Top-I2P: Explore Open-Domain Image-to-Point Cloud Registration Using Topology Relationship

Pei An[1], **Jiaqi Yang**[2], **Muyao Peng**[1], **You Yang**[1], **Qiong Liu**[1], **Jie Ma**[1] and **Liangliang Nan**[3]
[1]Huazhong University of Science and Technology, China
[2]Northwestern Polytechnical University, China
[3]Delft University of Technology, Netherlands
{anpei96, muyao99, yangyou, q.liu, majie}@hust.edu.cn, jqyang@nwpu.edu.cn,
liangliang.nan@tudelft.nl

Abstract

Image-to-point cloud (I2P) registration is a fundamental task in computer vision, which aims to align pixels in 2D images with corresponding points in 3D point clouds. While deep learning based methods dominate this field, they often fail to generalize to the open domain. In this paper, we address open-domain I2P registration from the topology relationship perspective. Firstly, we find that topology relationship reflect sparse connections between pixels and points, which shows the significant potential in enhancing cross-modality feature interaction in the open domain. Building on this insight, we develop an I2P registration framework using topology relationship. After that, to construct and leverage the topology relationship between the heterogeneous 2D and 3D spaces, we design a registration network, Top-I2P, with correction-based topology reasoning and fast topology feature interaction modules. Extensive experiments on 7-Scenes, RGBD-V2, ScanNet, and self-collected I2P datasets demonstrate that Top-I2P achieves superior registration performance in open-domain scenarios.

1 Introduction

Image-to-point cloud (I2P) registration is a fundamental task in computer vision [An *et al.*, 2024a]. Given an image and a point cloud, it establishes the pixel-to-point correspondences to estimate the six degrees of freedom (6-DoF) camera poses [Wang *et al.*, 2021]. I2P registration is critical for applications, such as visual localization, visual navigation, multi-sensor calibration, augmented reality (AR), object retrieval, and object pose estimation [Cheng *et al.*, 2023b].

Deep learning is a crucial technique for I2P registration, as it effectively bridges the modality gap between 2D images and 3D point clouds. To improve the capability of neural networks for I2P registration, researchers tend to revise the neural networks in image registration (i.e., SuperGlue [Sarlin *et al.*, 2020]) and point clouds registration (i.e., Geotrans [Qin *et al.*, 2022]) for I2P registration [Pham *et al.*, 2020].

Although current I2P registration approaches [Wang *et al.*, 2021; Li *et al.*, 2023] have made progress, domain generalization on unseen scenes remains a major challenge for existing learning based I2P methods, especially in the **open-domain** I2P registration scenario. For a better illustration, we conducted a case study of 2D3D-MATR [Li *et al.*, 2023] on the 7-Scenes [Glocker *et al.*, 2013], as shown in Fig. 1. We found that the quality of correspondences of 2D3D-MATR is not accurate in unseen scenarios, making it unsafe for downstream applications (i.e., 6 DoF visual localization [Kim *et al.*, 2023] and visual navigation [van Dijk *et al.*, 2024]). This underscores the need to explore open-domain I2P registration.

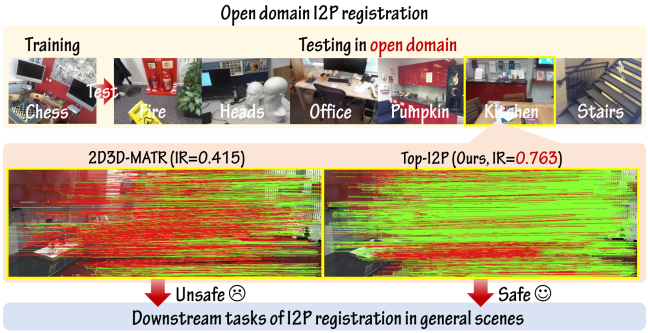

Figure 1: Open-domain I2P registration problem. The I2P registration model is tested on the unseen scenes (open domain). Compared with the state-of-the-art method, Top-I2P achieves a high-quality result, which is safe to the downstream tasks of I2P registration. IR is the acronym for inlier ratio.

