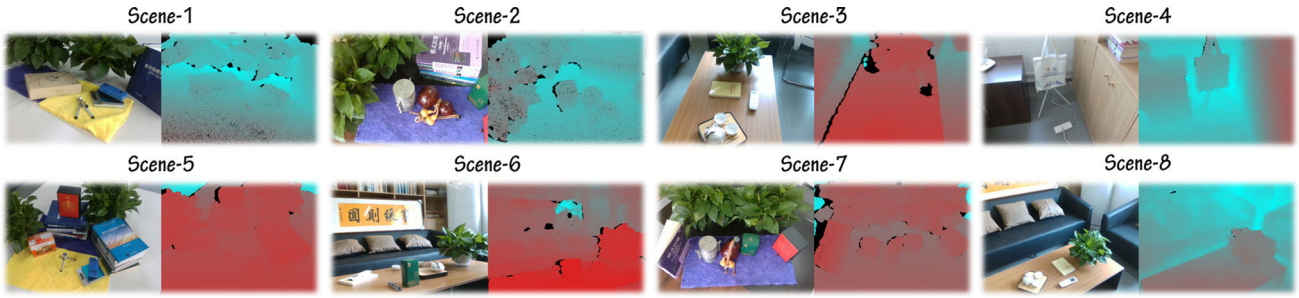

Fig. 11 Visualization of our self-collected dataset across 8 indoor office scenes. RGB images and point clouds were captured using the Intel RealSense D435i. The collected point clouds exhibit typical sensor imperfections, including measurement noise and incomplete geometry due to occlusions and limited depth sensing range

Domain generalization. To further examine the generalization ability of current I2P registration methods, we conducted an additional domain generalization experiment. The previous experiments reveal that **2D3D-MATR+SN** (Li et al., 2023) performs closest to **Flow-I2P** because (i) surface normals help bridge the feature gap between images and point clouds, and (ii) the pretrained surface normal estimator (Bae & Davison, 2024) is robust in open scenes (see Fig. 14). Without surface normals, the original **2D3D-MATR** fails in most unseen scenarios. Therefore, we compared **2D3D-MATR+SN** and **Flow-I2P** in this experiment. The results are reported in Table 8. **A**→**B** indicates that the model is trained on dataset **A** and tested on unseen dataset **B**. We can see that **Flow-I2P** outperforms **2D3D-MATR+SN**, which is attributed to the adaptive noise reduction of the proposed Beltrami flow layer during cross-modality feature interaction. In the self-collected dataset, most scenes involve short distances (i.e., below 1.0m), and the data distribution differs

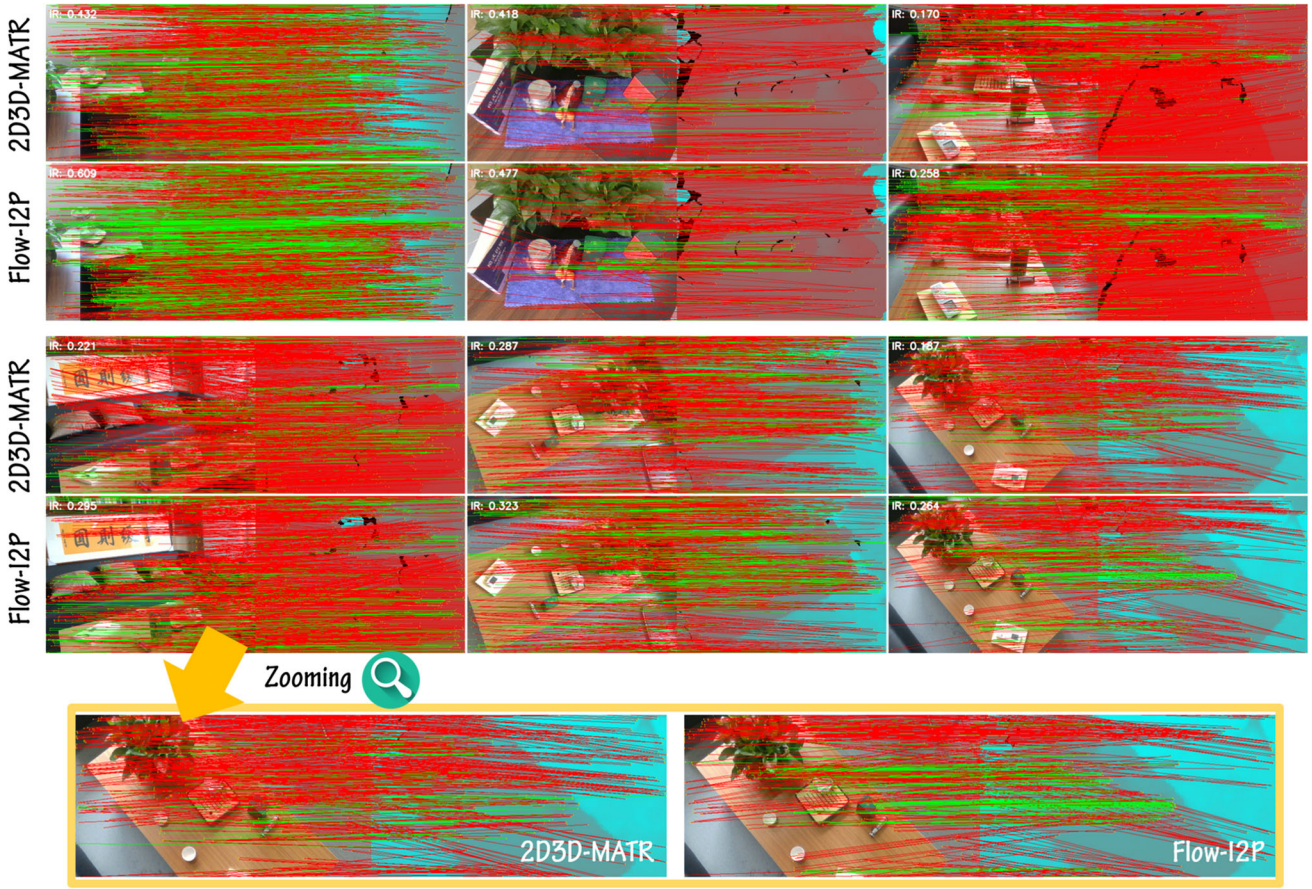

Fig. 12 Qualitative comparison of 2D–3D correspondences on the self-collected dataset. In scenes with closer objects, all methods yield sparse 2D-3D correspondences. In contrast, the distribution of 2D-3D correspondences is much denser in scenes with greater object distances. The proposed Flow-I2P consistently achieves the highest IR, although it has some incorrect correspondences under challenging conditions

Table 7 Noise robustness test on the self-collected dataset. The Bolditalic highlight indicates the best IR and RR

Noise δ_p	0.0cm	1.0cm	2.0cm	3.0cm	4.0cm
Inlier Ratio ↑					
2D3D-MATR	49.3%	48.1%	42.8%	33.7%	25.0%
2D3D-MATR+SN	51.8%	48.7%	43.2%	36.3%	28.2%
Flow-I2P	***53.0%***	***49.2%***	***44.1%***	***36.8%***	***28.4%***
Registration Recall ↑					
2D3D-MATR	91.3%	90.8%	90.2%	75.7%	64.8%
2D3D-MATR+SN	93.5%	93.0%	92.2%	76.4%	66.2%
Flow-I2P	***95.5%***	***95.2%***	***94.1%***	***79.2%***	***66.5%***

Table 8 Average results of learning-based I2P registration models in the context of domain generalization. The Bolditalic highlight indicates the best average for IR. ↑ means that a higher value indicates better performance

Model	PIR ↑	IR ↑	FMR ↑	RR ↑
7-Scenes → RGB-D Scenes V2				
2D3D-MATR+SN	49.1%	17.5%	66.2%	19.8%
Flow-I2P	***54.6%***	***18.3%***	***69.0%***	***23.3%***
(Gain)	+5.5%	+0.8%	+2.8%	+3.5%
7-Scenes → ScanNet				
2D3D-MATR+SN	47.8%	14.3%	70.2%	24.5%
Flow-I2P	***54.9%***	***16.1%***	***70.9%***	***29.0%***
(Gain)	+7.1%	+1.8%	+0.7%	+4.5%
7-Scenes → Self-collected Dataset				
2D3D-MATR+SN	25.2%	9.0%	33.2%	5.4%
Flow-I2P	***27.4%***	***10.8%***	***54.9%***	***7.1%***
(Gain)	+2.2%	+1.8%	+21.7%	+1.7%

from other datasets like 7-scenes and ScanNet, leading to suboptimal performance by the proposed method. Despite this, **Flow-I2P** still achieves higher IR and RR than **2D3D-MATR+SN**.