In 2024, Wang *et al.* were the first to discuss open-domain I2P registration with presenting a neural network named as FreeReg [Wang *et al.*, 2024]. Their method is computationally expensive (9 seconds per pair, $\geq$20 GB GPU memory) and lacks precision due to its use of a loose RMSE threshold (0.3m). These limitations make it unsuitable for real-time or high-accuracy applications. Thus, it is essential to explore a more effective and accurate I2P registration method in the open domain.

This paper addresses open-domain I2P registration from the perspective of topology relationship [Egenhofer, 1993]. We first explore the **connection between topology relationship and I2P registration**, and find that topology relation-

ship reflect the sparse connection between pixels and points. It shows the salient potential in enhancing cross-modality feature interaction in the open domain. Based on this insight, we address the problem of **learning and leveraging topology relationship** between the heterogeneous 2D and 3D spaces, and present a registration framework that incorporates topology relationship reasoning and topology feature interaction.

Furthermore, to effectively build and leverage topology relationship, we propose a dedicated neural network Top-I2P that incorporates correction-based topology reasoning (CTR) and fast topology feature interaction (FTI) modules to balance feature capability and inference speed. Extensive experiments are conducted on the 7-Scenes [Glocker *et al.*, 2013], RGBD-V2 [Lai *et al.*, 2014], ScanNet [Dai *et al.*, 2017], and self-collected datasets. Results demonstrate that the proposed Top-I2P achieves a superior inlier ratio (IR) and registration recall (RR) than state-of-the-art methods in open domain testing.

Our core contribution is **rethinking pixel-to-point matching mechanism from a topology relationship viewpoint**. To achieve this goal, we design an I2P registration network based on topology relationship, and it has better performance in the open domain.

2 Related Works

We briefly overview the mainstream literature related to I2P registration and topology relationship.

2.1 Learning based I2P registration

I2P registration aims to build pixel-to-point correspondences $\langle q, p\rangle$ from image and point cloud where $\langle q, p\rangle$ is defined as:

$$q = \pi(\mathbf{T} \circ p) \tag{1}$$

where $q \in \mathbb{R}^2$ represents a 2D pixel in image $\mathcal{I}$, and $p \in \mathbb{R}^3$ represents a 3D point in point cloud $\mathcal{P}$. $\mathbf{T}$ is a transformation from the world coordinate system to the camera coordinate system. $\pi(\cdot)$ is a projection operator. As $\mathbf{T}$ is unknown in I2P registration task setting, $\langle q, p\rangle$ is determined in a **feature-matching** manner:

$$d(\mathbf{f}_q^{\text{2D}}, \mathbf{f}_p^{\text{3D}}) \leq \delta \Rightarrow \langle q, p\rangle \text{ is a correspondence} \tag{2}$$

$$\mathbf{F}_{\text{I}}, \mathbf{F}_{\text{P}} = \varphi(\mathcal{I}, \mathcal{P}) \tag{3}$$

where $d(\cdot, \cdot)$ is a normalized L2 distance and δ is a threshold. $\mathbf{f}_q^{\text{2D}}$ and $\mathbf{f}_p^{\text{3D}}$ are features of q and p. $\varphi(\cdot, \cdot)$ is the network of I2P registration. $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$ is an image, $\mathcal{P} \in \mathbb{R}^{N \times C}$ is a point cloud with C-dimensional features. $\mathbf{f}_q^{\text{2D}}, \mathbf{f}_p^{\text{3D}}$ are queried from $\mathbf{F}_{\text{I}}, \mathbf{F}_{\text{P}}$ at q, p. Current works focus on designing $\varphi(\cdot, \cdot)$.