We further conduct a cross-domain generalization experiment on the 7-scenes dataset (Glocker et al., 2013). In this experiment, the inputs to the I2P registration model are fully overlapped images and point cloud pairs, where the point clouds are generated by back-projecting the depth maps into 3D space. In the **Chess→Other unseen six scenes** set-

Table 9 Cross-domain generalization performance on 7-scenes dataset where the fully overlapped images and colored point clouds are used as the input of the I2P registration model. The Bolditalic highlight indicates the best average for IR. ↑ means that a higher value indicates better performance

Model	Average IR ↑	Average RR ↑
Chess → Other six unseen scenes		
2D3D-MATR+SN	38.7%	47.8%
Flow-I2P	***41.2%***	***58.8%***
(Gain)	+2.5%	+11.0%
Office → Other six unseen scenes		
2D3D-MATR+SN	45.6%	49.1%
Flow-I2P	***46.0%***	***56.3%***
(Gain)	+0.4%	+7.2%
Kitchen → Other six unseen scenes		
2D3D-MATR+SN	53.7%	70.5%
Flow-I2P	***57.2%***	***77.4%***
(Gain)	+3.5%	+6.9%

ting, the model is trained on scene **Chess** and evaluated on the remaining six unseen scenes (i.e., **Fire**, **Heads**, **Office**, **Office**, **Pumpkin**, **Kitchen**, and **Stairs**). The results, shown in Table 9, indicate that **Flow-I2P** significantly outperforms **2D3D-MATR+SN** on the RR metric. This demonstrates the superior generalization capability of the proposed C-flow layers in aligning cross-modality manifolds.

Comparisons with other manifold smoothing methods. Since Beltrami flow is used to smooth the manifold $\mathbb{M}_H$, we compare C-flow with other manifold smoothing methods, including Laplacian eigenmaps (**LE**) (Bastico et al., 2024) and diffusion maps (**DM**) (Pillaud-Vivien & Bach, 2023). However, these methods involve high computational overhead and are not directly applicable to aligning pixel and point features across modalities. As a compromise, we apply them to smooth the extracted features from local 2D and 3D patches. Following the same experiment setup as in ablation studies, we present the comparative results in Table 10. The results demonstrate that the C-flow layers outperform these classical methods, as they can adaptively model correlations between cross-modality features. This confirms that C-flow is not only effective but also a learnable manifold smoothing approach for I2P registration.

These experimental results suggest that the proposed I2P registration method holds potential for domain generalization tasks.

4.3 Ablation Studies

Before comparing the proposed approach to state-of-the-art methods, we perform ablation studies to investigate the contribution of each module in Flow-I2P.

Fig. 13 Performance comparison on the self-collected dataset with varying overlap ratios. Flow-I2P demonstrates consistently superior performance

Fig. 14 Visualization of surface normal estimation results on the ScanNet (top) and self-collected (down) datasets. Despite measurement noise and structural variations, the major geometric structures in each scene are clearly preserved and identifiable

Feature inputs. As explained in Sec. 3.5, the inputs of Flow-I2P involve surface normals. This experiment examines the impact of surface normals on I2P registration. The **Baseline** configuration follows the previous work (Li et al., 2023), but it differs from the original setup, using RGB vectors for images and XYZ positions for point clouds instead of grayscale and full-one vectors. **Baseline+SN** refers to the addition of 2D and 3D surface normals to the original inputs. **Baseline+∇SN** incorporates surface normal gradients, while **Baseline+SN+∇SN** combines both surface normals and gradients with the original inputs. All approaches are trained under the same configuration (e.g., epoch, optimizer, learning rate, augmentation), with the epoch set to 20. To evaluate the performance under limited supervision, only 5% of the 7-Scene training data is used. The results are provided in Table 11. The ablation results reveal two key findings. First, as surface normals in image and point cloud have strong correlation (see Appendix D), models incorporating surface normals and their gradients (**Baseline+SN**, **Baseline+∇SN**, **Baseline+SN+∇SN**) demonstrate substantial improvements over the baseline. This confirms that both surface normals and their gradients contribute positively to learning 2D-3D correspondences. Second, while gradients of surface normals can enhance performance, they may also introduce sensitivity to estimation errors in SN_{2d} and SN_{3d}. This is evident in the observation that **Baseline+SN+∇SN** achieves a lower RR metric compared to **Baseline+SN**. Thus, surface normals are useful in improving the performance of I2P registration (See Fig. 15).

Table 10 I2P registration performance on the baseline using different feature interaction schemes on the 7-Scene validation dataset. PIR and PMR denote patch inlier ratio and patch matching recall, respectively

Methods	PIR	PMR	IR	RR
SCA (Li et al., 2023)	53.8%	62.9%	24.2%	42.8%
LE (Bastico et al., 2024)	55.2%	64.2%	25.1%	43.2%
DM (Pillaud-Vivien & Bach, 2023)	56.0%	64.7%	25.4%	43.5%
C-flow (Ours)	***59.7%***	***69.6%***	***28.2%***	***46.3%***

Table 11 I2P registration results of the baseline network using different feature inputs on the 7-Scene validation dataset. ↑ denotes that a higher value indicates better performance. SN denotes surface normals, and ∇SN represents their gradients

Feature Input	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Mean
Inlier Ratio ↑								
Baseline	39.7%	27.5%	3.0%	20.3%	17.0%	21.2%	5.4%	19.1%
Baseline+SN	42.0%	27.4%	9.0%	23.2%	17.5%	21.2%	5.9%	20.9%
Baseline+∇SN	41.5%	25.9%	10.1%	22.4%	17.1%	20.5%	6.2%	20.5%
Baseline+SN+∇SN	***41.9%***	***29.9%***	***7.1%***	***24.1%***	***18.4%***	***23.2%***	***3.6%***	***21.2%*** (↑***2.1%***)
Feature Matching Recall ↑								
Baseline	96.9%	93.2%	0.0%	89.9%	76.6%	86.3%	21.4%	66.3%
Baseline+SN	96.9%	89.0%	33.3%	86.0%	80.9%	88.1%	14.3%	69.9%
Baseline+∇SN	***98.5%***	***87.7%***	***50.0%***	***84.8%***	***78.7%***	***88.1%***	***21.4%***	***72.7%*** (↑***6.4%***)
Baseline+SN+∇SN	100.0%	90.4%	33.3%	89.9%	80.9%	89.3%	21.4%	72.2%
Registration Recall ↑								
Baseline	67.7%	27.4%	0.0%	52.5%	25.5%	54.8%	13.3%	34.6%
Baseline+SN	***70.8%***	***43.8%***	***8.3%***	***52.5%***	***34.0%***	***51.8%***	***14.3%***	***39.4%*** (↑***4.8%***)
Baseline+∇SN	72.3%	37.0%	8.3%	40.4%	38.3%	51.8%	0.0%	35.4%
Baseline+SN+∇SN	66.2%	43.8%	0.0%	55.6%	34.0%	64.9%	7.1%	38.8%

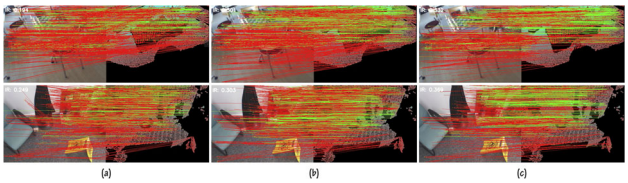

Fig. 15 Visualization of the ablation study on different feature interaction schemes. (a) Stacked histogram attention. (b) Stacked C-flow without the CSA layer. (c) Stacked C-flow with CSA. The proposed C-flow, particularly with CSA, significantly improves the quality of inlier correspondences under challenging registration conditions

Feature interaction schemes. As illustrated in Sec. 2.2, feature interaction plays a crucial role in I2P registration. This experiment examines the effectiveness of the proposed feature interaction schemes. B-flow demands significant memory and cannot be trained on a single GPU. Given these GPU limitations, we evaluate feature interaction accuracy using I2P patch registration for a fair comparison with the previous work (Li et al., 2023). **SCA** (Li et al., 2023) refers to the use of a series of standard transformer-based self- and cross-attention for LFI. **B-flow** and **C-flow** are the proposed LFI methods, as detailed in Sec. 3.3 and 3.4. **Deep-I2P** is a representative GFI method (Li & Lee, 2021), discussed in Sec. 2.2. **None** serves as a reference, and does not use any feature interaction modules. All methods are implemented based on **Baseline+SN**. The results are reported in Table 12. Patch inlier ratio (PIR) and patch matching ratio (PMR) are used as metrics, with a PMR threshold of 0.5. First, results in Table 12 verify that (i) feature interaction improves performance (i.e., all approaches are superior to **None**); (ii) LFI is better than GFI (**SCA**, **B-flow**, and **C-flow** outperform **Deep-I2P**), thanks to its support for fine-grained pixel-level feature interaction. Second, **B-flow** is more accurate than **SCA** (Li et al., 2023), because **B-flow** is derived from Beltrami flow and enabled to adaptively represent the correlation between the features of 2D pixels and 3D points. So, it enhances the alignment of manifolds $\mathbb{M}_I$ and $\mathbb{M}_P$, and leads to the superior results compared to **SCA**. Third, **C-flow** performs nearly as well as **B-flow**, indicating that dimension-wise attention can achieve a similar result to spatial-wise attention in the context of I2P registration. The major advantage of **C-flow** is its reduced memory consumption compared to **B-flow**, making it more feasible for pixel-level cross-modality feature representation in I2P registration architecture.