In 2019, Feng *et al*. [Feng *et al.*, 2019] was the first to develop a network to learn the descriptors of image patches and point cloud patches. Triplet loss [Schroff *et al.*, 2015] is used to eliminate the adverse effects of a limited number of inliers. Li *et al*. presented Deep-I2P to interact with the global features from images and point clouds [Li and Lee, 2021]. Their approach predicts the overlap region in 3D space, which improves the stability of I2P registration. Wang *et al*. polished the image registration neural network [Dusmanu *et al.*, 2019] as P2-Net for I2P registration [Wang *et al.*, 2021]. Building on work [Li and Lee, 2021], Ren *et al*. refined feature interaction scheme with transformers and established a neural network Corr-I2P [Ren *et al.*, 2023], which improves visual localization accuracy in the outdoor scenes. Kim *et al*. established a neural network EP2P-Loc to deal with visual localization in the pre-built map [Kim *et al.*, 2023]. Li *et al*. found that P2-Net [Wang *et al.*, 2021] suffers from a few inliers, and they attempted to solve it by designing 2D3D-MATR [Li *et al.*, 2023]. It takes 2D-3D patch matching as guidance for the accurate registration. Taking Corr-I2P [Ren *et al.*, 2023] as baseline, Zhou *et al*. exploited a refined circle loss [Sun *et al.*, 2020] and a differential perspective-n-point (PnP) loss to improve the geometric consistency of I2P correspondences [Zhou *et al.*, 2023]. Wu *et al*. leveraged diffusion model in de-nosing pixel-to-point correspondences [Wu *et al.*, 2024b].

Despite the advancements in I2P registration, there is limited research targeting open-domain scenarios. To address this gap, we propose Top-I2P to leverage topology relationship for robust I2P registration in a challenging environment.

2.2 Learning with topology relationship

Topology relationship are the fundamental relationships[1] of the space [Egenhofer, 1993]. As these relationships are invariant to topology transformation, many researchers exploit the topology relationship in computer vision and robotics.

Wang *et al*. developed a topology reasoning benchmark for high definition (HD) mapping task [Wang *et al.*, 2023]. They developed an entity-specific similarity measure to evaluate the similarity of two topology graphs. Chen *et al*. studied semantic topology reasoning for the domain generalization task [Chen *et al.*, 2022]. Domain invariance features are learnt by reasoning semantic topology graphs from cross-domain data. Cheng *et al*. used topology relationship for pedestrian re-identification [Cheng *et al.*, 2023a]. A multi-camera logical topology graph is built according to the cameras' localizations and image features. This topology graph is used in a graph neural network (GNN) to predict pedestrian detection results. In the autonomous driving scenario, Wu *et al*. used multilayer perceptron (MLP) to construct topology reasoning networks to predict lane-lane and lane-traffic topology relationship [Wu *et al.*, 2024a]. Li *et al*. designed a graph contrastive learning scheme. Its highlight is to learn both graph-level and subgraph-level topology isomorphism knowledge [Li *et al.*, 2024]. Towards 3D indoor scene understanding, Zhang *et al*. designed a graph to record the relation of place-place and place-object. This graph benefits various downstream tasks, such as place recognition and 3D reconstruction [Zhang *et al.*, 2024].

In summary, topology relationship provide a lightweight, fundamental, and stable representation of the 3D space. This concept has garnered attention across various fields, not limited to computer vision, robotics, and autonomous driving. In this paper, we study topology relationship for open-domain I2P registration.

[1]This term is used to describe the *disjoint*, *meet*, *overlaps*, *covers*, *contains*, *equal*, *converBy*, and *inside* relationships of two sets.

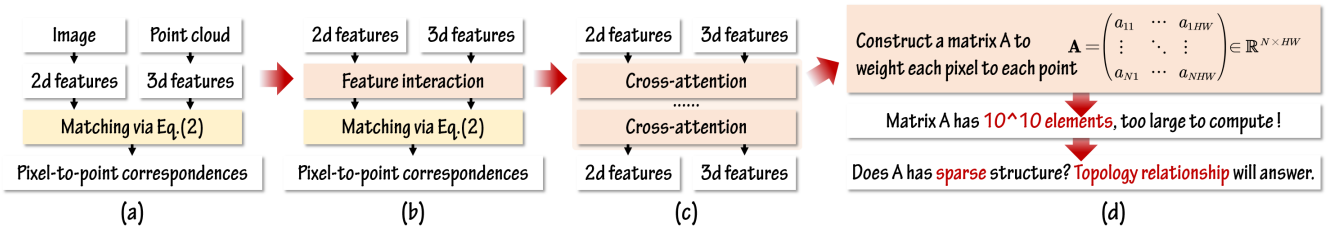

Figure 2: Bottleneck analysis of I2P registration. (a) Overview of the mainstream architecture. (b) Feature interaction is integrated to reduce modality differences and mitigate overfitting. (c) Pixel-level cross-attention forms the central component of feature interaction. (d) Bottleneck analysis of the bottleneck in cross-attention, highlighting challenges in computation and feature alignment.