The above experiments were conducted on I2P patch registration, a specific case of I2P registration. To further explore the effect of C-flow in general I2P registration, we included

Table 12 I2P registration results of the baseline using different feature interaction schemes on the 7-Scene validation dataset. ↑ denotes that a higher value indicates better performance

Feature Interaction Scheme	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Mean
Patch Inlier Ratio ↑								
None	63.9%	59.1%	26.0%	51.5%	53.2%	57.3%	11.6%	46.1%
Deep-I2P (Li & Lee, 2021)	67.3%	63.9%	29.2%	56.1%	57.3%	59.6%	11.9%	49.3%
SCA (Li et al., 2023)	74.8%	65.4%	34.5%	60.6%	59.0%	65.3%	17.1%	53.8%
B-flow (Ours)	***81.3%***	***71.8%***	***33.9%***	***69.2%***	***68.7%***	***73.8%***	***21.0%***	***59.9%*** (↑***13.8%***)
C-flow (Ours)	80.3%	71.7%	34.2%	69.2%	68.2%	73.5%	20.5%	59.7%
Patch Matching Ratio ↑								
None	89.2%	75.3%	16.7%	64.6%	66.0%	78.0%	0.0%	55.7%
Deep-I2P (Li & Lee, 2021)	93.8%	82.2%	0.0%	72.7%	72.3%	75.0%	0.0%	56.6%
SCA (Li et al., 2023)	95.4%	87.7%	16.7%	81.8%	72.3%	86.3%	0.0%	62.9%
B-flow (Ours)	98.5%	89.0%	16.7%	89.9%	91.5%	92.3%	0.0%	68.3%
C-flow (Ours)	***98.5%***	***90.4%***	***25.0%***	***90.9%***	***89.4%***	***92.9%***	***0.0%***	***69.6%*** (↑***13.9%***)

an additional comparison. **Stacked GFI** refers to using GFI modules (Li & Lee, 2021) for feature interaction. **Stacked kernel distribution estimation (KDE) attention** replaces the covariance matrix with KDE-based attention. More detail is shown in Appendix C. **Stacked Linformer** (Wang et al., 2020), as a lightweight variant of the standard transformer layers, is also compared in this experiment. **Stacked C-flow** employs a series of proposed C-flow layers. **w/o CSA** indicates the C-flow configuration without the CSA layer in Eq. (12). The stack number for all methods was set to 3. The results are shown in Table 13. First, according to the FMR definition (Li et al., 2023), it is not a strict metric for assessing the quality of 2D-3D correspondence. Thus, the methods show similar FMR values. Then, both **stacked KDE attention** and **stacked C-flow**, being LFI-based methods, outperform GFI-based methods. This result is consistent with the findings in Table 12. **stacked C-flow** outperforms **stacked KDE attention** in both IR and RR metrics, as KDE is only an approximation of feature distribution. The resolution of KDE limits the performance of **stacked KDE attention**. If set too high, it demands excessive memory for training. Thus, KDE-based attention is less efficient than covariance-based attention in Eqs. (12) and (13). While **stacked C-flow** and **stacked C-flow (w/o CSA)** exhibit similar IR, **stacked C-flow** achieves the superior performance in RR. Compared to IR, RR is more closely tied to RANSAC-based PnP accuracy, which depends on the alignment quality of 2D-3D correspondences. This result reveals that the CSA layer enhances correspondence learning quality. Furthermore, we observe that **stacked C-flow** layers perform better when τ is set to a fixed value. Learning τ dynamically introduces instability in training, as it interferes with the optimization of latent feature $\mathbf{z}[t]$ and the attention function $\mathbf{A}(\cdot)$. When τ approaches zero, the gradient $\partial L_{\text{I2P}}/\partial \mathbf{A}$ tends to vanish, which weakens the impact of C-flow on I2P registration. Conversely, if τ approaches one, the contribution of $\mathbf{z}[t-1]$ becomes negligible, causing the model to forget historical features and impairing correspondence learning. Thus, a properly selected fixed τ is recommended in practice. We also include **stacked Linformer** as a transformer-based baseline that learns a low-rank attention matrix. While computationally efficient for pixel-level interactions, Linformer does not incorporate cross-model correlation priors, leading to inferior performance in I2P registration compared to the proposed **stacked C-flow**.

In summary, results from Tables 12 and 13 confirm that the proposed **C-flow** outperforms current GFI and LFI modules. Rooted in Beltrami-flow theory, C-flow effectively captures the correlations between different modality feature manifolds. As a result, **C-flow** provides more semantically meaningful interactions than traditional attention layers, especially in multi-modal scenarios. Also, **C-flow** only has the complexity much less than standard transformer-based attention schemes, which enables an effective and lightweight cross-modality feature interaction on I2P registration.

Hyper-parameters in C-flow. The above experiments have validated the effectiveness of the proposed C-flow layer. To optimize performance, we examined the hyperparameters in C-flow. We leveraged a two-stage training scheme, where the first and the second stages require 20 and 10 epochs for training, respectively.

First, the time step τ in Eq. (7) is a crucial parameter, as it determines the weighting of the feature interaction result. To determine the optimal τ, we uniformly sampled the interval [0, 0.5] with a step size of 0.1 and recorded the performance in Table 14, where the stack number was set to 3. We observed that the training procedure fails due to vanishing gradient when τ exceeds 0.5. A large τ causes the neural network to

Table 13 I2P registration results of the baseline network using different stacked feature interaction schemes on the 7-Scene validation dataset. ↑ denotes that a higher value indicates better performance. w/o CSA is C-flow layer without CSA module. w/ τ learning is C-flow layer where τ in each layer is learnable

Stacked Feature Interaction Scheme	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Mean
Inlier Ratio ↑								
Stacked GFI	52.7%	37.5%	9.2%	31.6%	24.8%	29.9%	7.0%	27.5%
Stacked KDE Attention	53.8%	38.1%	10.9%	31.6%	23.1%	29.4%	7.0%	27.7%
Stacked Linformer	52.4%	37.5%	11.5%	28.7%	26.2%	30.1%	6.8%	27.6%
Stacked C-flow (Ours, w/o CSA)	52.1%	38.0%	12.3%	31.3%	25.9%	29.3%	7.9%	28.1%
Stacked C-flow (Ours, w/ τ learning)	52.2%	37.8%	11.2%	31.0%	24.7%	29.0%	7.2%	27.5%
Stacked C-flow (Ours)	***52.5%***	***37.9%***	***12.5%***	***31.3%***	***26.0%***	***29.2%***	***7.7%***	***28.2%***
Feature Matching Recall ↑								
Stacked GFI	***100.0%***	***93.2%***	***50.0%***	***97.0%***	***87.2%***	***95.2%***	***28.6%***	***78.7%***
Stacked KDE Attention	100.0%	93.2%	41.7%	96.0%	83.0%	91.7%	28.6%	76.3%
Stacked Linformer	100.0%	93.2%	45.8%	92.2%	93.0%	90.3%	30.6%	77.9%
Stacked C-flow (Ours, w/o CSA)	100.0%	93.2%	50.0%	94.9%	91.5%	91.7%	28.6%	78.5%
Stacked C-flow (Ours, w/ τ learning)	100.0%	93.2%	42.3%	91.7%	89.8%	92.3%	27.3%	76.7%
Stacked C-flow (Ours)	100.0%	93.2%	41.7%	94.9%	91.5%	91.7%	35.7%	78.4%
Registration Recall ↑								
Stacked GFI	80.0%	56.2%	0.0%	61.6%	40.4%	64.3%	0.0%	43.2%
Stacked KDE Attention	81.5%	47.9%	0.0%	64.6%	46.8%	66.7%	7.1%	45.0%
Stacked Linformer	77.2%	45.2%	0.0%	63.8%	42.3%	66.8%	4.2%	42.8%
Stacked C-flow (Ours, w/o CSA)	76.9%	46.6%	0.0%	62.6%	38.3%	67.3%	7.1%	42.7%
Stacked C-flow (Ours, w/ τ learning)	78.3%	40.2%	0.0%	61.2%	40.0%	66.2%	7.1%	41.8%
Stacked C-flow (Ours)	***89.2%***	***41.1%***	***0.0%***	***64.6%***	***51.1%***	***70.8%***	***7.1%***	***46.3%***

Table 14 Ablation study of the time step parameter τ in C-flow layers on the 7-Scene validation dataset. The registration recall for all scenes is recorded. The Bolditalic highlight indicates the best RR

Time step τ	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Mean
0.0	86.2%	39.7%	0.0%	64.6%	29.8%	72.0%	7.1%	42.8%
0.1	78.5%	42.5%	0.0%	65.7%	44.7%	66.1%	14.3%	44.5%
0.2	***89.2%***	***41.1%***	***0.0%***	***64.6%***	***51.1%***	***70.8%***	***7.1%***	***46.3%*** (↑***3.5%***)
0.3	81.5%	43.8%	0.0%	60.6%	46.8%	69.6%	7.1%	44.2%
0.4	83.1%	41.1%	8.3%	59.6%	51.1%	64.3%	7.1%	44.9%
0.5	76.9%	46.6%	0.0%	63.6%	48.9%	63.1%	0.0%	42.7%

forget a significant portion of features in the last iteration (i.e., $\mathbf{x}[t-1]$ and $\mathbf{y}[t-1]$), increasing the risk of vanishing gradient. Conversely, a small τ reduces the effect of feature interaction, hindering the discriminative ability of cross-modality features. Thus, there must be an optimal τ for the C-flow layer. As shown in Table 14, the best RR was achieved when $\tau = 0.2$.