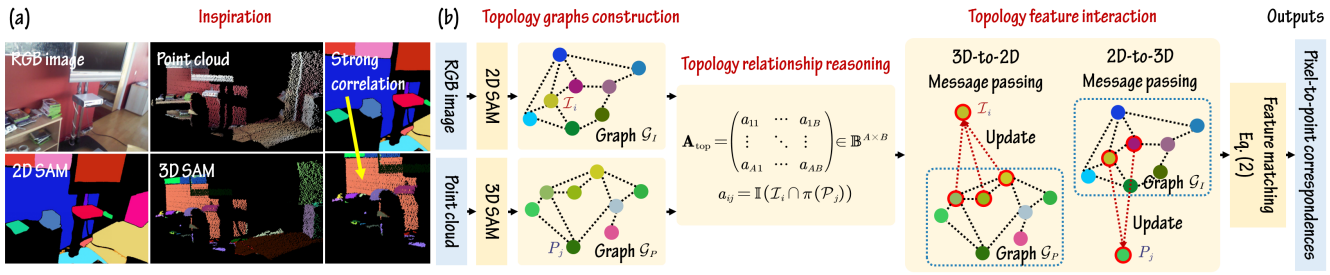

Figure 3: Proposed I2P open-domain registration framework leveraging topology relationship. (a) Observations reveal a strong correlation between patches segmented using the 2D and 3D SAM (Segment Anything Model). This is because existing 3D SAM models are often trained using labels generated by 2D SAM. As a result, 3D SAM models predict segments with structures highly similar to 2D SAM. (b) Building on this insight, we aim to predict the topology relationship and leverage it to enhance feature interaction for robust I2P registration.

2.3 Discussions

The use of topology relationship in I2P registration remains an open question of interest. Li *et al* were the first to apply topology relationship by predicting the intersection relation of 2D and 3D patches [Li *et al.*, 2023]. However, their approach employs a simplistic patch partitioning scheme, which relies on 2D square patches and 3D spherical patches. This limitation results in suboptimal performance, particularly in unseen scenes, as shown in Fig. 1. To address these shortcomings, we consider a patch with the general shape and use a visual fundamental model for patch partition, enabling a broader applicability and improved performance compared to the prior method [Li *et al.*, 2023].

3 Open-Domain I2P Registration

We first analyze the bottleneck in open-domain I2P registration. Then, we illustrate why topology relationship can solve this bottleneck and provide a complete framework. After that, we implement this framework by designing Top-I2P.

3.1 Bottleneck analysis and topology relationship

Why is the current I2P registration method unstable (as seen in Fig. 1(b))? The instability arises primarily from **the lack of pixel-level feature interaction**. We illustrate this issue in depth. Without feature interaction, image or point cloud feature extractors operate independently (as shown in Fig. 2(a)), making the 2D and 3D features prone to overfitting on training data. To overcome the overfitting, one scheme is to incorporate a feature interaction module as shown in Fig. 2(b). It facilitates message exchange between the cross-modality data, and minimizes the feature disparity between correspondences, aiding in **domain-invariant** features learning [Wu *et al.*, 2021].

Pixel-level feature interaction is the ideal scheme to model the relationships of each pixel to each point. Pixel-level cross-attention is its core component, and it constructs a matrix $\mathbf{A} \in \mathbb{R}^{N \times HW}$ that records the weights between each pixel in $\mathcal{I}$ and each point in $\mathcal{P}$ (as shown in Fig. 2(c), (d)).

However, $\mathbf{A}$ becomes prohibitively large. For a 640×480 image and a point cloud with 10^5 points, $\mathbf{A}$ **contains 10^{10} elements**, which is impractical to store and compute. Hence, pixel-level feature interaction represents a major bottleneck for open-domain I2P registration.