Second, the stack number T of the C-flow layer is closely related to I2P registration performance. To find the optimal T, we tested the value of T from 0 to 5 with $\tau = 0.2$. The registration results are presented in Table 15. As T increases, all metrics improve up to $T = 3$, after which they begin to decline. The weight of the original feature from $\mathbf{F}(\cdot)$ and $\mathbf{G}(\cdot)$ is $(1-\tau)^T$. As T increases, the contribution of $\mathbf{x}[0]$ and $\mathbf{y}[0]$ to $\mathbf{x}[T]$ and $\mathbf{y}[T]$ diminishes. This increases the risk of degradation in the manifold structures $\mathbb{M}_I$ and $\mathbb{M}_P$. Thus, we select $T = 3$ for optimal I2P registration performance.

These experiments indicate that the performance of I2P registration declines, or the training may fail if τ or T is too large. We attempt to explain this in-depth from the perspective of manifold alignment. As shown in Fig. 3, manifold alignment can be interpreted as a manifold smoothing problem. Both τ and T control the *strength* of manifold smoothing. If either τ or T is too large, the manifold becomes **over-smoothing**, causing a sharp drop in the discriminative ability of cross-modality features. Specifically, if τ exceeds 0.5, the manifold structure may degrade, resulting in vanishing gradient. Therefore, τ and T should be carefully selected in practical applications.

Table 15 Ablation study of the stack number T in the proposed C-flow layer on the 7-Scene validation dataset. Mean metrics for all scenes are recorded

T	IR	FMR	RR
0	26.5%	75.9%	42.8%
1	27.6%	76.2%	44.7%
2	28.1%	**77.4%**	46.0%
3	***28.2%*** (↑***1.7%***)	***76.3%*** (↑***0.4%***)	***46.3%*** (↑***3.5%***)
4	26.2%	74.8%	45.8%
5	25.4%	72.8%	41.6%

Table 16 Ablation study of different training schemes on the 7-Scene validation dataset. Mean metrics for all scenes are recorded

Scheme	IR	FMR	RR
One-Stage	27.3%	75.1%	41.9%
Two-Stage	***28.2%*** (↑***0.9%***)	***75.9%*** (↑***0.8%***)	***46.3%*** (↑***4.4%***)

Table 17 Model efficiency analysis in several aspects, like GFLOPs, average inference memory, parameter size, and FPS

Models	GFLOPs	Memory (GB)	Param (M)	FPS
Diff-Reg	1165.3	5.637	358.2	0.25
Baseline	388.1	2.732	28.2	7.874
+B-flow	N/A	Out of memory	28.4	N/A
+C-flow×1	433.8	2.769	28.4	6.289
+C-flow×2	479.4	2.806	28.5	6.173
+C-flow×3	***525.1***	***2.844***	***28.7***	***6.061***

Training scheme. We evaluated the performance of the training scheme outlined in Sec. 3.5. The **One-stage** approach involves training the entire neural network from scratch for 30 epochs. The **Two-stage** approach follows the proposed scheme, with 20 epochs in the first stage and 10 epochs in the second. Results are provided in Table 16. Compared to the **One-stage** approach, the **Two-stage** training significantly improves the RR metric. This improvement arises from the pre-training of the feature extractors and increases the alignment accuracy of $\mathcal{M}_I[0]$ and $\mathcal{M}_P[0]$. As a result, the **Two-stage** setup better satisfies the manifold alignment prior condition described in Sec. 3.2, thereby improving the training efficiency and effectiveness.

Model efficiency analysis. We evaluate the efficiency of the proposed model in terms of gigabyte floating point operations per second (GFLOPs), average inference memory, parameter size, and runtime (frames per second, FPS). The results are summarized in Table 17. The original B-flow suffers from computational overhead due to large matrix multiplications. As shown in Eq. (10), the self-attention operation on image features requires constructing a feature matrix of size $HW \times HW$. For a 640 × 480 image, this leads to a matrix with approximately 9.4×10^{10} elements, making it infeasible to store or compute on commodity GPUs such as NVidia GTX 3080. To address this issue, C-flow is proposed, which requires only a feature matrix of size $c \times c$, with $c = 256$ in our experiment. This reduces the computation load to approximately 10^{-6} of that of B-flow. Since $c < 10^3$, the stacked C-flow layers significantly reduce GFLOPs, memory consumption, and inference latency, making them lightweight yet effective for I2P registration.

Moreover, we assess the practical deployment of Flow-I2P on edge devices. As shown in Table 17, Flow-I2P consumes approximately 0.48 TFLOPs and less than 3 GB of GPU memory during inference. These requirements are well within the capabilities of modern low-power GPUs, such as the NVIDIA Jetson Orin Nano (40 TFLOPs peak, 8 GB GPU memory, 7-25W power consumption). This makes Flow-I2P suitable for edge computing scenarios on mobile and embedded robotic platforms. In addition, we compare the runtime efficiency of Flow-I2P with state-of-the-art methods in Fig. 2. Flow-I2P strikes a favorable balance between registration accuracy and runtime, achieving a runtime of nearly 6 FPS. In real-world robotic applications, I2P registration is typically used when the robot loses localization within a 3D map or needs to establish accurate 2D-3D loop-closure constraints during visual SLAM. Thus, the runtime efficiency of Flow-I2P is sufficient for the majority of robot-centric scenarios.

Finally, we discuss model efficiency in large-scale environments, as encountered in practical deployments (Kim et al., 2023). Large-scale I2P registration entails matching images to extensive point cloud maps. However, in real-world robotics and autonomous navigation settings (An et al., 2024a; Kim et al., 2023), the I2P registration task is commonly reduced to small-scale local registration tasks using pose priors (e.g., from other sensors such as GNSS or IMU) (An et al., 2024b; Yuan et al., 2021). These priors enable coarse localization, allowing a relevant subset of the 3D map to be segmented for registration. This localized context dramatically reduces the search space, enabling Flow-I2P to operate efficiently even in large-scale environments.

4.4 Limitations and Future Works

The proposed Flow-I2P demonstrates superior performance in both model comparison and cross-domain generalization across public datasets that encompass a variety of indoor and outdoor scenes. These results indicate that the proposed Beltrami flow is an effective approach for enhancing I2P registration. However, despite its promising results, Flow-I2P still has room for improvement before it can be fully deployed in practical applications.

Lack 2D-3D correspondences pruning. While Flow-I2P achieves strong overall performance, it may produce noisy correspondences in sparse or out-of-distribution scenarios. To mitigate this, we plan to integrate a dedicated pruning module inspired by 3D-3D correspondence pruning (Cheng et al., 2023) to remove unreliable matches in future work.

Accurate I2P registration in real-time. Our current implementation runs at nearly 6 FPS, which may be insufficient for latency-sensitive applications such as AR/VR or high-speed robotics. We outline the future directions to develop a lightweight variant to improve runtime while maintaining accuracy.

Generalization to unseen domain. As shown in Tables 4 and 8, generalization across domains is still challenging. We plan to explore the development of a visual foundation model tailored for I2P registration, capable of generalizing across different domains and sensor modalities.

Extend C-flow to general registration. Theoretically, the proposed C-flow can be extended to aligning features manifolds generated from the single and multiple modality data. It suggests that C-flow can be embedded into deep neural networks for broader registration tasks, including image registration, point cloud registration, and cross-modal registration.

5 Conclusion

This paper addresses improving the performance of I2P registration from the perspective of information geometry. We began by analyzing how Beltrami flow improves the capacity of I2P registration by correcting the misalignment between image and point cloud feature manifolds. Building on this insight, we proposed B-flow, Beltrami flow based feature interaction layers, to progressively align feature manifolds. To mitigate its high computational cost, we introduced C-flow, a lightweight and efficient variant of B-flow. We then developed the Flow-I2P registration network to fully leverage the advantages of C-flow in I2P registration tasks. Extensive experiments across five indoor and outdoor datasets demonstrate that Flow-I2P significantly outperforms existing methods, especially under limited training data. These results validate that incorporating Beltrami flow effectively enhances the generalization ability of I2P registration.

Appendix

A. Interpretation of manifold alignments in I2P registration

We present an interpretation of I2P registration from the perspective of manifold theory. Let $\mathbb{M}_I$ and $\mathbb{M}_P$ be differentiable manifolds in a c-dimensional Euclidean space, containing $\{\mathbf{x}_i\}_{i=1}^{M}$ (M=HW) and $\{\mathbf{y}_j\}_{j=1}^{N}$, respectively. To reveal the geometrical relation between $\mathbb{M}_I$ and $\mathbb{M}_P$ under the ideal correspondence condition stated in Eq. (2), we introduce the following:

Lemma 1 *Let image $\mathcal{I}$ be a bounded continuous surface on the 2D region $[0,1]\times[0,1]$, and point cloud $\mathcal{P}$ be a bounded continuous surface in the 3D space. If $\mathbf{F}(\cdot)$ and $\mathbf{G}(\cdot)$ are differential functions that satisfy the ideal correspondence condition for all correspondences, then for any ideal correspondence $\langle I_i, P_j\rangle$, there exist open sets $\mathcal{O}_I(I_i)$ on I_i and $\mathcal{O}_P(P_j)$ on P_j such that*

$$\forall P \in \mathcal{O}_P(P_j),\ \exists! I \in \mathcal{O}_I(I_i), \mathbf{F}(I|\mathcal{I}) = \mathbf{G}(P|\mathcal{P}). \tag{18}$$

Proof We can select an open set $\mathcal{O}_P(P_j)$ on P_j that lies within the field of view (FoV) of the camera without occlusion. For $P \in \mathcal{O}_P(P_j)$, $I = \pi(P)$, where $\pi(\cdot)$ is the projection operator onto 3D space to image plane, as presented in Eq. (1). Since I is uniquely determined by P, $\langle I, P\rangle$ forms a correspondence. Given that $\mathbf{F}(\cdot)$ and $\mathbf{G}(\cdot)$ satisfy the ideal correspondence condition for all correspondences, we have $\mathbf{F}(I|\mathcal{I}) = \mathbf{G}(P|\mathcal{P})$. Thus, this proof is completed.

Furthermore, since $\mathbf{F}(\cdot)$ and $\mathbf{G}(\cdot)$ are differential functions, the mapping results $\mathbf{F}(\mathcal{O}_I(I_i))$ and $\mathbf{G}(\mathcal{O}_P(P_j))$ are also open sets (i.e., sub-manifolds of $\mathbb{M}_I$ and $\mathbb{M}_P$). Based on **Lemma 1**, $\mathbf{F}(\mathcal{O}_I(I_i))$ and $\mathbf{G}(\mathcal{O}_P(P_j))$ coincide, and this geometrical relationship can be generalized.

Theorem 1 ***(Manifold structure in ideal correspondence).*** *Let image $\mathcal{I}$ be a bounded continuous surface on the region $[0,1]\times[0,1]$; point cloud $\mathcal{P}$ be a bounded continuous surface in the 3D space. If $\mathbf{F}(\cdot)$ and $\mathbf{G}(\cdot)$ are differential functions that satisfy the ideal correspondence condition for all correspondences, there exist sets $\{\mathcal{O}_I(I_i)\}_{i=1}^{M}$ and $\{\mathcal{O}_P(P_j)\}_{j=1}^{N}$, such that $\mathbb{M}_I \cap \mathbb{M}_P$ can be decomposed as:*

$$\mathbb{M}_I \cap \mathbb{M}_P = \bigcup_{\langle I_i, P_j\rangle \in \mathcal{C}} \mathbf{F}(\mathcal{O}_I(I_i)) \cap \mathbf{G}(\mathcal{O}_P(P_j)) \tag{19}$$

where $\mathcal{C}$ represents the set of all correspondences.

Proof From **Lemma 1**, we directly obtain:

$$\mathbb{M}_I \cap \mathbb{M}_P \supseteq \bigcup_{\langle I_i, P_j\rangle \in \mathcal{C}} \mathbf{F}(\mathcal{O}_I(I_i)) \cap \mathbf{G}(\mathcal{O}_P(P_j)) \tag{20}$$

Conversely, for $\mathbf{x} = \mathbf{y} \in \mathbb{M}_I \cap \mathbb{M}_P$, we can identify the correspondence $\langle I, P\rangle$ related to $\mathbf{x}$ and $\mathbf{y}$, as stated in Eq. (2). Thus, we can define local regions $\mathcal{O}_I(I)$ and $\mathcal{O}_P(P)$ such that $\mathbf{x} = \mathbf{y} \in \mathbf{F}(\mathcal{O}_I(I)) \cap \mathbf{G}(\mathcal{O}_P(P))$. This implies that

$$\mathbb{M}_I \cap \mathbb{M}_P \subseteq \bigcup_{\langle I_i, P_j\rangle \in \mathcal{C}} \mathbf{F}(\mathcal{O}_I(I_i)) \cap \mathbf{G}(\mathcal{O}_P(P_j)) \tag{21}$$

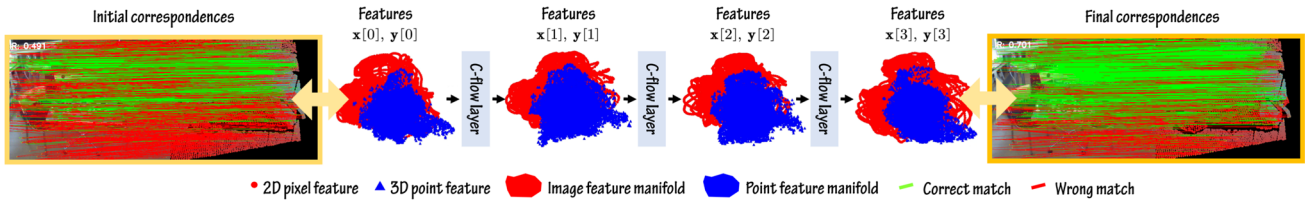

Fig. 16 Visualization of manifold alignment after applying stacked C-flow layers. As the number of iterations increases, the accuracy of manifold alignment improves, leading to higher-quality 2D-3D correspondences

Combing Eqs. (20) and (21), the proof is complete. Furthermore, as discussed in Lemma 1, $\mathbf{F}(\mathcal{O}_I(I_i)) \cap \mathbf{G}(\mathcal{O}_P(P_j) = \mathbf{F}(\mathcal{O}_I(I_i)) = \mathbf{G}(\mathcal{O}_P(P_j)$ (as discussed in **Lemma 1**). **Theorem 1** indicates that $\mathbb{M}_I \cap \mathbb{M}_P$ encapsulates all the 2D-3D correspondences.

B. Explanation of Beltrami flow based I2P registration

The proposed Beltrami flow based I2P registration model in Fig. 4 can be reformulated as follows:

Definition 2 (Neural diffusion based I2P registration model). The discrete neural diffusion based I2P registration model can be expressed as the following optimization problem:

$$\min_{\Theta} \sum_{t=0}^{T} L_{\text{I2P}}(\mathbf{x}[t], \mathbf{y}[t] | \mathcal{C}_{\text{GT}}) \tag{22}$$

$$\begin{cases} \mathbf{x}[t+1] = (1-\tau)\mathbf{x}[t] + \tau(\text{SA}_I(\mathbf{x}[t]) + \text{CA}_I(\mathbf{x}[t], \mathbf{y}[t])) \\ \mathbf{y}[t+1] = (1-\tau)\mathbf{y}[t] + \tau(\text{SA}_P(\mathbf{y}[t]) + \text{CA}_P(\mathbf{y}[t], \mathbf{x}[t])) \end{cases} \tag{23}$$

$$\begin{aligned} &t = 0, ..., T-1 \\ &\mathbf{x}[0] = (\mathbf{x}_1^T, ..., \mathbf{x}_M^T)^T = (\mathbf{F}(I_1|\mathcal{I})^T, ..., \mathbf{F}(I_M|\mathcal{I})^T)^T \\ &\mathbf{y}[0] = (\mathbf{y}_1^T, ..., \mathbf{y}_N^T)^T = (\mathbf{G}(P_1|\mathcal{P})^T, ..., \mathbf{G}(P_N|\mathcal{P})^T)^T \end{aligned} \tag{24}$$

where Θ represents the parameters of the learnable functions $\mathbf{F}(\cdot)$, $\mathbf{G}(\cdot)$, $\text{SA}_I(\cdot)$, $\text{SA}_P(\cdot)$, $\text{CA}_I(\cdot)$, and $\text{CA}_P(\cdot)$. L_{I2P} is the regular correspondence loss (Li et al., 2023), and $\mathcal{C}_{\text{GT}}$ denotes the ground truth (GT) correspondences.

We interpret the model in Eqs. (22-24) through the viewpoint of optimal control theory. Eq. (23) can be reformulated as a time-continuous system:

$$\min_{\mathbf{u}_x(t), \mathbf{u}_y(t)} J = \int_0^1 L_{\text{I2P}}(\mathbf{x}(t), \mathbf{y}(t) | \mathcal{C}_{\text{GT}}) \tag{25}$$

$$\begin{cases} \dfrac{\partial \mathbf{x}(t)}{\partial t} = (\text{SA}_I(\mathbf{x}(t)) - \mathbf{I})\mathbf{x}(t) + \mathbf{u}_x(t) \\ \dfrac{\partial \mathbf{y}(t)}{\partial t} = (\text{SA}_P(\mathbf{y}(t)) - \mathbf{I})\mathbf{y}(t) + \mathbf{u}_y(t) \end{cases} \tag{26}$$

Here, Eq. (23) can be interpreted as a differential system involving $\mathbf{x}(t)$ and $\mathbf{y}(t)$. When $\mathbf{u}_x(t) = 0$ and $\mathbf{u}_y(t) = 0$, this system breaks down into two independent diffusion subsystems that smooth the manifolds of $\mathbf{x}(t)$ and $\mathbf{y}(t)$, respectively. However, these independent diffusions are inefficient at minimizing L_{I2P}, because they do not share information between $\mathbf{x}(t)$ and $\mathbf{y}(t)$. For optimal control, the external forces in this system (i.e., $\mathbf{x}(t)$ and $\mathbf{y}(t)$) must include the similarity information between $\mathbf{x}(t)$ and $\mathbf{y}(t)$. In the neural diffusion based I2P registration model, this similarity information is captured using learnable cross-attention layers. The control variables are defined as:

$$\mathbf{u}_x(t) = \text{CA}_I(\mathbf{x}[t], \mathbf{y}[t])), \quad \mathbf{u}_y(t) = \text{CA}_P(\mathbf{y}[t], \mathbf{x}[t])) \tag{27}$$

The parameters in $\text{CA}_I(\cdot)$ and $\text{CA}_P(\cdot)$ are updated through backward propagation by minimizing J in Eq. (22).