To address this bottleneck, we find that **topology relationship between 2D and 3D spaces reflect the sparsity of $\mathbf{A}$**. For one pixel q and one point p, if their neighboring regions have no interaction (i.e., $\mathcal{N}_{2D}(q) \cap \pi(\mathbf{T} \circ \mathcal{N}_{3D}(p)) = \emptyset$), there is no need to interact features between q and p. From Eq. (5) in Sec. 3.2, we construct $\mathbf{A}$ only with nearly 10^4 elements, 10^6 **times smaller** than the original scheme. Hence, topology relationship significantly reduce the computational burden of pixel-level feature interaction.

3.2 A framework with topology relationship

As topology relationship show significant potential in feature interaction, we attempt to leverage topology relationship and derive a lightweight and efficient framework for open-domain I2P registration, where an overview of the framework is shown in Fig. 3. Its sub-modules are discussed as follows.

Topology graphs construction. Using the segmentation anything model (SAM) in image and point cloud [Kirillov *et al.*, 2023; Zhou *et al.*, 2024], $\mathcal{I}$ and $\mathcal{P}$ are decomposed as arbitrary shapes without overlaps:

$$\mathcal{P} = \bigcup_{i}^{A} \mathcal{P}_i,\ \mathcal{I} = \bigcup_{j}^{B} \mathcal{I}_j \tag{4}$$

where A and B are the patch numbers. $\mathcal{P}_i$ and $\mathcal{I}_j$ are visualized with different colors in Fig. 3. Complete graphs $\mathcal{G}_P$, $\mathcal{G}_I$ are constructed from $\{\mathcal{P}_i\}_{i=1}^{A}$ and $\{\mathcal{I}_j\}_{j=1}^{B}$ where $\mathcal{P}_i$ and $\mathcal{I}_j$ are graph nodes. $\mathcal{G}_I$ and $\mathcal{G}_P$ are named as 2D and 3D graphs.

Besides, we discuss the reason why 2D and 3D patches are obtained by SAM instead of the regular shapes (i.e., 2D square and 3D sphere) or 3D superpoint and 2D superpixel. First, unlike regular shapes, patches segmented by SAM have a semantic correlation, helpful to topology relationship reasoning. Second, SAM is faster than the superpoint and superpixel algorithms. Third, as pixel distance is not aligned with metric distance, regular shapes suffer from the scale selection [Li *et al.*, 2023], while patches segmented by SAM do not.

Topology relationship reasoning. This step is to learn a sparse $\mathbf{A}$ by topology relationship reasoning on $\mathcal{G}_I$ and $\mathcal{G}_P$. As $\mathcal{G}_I$ and $\mathcal{G}_P$ are heterogeneous graphs on 2D and 3D spaces, learning $\mathbf{A}$ is a challenging task. We simplify this task in such an approximate way. For $p \in \mathcal{P}_i$ and $q \in \mathcal{I}_j$, if $\pi(\mathcal{P}_i) \cap \mathcal{I}_j = \emptyset$, $\mathbf{f}_q^{\text{2D}}$ has no contribution to $\mathbf{f}_p^{\text{3D}}$. Thus, we can approximately construct a sparse $\mathbf{A}$ as:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \cdots & \mathbf{A}_{1B} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{A1} & \cdots & \mathbf{A}_{AB} \end{pmatrix}, \mathbf{A}_{ij} = \begin{cases} \mathbf{1}, & \pi(\mathcal{P}_i) \cap \mathcal{I}_j \neq \emptyset \\ \mathbf{0}, & \pi(\mathcal{P}_i) \cap \mathcal{I}_j = \emptyset \end{cases} \tag{5}$$

where $\mathbf{1}$ and $\mathbf{0}$ are matrices with full of ones or zeros. $\mathbf{A}$ can be further collapsed into a small-sized boolean matrix $\mathbf{A}_{\text{top}} \in \mathbb{B}^{A \times B}$, $\mathbf{A}_{\text{top},ij} = \mathbb{I}(\pi(\mathcal{P}_i) \cap \mathcal{I}_j)$ where $\mathbb{I}(\cdot)$ is an indicator function. After figuring out the aim of topology relationship reasoning, this procedure can be represented as:

$$\hat{\mathbf{A}}_{\text{top}} = \phi_{\text{top}}(\mathcal{G}_I, \mathcal{G}_P) \tag{6}$$

where $\phi_{\text{top}}(\cdot,\cdot)$ is a self-defined topology relationship reasoning network. $\hat{\mathbf{A}}_{\text{top}} \in [0,1]^{A \times B}$ describes confidence score of whether $\pi(\mathcal{P}_i) \cap \mathcal{I}_j$.

Topology feature interaction. For the open-domain I2P registration, topology feature interaction is a crucial step to eliminate cross modality difference from $\mathbf{f}_q^{\text{2D}}$ and $\mathbf{f}_p^{\text{3D}}$ with the guidance of $\mathcal{G}_I$, $\mathcal{G}_P$, and $\hat{\mathbf{A}}_{\text{top}}$. This step contains (i) 3D-to-2D message passing and (ii) 2D-to-3D message passing.

At first, 3D-to-2D message passing is defined in a moving average manner (suppose $q \in \mathcal{I}_j$):

$$\mathbf{f}_q^{\text{2D}} \leftarrow \alpha \mathbf{f}_q^{\text{2D}} + (1-\alpha) \frac{\sum_i^A \hat{\mathbf{A}}_{\text{top},ij} \cdot \bar{\mathbf{f}}_i^{\text{3D}}}{\sum_i^A \hat{\mathbf{A}}_{\text{top},ij}} \tag{7}$$

$$\bar{\mathbf{f}}_i^{\text{3D}} = \frac{\sum_{k=1}^{N} \mathbb{I}(p_k \in \mathcal{P}_i) \cdot \mathbf{f}_k^{\text{3D}}}{\sum_{k=1}^{N} \mathbb{I}(p_k \in \mathcal{P}_i)} \tag{8}$$

where α is a smooth coefficient. $\bar{\mathbf{f}}_i^{\text{3D}}$ is the average feature in region $\mathcal{P}_i$. Thus, 3D-to-2D message passing is to update $\mathbf{f}_q^{\text{2D}}$ with average 3D features in regions that $\pi(\mathcal{P}_i) \cap \mathcal{I}_j \neq \emptyset$.

Similar to Eqs. (7) and (8), 2D-to-3D message passing is defined as:

$$\mathbf{f}_p^{\text{3D}} \leftarrow \alpha \mathbf{f}_p^{\text{3D}} + (1-\alpha) \frac{\sum_j^B \hat{\mathbf{A}}_{\text{top},ij} \cdot \bar{\mathbf{f}}_j^{\text{2D}}}{\sum_j^B \hat{\mathbf{A}}_{\text{top},ij}} \tag{9}$$

$$\bar{\mathbf{f}}_j^{\text{2D}} = \frac{\sum_{k=1}^{HW} \mathbb{I}(q_k \in \mathcal{I}_j) \cdot \mathbf{f}_k^{\text{2D}}}{\sum_{k=1}^{HW} \mathbb{I}(q_k \in \mathcal{I}_j)} \tag{10}$$

Finally, with the updated $\mathbf{f}_q^{\text{2D}}$ and $\mathbf{f}_p^{\text{3D}}$, whether $\langle q, p \rangle$ is a pixel-to-point correspondence can be determined via Eq. (2).

More discussions. We further discuss two issues of the proposed method. First, we only utilize a single topology relationship (i.e., intersection) in this paper. Actually, the general topology relationship has multiple relationships. 4 intersection model (4IM) [Egenhofer, 1993] is a traditional model to represent the multiple topology relations of two 2D sets $\mathcal{A}$ and $\mathcal{B}$:

$$\mathcal{R}_{\text{4IM}}(\mathcal{A}, \mathcal{B}) = \begin{pmatrix} \mathcal{A}^\circ \cap \mathcal{B}^\circ & \mathcal{A}^\circ \cap \partial\mathcal{B} \\ \partial\mathcal{A} \cap \mathcal{B}^\circ & \partial\mathcal{A} \cap \partial\mathcal{B} \end{pmatrix} \tag{11}$$