We further analyze the role of τ in Eq. (7), which is a time step to describe the discrete Beltrami flow in Eqs. (4)-(8). First, τ is a unified time step for image and point cloud features, because the Beltrami flow formulation treats both modalities as part of a unified manifold. This design ensures consistent flow dynamics across both data types. Second, it is not recommended to set τ as a learnable parameter, because learning τ poses the risk of inhibiting the learning efficiency of latent feature $\mathbf{z}[t]$ or attention function $\mathbf{A}$. Third, the stacked layers denote the continuous forward propagation of discrete differential equations (i.e., in a manner of explicit Euler method). Allowing τ to vary across layers would compromise the consistency of the discretization and undermine the interpretation of the network as a numerical solver for a continuous system. Thus, τ should be a constant parameter. In future work, we will explore a general Beltrami flow with the dynamic τ.

Besides, we further provide an in-depth empirical justification of the proposed Beltrami flow on I2P registration. As discussed in Fig. 1, the motivation of Beltrami flow is to enhance the alignment accuracy of manifolds generated from pixel and point features. It is recalled that the proposed Beltrami flow layers are embedded into the I2P registration architecture in a stacked manner. We first visualize the process of manifold alignment with the stacked C-flow layers, and the results are provided in Fig. 16. As pixel and point features are in a 256-dimensional space, we project these features into a 2D space for visualization using t-SNE. This

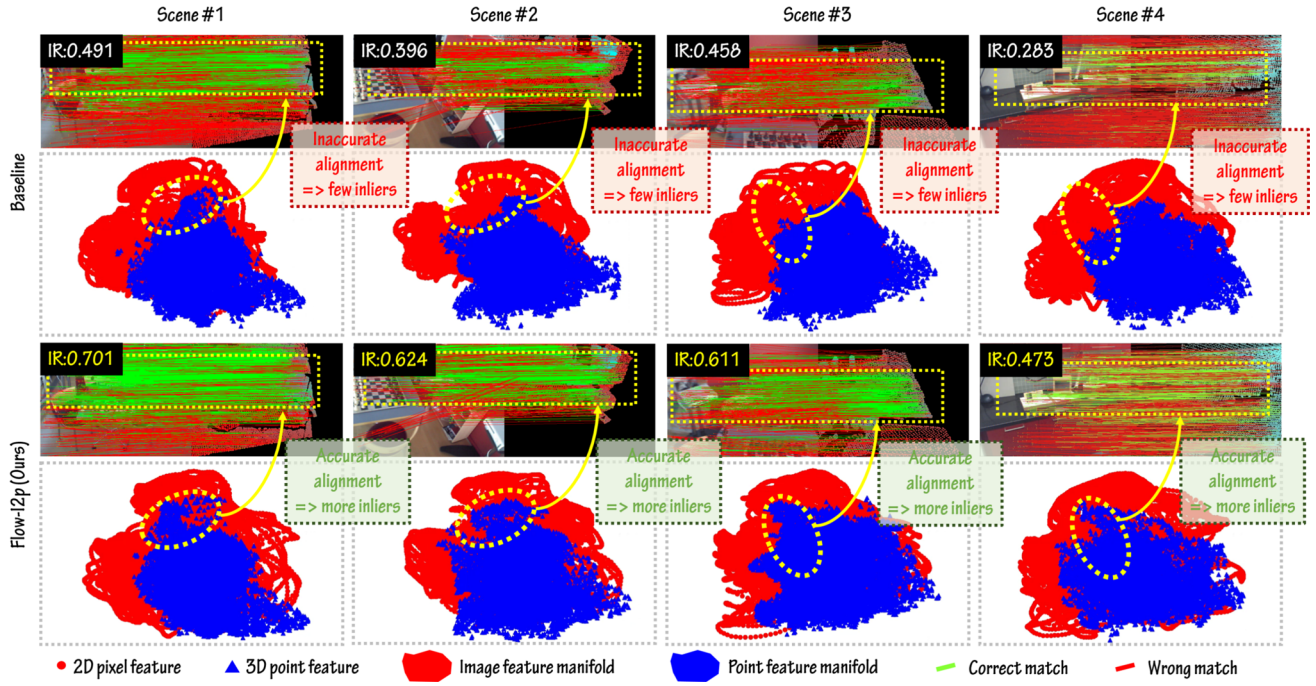

Fig. 17 Relationship between manifold alignment and I2P registration. Manifolds of pixel features and point features are visualized in 2D space using the technique of t-SNE. An accurate manifold alignment indicates the more correct 2D-3D correspondences

visualization shows that C-flow layers enhance the quality of manifold alignment in each iteration, significantly improving the quality of 2D-3D correspondences. We then visualize the relationship between the manifold alignment and I2P registration in Fig. 17. Comparing the manifold alignment of 2D3D-MATR (baseline) (Li et al., 2023) and the proposed Flow-I2P, it can be observed that Beltrami flow improves the alignment quality and enlarges the inliers in 2D-3D correspondences, demonstrating the effectiveness of the proposed Beltrami flow on I2P registration.

At the end of this appendix, we discuss the issue of Eqs. (12) and (13) with explaining why the following equation satisfies:

$$\begin{aligned}&\text{Softmax}\left(\frac{(\mathbf{x}[t]\mathbf{w}_x)(\mathbf{x}[t]\mathbf{w}_x)^T}{\sqrt{c}}\right)\cdot\mathbf{x}[t]\\&\quad=\mathbf{x}[t]\cdot\text{Softmax}\left(\frac{(\mathbf{x}[t]\mathbf{v}_x)^T(\mathbf{x}[t]\mathbf{v}_x)}{\sqrt{c}}\right)\end{aligned}\tag{28}$$

Eq. (28) can be seen as an instance of the matrix equation $\mathbf{AX}=\mathbf{XB}$, where $\text{rank}(\mathbf{A})\le c$ and $\text{rank}(\mathbf{B})\le c$. From the theory of linear equations, $\mathbf{X}$ is solvable if $\mathbf{A}$ and $\mathbf{B}$ have the same eigenvalues. As $\text{rank}(\mathbf{A})$ and $\text{rank}(\mathbf{B})$ are below c, it is possible to learn $\mathbf{w}_x$, $\mathbf{w}_y$, $\mathbf{v}_x$, and $\mathbf{v}_y$ to make sure $\mathbf{X}$ solvable. Hence, it is feasible to convert spatial-wise attention to channel-wise attention like Eq. (28).

C. Implementation details of the KDE attention layer

KDE attention is one of the lightweight variants of the proposed B-flow. Its details are provided in the following.

KDE approximation. To reduce the memory usage, we design a KDE attention layer by approximating $\mathbf{x}[t]$ and $\mathbf{y}[t]$ using the multi-dimensional KDE:

$$\begin{aligned}\hat{f}(\mathbf{x}|\mathbf{x}[t])&=\frac{1}{Mh^c}\sum_{i=1}^{M}\prod_{c=1}^{C}K\left(\frac{\mathbf{x}^c-\mathbf{x}_i^c[t]}{h_{c,x}}\right)\\\hat{g}(\mathbf{y}|\mathbf{y}[t])&=\frac{1}{Nh^c}\sum_{i=1}^{N}\prod_{c=1}^{C}K\left(\frac{\mathbf{y}^c-\mathbf{y}_i^c[t]}{h_{c,y}}\right)\end{aligned}\tag{29}$$

$$\begin{aligned}h_{c,x}&=\frac{\max(\mathbf{x}^c[t])-\min(\mathbf{x}^c[t])}{\text{Bin}}\\h_{c,y}&=\frac{\max(\mathbf{y}^c[t])-\min(\mathbf{y}^c[t])}{\text{Bin}}\end{aligned}\tag{30}$$

where $K(\cdot)$ represents the standard Gaussian kernel with a covariance matrix $\sigma^2\mathbf{I}$ (σ is set to 0.1 by default). $\mathbf{x}^c\in\mathbb{R}^{M\times 1}$ and $\mathbf{y}^c\in\mathbb{R}^{N\times 1}$ represent the c-th column of $\mathbf{x}$ and $\mathbf{y}$, respectively. $h_{c,x}$ and $h_{c,y}$ define the kernel window sizes. They are determined using Eq. (30) with $\text{Bin}\in\mathbb{N}_+$ as a constant. Eq. (29) indicates that $\hat{f}(\mathbf{x}|\mathbf{x}[t])$ and $\hat{g}(\mathbf{y}|\mathbf{y}[t])$ are the smoothed continuous distribution of $\mathbf{x}[t]$ and $\mathbf{y}[t]$. Bin regulates the resolution of the distribution. However, $\hat{f}(\mathbf{x}|\mathbf{x}[t])$ and $\hat{g}(\mathbf{y}|\mathbf{y}[t])$ cannot be directly used for computing self- and cross-attention calculations since the attention layer captures

the correlation between discrete elements, whereas distributions $\hat{f}(\mathbf{x}|\mathbf{x}[t])$ and $\hat{g}(\mathbf{y}|\mathbf{y}[t])$ are continuous.

KDE based attention layer. To solve the above problem, we develop a KDE-based neural diffusion approach for I2P registration. Since $\hat{f}(\mathbf{x}|\mathbf{x}[t])$ and $\hat{g}(\mathbf{y}|\mathbf{y}[t])$ reflect distributions in the high-dimensional space, we design self- and cross-attention mechanisms to capture correlations across each dimension. This enables the construction of a correlation matrix as follows:

$$\mathbf{A}_{\hat{f}\hat{g}} = \text{Softmax}\left\{\begin{pmatrix} \langle \hat{f}_1, \hat{g}_1 \rangle & \cdots & \langle \hat{f}_1, \hat{g}_c \rangle \\ \vdots & \ddots & \vdots \\ \langle \hat{f}_c, \hat{g}_1 \rangle & \cdots & \langle \hat{f}_c, \hat{g}_c \rangle \end{pmatrix}\right\} \in \mathbb{R}^{c\times c} \quad (31)$$

$$\langle \hat{f}_i, \hat{g}_j \rangle = \int_{-\infty}^{+\infty} \omega_x(\hat{f}_i(\mathbf{x}|\mathbf{x}[t])) \cdot \omega_y(\hat{g}_j(\mathbf{x}|\mathbf{y}[t]))dx \quad (32)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in the L^2 function space. $\omega_x(\cdot)$ and $\omega_y(\cdot)$ are learnable functions corresponding to $\mathbf{w}_x$ and $\mathbf{w}_y$ in Eq. (10). Since $\hat{f}$ and $\hat{g}$ are multi-dimensional distributions, $\hat{f}_i$ and $\hat{g}_j$ represent the distributions in the i-th and j-th dimensions. Similar to Eq. (31), we can construct $\mathbf{A}_{\hat{f}\hat{f}}$, $\mathbf{A}_{\hat{g}\hat{g}}$, and $\mathbf{A}_{\hat{g}\hat{f}}$. To reduce the computation burden in Eq. (32), we normalize the distributions and compute the integration as:

$$\langle \hat{f}_i, \hat{g}_j \rangle \approx \begin{pmatrix} \hat{f}_i^1 \\ \vdots \\ \hat{f}_i^{\text{Bin}} \end{pmatrix}^T \Omega_x^T \Omega_y \begin{pmatrix} \hat{g}_j^1 \\ \vdots \\ \hat{g}_j^{\text{Bin}} \end{pmatrix} \quad (33)$$

where $\hat{f}_i^1$, ..., $\hat{f}_i^{\text{Bin}}$ are values of $\hat{f}_i(\mathbf{x}|\mathbf{x}[t])$ uniformly sampled over $[\min(\mathbf{x}^i[t]), \max(\mathbf{x}^i[t])]$ and $\hat{g}_j^1$, ..., $\hat{g}_j^{\text{Bin}}$ are values of $\hat{g}_j(\mathbf{y}|\mathbf{y}[t])$ uniformly sampled over $[\min(\mathbf{y}^j[t])$ and $\max(\mathbf{y}^j[t])]$. $\Omega_x \in \mathbb{R}^{\text{Bin}\times\text{Bin}}$ and $\Omega_y \in \mathbb{R}^{\text{Bin}\times\text{Bin}}$ are learnable matrices denoting the discretization of $\omega_x(\cdot)$ and $\omega_y(\cdot)$.

Next, we construct layers of KDE-based self-attention (KSA) and KDE-based cross-attention (KCA) as follows:

$$\text{KSA}_I(\mathbf{x}[t]) = \mathbf{x}[t] \cdot \mathbf{A}_{\hat{f}\hat{f}}, \quad \text{KSA}_P(\mathbf{y}[t]) = \mathbf{y}[t] \cdot \mathbf{A}_{\hat{g}\hat{g}} \quad (34)$$

$$\begin{aligned} \text{KCA}_I(\mathbf{x}[t], \mathbf{y}[t]) &= \mathbf{x}[t] \cdot \mathbf{A}_{\hat{f}\hat{g}} \\ \text{KCA}_P(\mathbf{y}[t], \mathbf{x}[t]) &= \mathbf{y}[t] \cdot \mathbf{A}_{\hat{g}\hat{f}} \end{aligned} \quad (35)$$

By replacing SA$(\cdot)$ and CA$(\cdot)$ with KSA$(\cdot)$ and KCA$(\cdot)$ in **Definition 2**, we construct a KDE based neural diffusion model for I2P registration. This model is similar to Eqs. (22-24), with only Eq. (23) replaced as follows:

$$\begin{cases} \mathbf{x}[t+1] = (1-\tau)\mathbf{x}[t] \\ \qquad\qquad + \tau(\text{KSA}_I(\mathbf{x}[t]) + \text{KCA}_I(\mathbf{x}[t], \mathbf{y}[t])) \\ \mathbf{y}[t+1] = (1-\tau)\mathbf{y}[t] \\ \qquad\qquad + \tau(\text{KSA}_P(\mathbf{y}[t]) + \text{KCA}_P(\mathbf{y}[t], \mathbf{x}[t])) \end{cases} \quad (36)$$

D. Analyzing surface normals for I2P registration

In this section, we discuss why surface normals, specifically, the image derived normals (SN_{2d}), and point cloud derived normals (SN_{3d}), are beneficial for I2P registration. In general, SN_{2d} and SN_{3d} are given in the camera coordinate system and world coordinate system, respectively. For a correspondence $\langle I_i, P_j \rangle$, where $\mathbf{n}_i^{2d} \in \text{SN}_{2d}$ corresponds to pixel I_i and $\mathbf{n}_j^{3d} \in \text{SN}_{3d}$ corresponds to point P_j, the following constraint must be satisfied:

$$\mathbf{n}_i^{2d} = \mathbf{R} \cdot \mathbf{n}_j^{3d}, \ \mathbf{n}_i^{2d} \in \text{SN}_{2d}, \ \mathbf{n}_j^{3d} \in \text{SN}_{3d} \quad (37)$$

where $\mathbf{R}$ is the rotation matrix defined in Eq. (1). This constraint implies that valid correspondences must preserve consistency in surface normals under rotation, thus serving as a valuable correspondence cue during registration.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-025-02575-4.

Acknowledgements This work is partially supported by the National Key R&D Program of China (Grant ID: 2024YFC3015302), National Natural Science Foundation of China (62372377), and China Postdoctoral Science Foundation (Grant ID: 2024M761014).

Data Availability All data used in this paper are available upon request by contacting the corresponding authors.

Declarations

Conflicts of Interest The authors declare no conflict of interest.

References

An, P., Gao, Y., Ma, T., Yu, K., Fang, B., Zhang, J., & Ma, J. (2020). Lidar-camera system extrinsic calibration by establishing virtual point correspondences from pseudo calibration objects. *Opt. Express, 28*, 18261–18282.

An, P., Liang, J., Yu, K., Fang, B., & Ma, J. (2022). Deep structural information fusion for 3d object detection on lidar-camera system. *Comput. Vis. Image Underst., 214*, Article 103295.

An P, Ding J, Quan S, Yang J, Yang Y, Liu Q, & Ma J (2024a) Survey of extrinsic calibration on lidar-camera system for intelligent vehicle: Challenges, approaches, and trends. IEEE Trans Intell Transp Syst Early Access(1):1–25

An, P., Hu, X., Ding, J., Zhang, J., Ma, J., Yang, Y., & Liu, Q. (2024). Ol-reg: Registration of image and sparse lidar point cloud with object-level dense correspondences. *IEEE Trans. Circuits Syst. Video Technol., 1*(1), 1–15.

Bae G, & Davison AJ (2024) Rethinking inductive biases for surface normal estimation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 1–10

Bastico M, Decencière E, Corté L, Tillier Y, & Ryckelynck D (2024) Coupled laplacian eigenmaps for locally-aware 3d rigid point cloud matching. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3447–3458

Brachmann E, Cavallari T, & Prisacariu VA (2023) Accelerated coordinate encoding: Learning to relocalize in minutes using RGB and poses. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5044–5053

Campbell D, Liu L, & Gould S (2020) Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization. In: Proceedings of European Conference on Computer Vision, pp 244–261

Chamberlain B, Rowbottom J, Eynard D, Giovanni FD, Dong X, & Bronstein MM (2021a) Beltrami flow and neural diffusion on graphs. In: Proceedings of Advances in Neural Information Processing Systems, pp 1594–1609

Chamberlain B, Rowbottom J, Gorinova MI, Bronstein MM, Webb S, & Rossi E (2021b) GRAND: graph neural diffusion. In: Proceedings of the 38th International Conference on Machine Learning, vol 139, pp 1407–1418

Chang M, Mangelson JG, Kaess M, & Lucey S (2021) Hypermap: Compressed 3d map for monocular camera registration. In: Proceedings of IEEE International Conference on Robotics and Automation, pp 11739–11745

Chen H, Wang P, Wang F, Tian W, Xiong L, & Li H (2022) Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2771–2780

Cheng, Y., Huang, Z., Quan, S., Cao, X., Zhang, S., & Yang, J. (2023). Sampling locally, hypothesis globally: accurate 3d point cloud registration with a ransac variant. *Visual Intelligence, 20*, 1–15.

Cheng Z, Deng J, Li X, Yin B, & Zhang T (2025) Bridge 2d-3d: Uncertainty-aware hierarchical registration network with domain alignment. In: Proceedings of AAAI Conference on Artificial Intelligence, pp 2491–2499

Choy CB, Park J, & Koltun V (2019) Fully convolutional geometric features. In: Proceedings of IEEE/CVF International Conference on Computer Vision, pp 8957–8965

Dai A, Chang AX, Savva M, Halber M, Funkhouser TA, & Nießner M (2017) Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 2432–2443

David, P., DeMenthon, D., Duraiswami, R., & Samet, H. (2004). Softposit: Simultaneous pose and correspondence determination. *Int. J. Comput. Vis., 59*(3), 259–284.

Dusmanu M, Rocco I, Pajdla T, Pollefeys M, Sivic J, Torii A, & Sattler T (2019) D2-net: A trainable CNN for joint description and detection of local features. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 8092–8101

Feng M, Hu S, Ang MH, & Lee GH (2019) 2D3D-Matchnet: Learning to match keypoints across 2D image and 3D point cloud. In: Proceedings of IEEE International Conference on Robotics and Automation, pp 4790–4796

Geiger A, Lenz P, & Urtasun R (2012) Are we ready for autonomous driving? the KITTI vision benchmark suite. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 3354–3361

Glocker B, Izadi S, Shotton J, & Criminisi A (2013) Real-time RGB-D camera relocalization. In: Proceedings of IEEE International Symposium on Mixed and Augmented Reality, pp 173–179

He K, Zhang X, Ren S, & Sun J (2016) Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778

Huang S, Gojcic Z, Usvyatsov M, Wieser A, & Schindler K (2021) Predator: Registration of 3d point clouds with low overlap. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 4267–4276

Kang, S., Liao, Y., Li, J., Liang, F., Li, Y., Zou, X., Li, F., Chen, X., Dong, Z., & Yang, B. (2024). Cofii2p: Coarse-to-fine correspondences-based image to point cloud registration. *IEEE Robotics Autom Lett, 9*(11), 10264–10271.

Kim M, Koo J, & Kim G (2023) Ep2p-loc: End-to-end 3d point to 2d pixel localization for large-scale visual localization. In: Proceedings of IEEE/CVF International Conference on Computer Vision, pp 21470–21480

Lai K, Bo L, & Fox D (2014) Unsupervised feature learning for 3d scene labeling. In: Proceedings of IEEE International Conference on Robotics and Automation, pp 3050–3057

Lepetit, V., Moreno-Noguer, F., & Fua, P. (2009). EPnP: An accurate O(n) solution to the pnp problem. *Int. J. Comput. Vis., 81*(2), 155–166.

Li J, & Lee GH (2021) DeepI2P: Image-to-point cloud registration via deep classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 15960–15969

Li M, Qin Z, Gao Z, Yi R, Zhu C, Guo Y, & Xu K (2023) 2D3D-MATR: 2D-3D matching transformer for detection-free registration between images and point clouds. In: Proceedings of IEEE Conference on Computer Vision, pp 1–10

Liu, S., Suganuma, M., & Okatani, T. (2024). Symmetry-aware neural architecture for embodied visual navigation. *Int. J. Comput. Vis., 132*(4), 1091–1107.

Liu Z, Tang H, Zhu S, & Han S (2021) SemAlign: Annotation-free camera-lidar calibration with semantic alignment loss. In: Proceedings of IEEE International Conference on Intelligent Robots and Systems, pp 8845–8851

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis., 60*(2), 91–110.

Lv, J., Lang, X., Xu, J., Wang, M., Liu, Y., & Zuo, X. (2023). Continuous-time fixed-lag smoothing for lidar-inertial-camera slam. *IEEE/ASME Trans. Mechatron., 28*(4), 2259–2270.

Matsumoto Y, Nakano G, & Ogura K (2024) Indoor visual localization using point and line correspondences in dense colored point cloud. In: Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision, pp 3604–3613

Miao J, Jiang K, Wen T, Wang Y, Jia P, Zhao X, Xiao Z, Huang J, Zhong Z, & Yang D (2023) A survey on monocular re-localization: From the perspective of scene map representation. CoRR abs/2311.15643

Moreno-Noguer, F., Lepetit, V., & Fua, P. (2008). Pose priors for simultaneously solving alignment and correspondence. *Proceedings of European Conference on Computer Vision, 5303*, 405–418.

Pham Q, Uy MA, Hua B, Nguyen DT, Roig G, & Yeung S (2020) LCD: learned cross-domain descriptors for 2d-3d matching. In: Proceedings of AAAI Conference on Artificial Intelligence, pp 11856–11864

Pillaud-Vivien, L., & Bach, F. R. (2023). Kernelized diffusion maps. *Proceedings of The Thirty Sixth Annual Conference on Learning Theory, 195*, 5236–5259.

Qin Z, Yu H, Wang C, Guo Y, Peng Y, & Xu K (2022) Geometric transformer for fast and robust point cloud registration. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 11133–11142

Ren, S., Zeng, Y., Hou, J., & Chen, X. (2023). CorrI2P: Deep image-to-point cloud registration via dense correspondence. *IEEE Trans. Circuits Syst. Video Technol., 33*(3), 1198–1208.

Sarlin P, DeTone D, Malisiewicz T, & Rabinovich A (2020) Superglue: Learning feature matching with graph neural networks. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4937–4946

Schroff F, Kalenichenko D, & Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 815–823

Sun Y, Cheng C, Zhang Y, Zhang C, Zheng L, Wang Z, & Wei Y (2020) Circle loss: A unified perspective of pair similarity optimization.

In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6397–6406

Thomas H, Qi CR, Deschaud J, Marcotegui B, Goulette F, & Guibas LJ (2019) Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of IEEE/CVF International Conference on Computer Vision, pp 6410–6419

Thomas, H., Zhang, J., & Barfoot, T. D. (2023). The foreseeable future: Self-supervised learning to predict dynamic scenes for indoor navigation. *IEEE Trans Robotics, 39*(6), 4581–4599.

Vechersky, P., Cox, M., Borges, P. V. K., & Lowe, T. (2018). Colourising point clouds using independent cameras. *IEEE Robotics Autom Lett, 3*(4), 3575–3582.

Wang B, Chen C, Cui Z, Qin J, Lu CX, Yu Z, Zhao P, Dong Z, Zhu F, Trigoni N, & Markham A (2021) P2-Net: Joint description and detection of local features for pixel and point matching. In: Proceedings of IEEE International Conference on Computer Vision, pp 15984–15993

Wang H, Liu Y, Wang B, Sun Y, Dong Z, Wang W, & Yang B (2024) Freereg: Image-to-point cloud registration leveraging pretrained diffusion models and monocular depth estimators. In: Proceedings of International Conference on Learning Representation, pp 1–24

Wang S, Li BZ, Khabsa M, Fang H, & Ma H (2020) Linformer: Self-attention with linear complexity. CoRR abs/2006.04768

Wu B, Ma J, Chen G, & An P (2021) Feature interactive representation for point cloud registration. In: Proceedings IEEE/CVF International Conference on Computer Vision, pp 5510–5519

Wu, Q., Jiang, H., Luo, L., Li, J., Ding, Y., Xie, J., & Yang, J. (2024). Diff-reg: Diffusion model in doubly stochastic matrix space for registration problem. *Proceedings of European Conference on Computer Vision, 15123*, 160–178.

Xu, Y., Lin, K., Zhang, G., Wang, X., & Li, H. (2024). Rnnpose: 6-dof object pose estimation via recurrent correspondence field estimation and pose optimization. *IEEE Trans. Pattern Anal. Mach. Intell., 46*(7), 4669–4683.

Yang, H., & Carlone, L. (2023). Certifiably optimal outlier-robust geometric perception: Semidefinite relaxations and scalable global optimization. *IEEE Trans. Pattern Anal. Mach. Intell., 45*(3), 2816–2834.

Yang, J., Xian, K., Wang, P., & Zhang, Y. (2021). A performance evaluation of correspondence grouping methods for 3d rigid data matching. *IEEE Trans. Pattern Anal. Mach. Intell., 43*(6), 1859–1874.

Yao G, Xuan Y, Li X, & Pan Y (2024) Cmr-agent: Learning a cross-modal agent for iterative image-to-point cloud registration. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, pp 13458–13465

Ye H, Huang H, & Liu M (2020) Monocular direct sparse localization in a prior 3d surfel map. In: Proceedings of IEEE International Conference on Robotics and Automation, pp 8892–8898

Yin, H., Xu, X., Lu, S., Chen, X., Xiong, R., Shen, S., Stachniss, C., & Wang, Y. (2024). A survey on global lidar localization: Challenges, advances and open problems. *Int. J. Comput. Vis., 1*(1), 1–33.

Yu H, Ye W, Feng Y, Bao H, & Zhang G (2020) Learning bipartite graph matching for robust visual localization. In: Proceedings of IEEE International Symposium on Mixed and Augmented Reality, pp 146–155

Yuan, C., Liu, X., Hong, X., & Zhang, F. (2021). Pixel-level extrinsic self calibration of high resolution lidar and camera in targetless environments. *IEEE Robotics Autom Lett, 6*(4), 7517–7524.

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell., 22*(11), 1330–1334.

Zhou J, Ma B, Zhang W, Fang Y, Liu Y, & Han Z (2023) Differentiable registration of images and lidar point clouds with voxelpoint-to-pixel matching. In: Proceedings of Advances in Neural Information Processing Systems, pp 1–10

Zhou, Q., Agostinho, S., Osep, A., & Leal-Taixé, L. (2022). Is geometry enough for matching in visual localization? *Proceedings of European Conference on Computer Vision, 13670*, 407–425.