where $\mathcal{A}^\circ$ and $\partial\mathcal{A}$ are interior and boundary sets of $\mathcal{A}$. However, boundary-to-boundary and boundary-to-interior relation is difficult to predict from $\pi(\mathcal{P}_i)$ and $\mathcal{I}_j$. An *et al.* once used boundary-to-interior relationship for I2P registration [An *et al.*, 2024b], but they needed 3D object detector to understand 3D scene. Their method cannot be utilized in the general open scene. So, we simplify the 4IM model and only consider the intersection of two sets. It increases the prediction accuracy of topology reasoning in Eq. (6).

Then, we analyze why the proposed framework is robust to the open domain. Actually, topology relationship is a **scene-independent** feature, because it is only variant to the shapes relationships and robust to the unseen scenes. Thus, the topology relationship can be used as an important cue to the open-domain I2P registration.

3.3 Framework implementation

Sec. 3.2 provides an abstracted framework to leverage the topology relationship. To implement this framework in an actual scene, we propose a neural network Top-I2P with detailed computation procedures. Pipeline is provided in Fig. 4.

Topology graphs. We begin by discussing the construction of graphs $\mathcal{G}_P$ and $\mathcal{G}_I$. 2D and 3D backbones are used to extract features from $\mathcal{I}$ and $\mathcal{P}$. Following 2D3D-MATR [Li *et al.*, 2023], ResNet [He *et al.*, 2016] and KPConv [Thomas *et al.*, 2019] are used to construct 2D and 3D backbones. Backbone features of $\mathcal{I}$ and $\mathcal{P}$ are extracted as $\mathbf{F}_{\text{I}}^B$ and $\mathbf{F}_{\text{P}}^B$, respectively. In the meanwhile, using 2D and 3D SAM, $\mathcal{I}$ and $\mathcal{P}$ are segmented into non-overlapping regions, as shown in Eq. (4). For each point cloud patch $\mathcal{P}_i$, we compute the average coordinate $\bar{\mathbf{p}}_i$ and the average feature $\bar{\mathbf{f}}_i^{\text{3D}}$; For each image patch $\mathcal{I}_j$, we compute the average coordinate $\bar{\mathbf{q}}_j$ and average

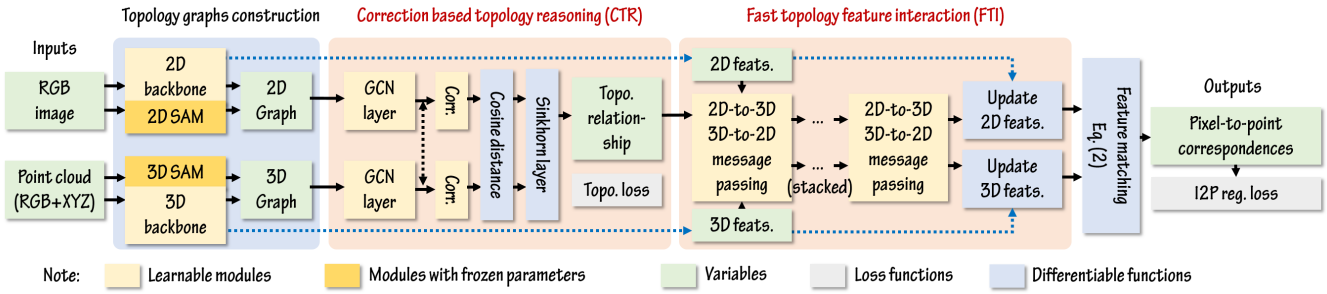

Figure 4: Top-I2P includes CTR and FTI modules, designed to leverage topology relationship for effective and efficient feature interaction.

feature $\bar{\mathbf{f}}_j^{2D}$. $\bar{\mathbf{f}}_i^{3D}$ $(i = 1,...,A)$ are vertices in $\mathcal{G}_P$, its adjacent matrix $\mathbf{A}_{3D}$ is constructed by $\exp(-\|\bar{\mathbf{p}}_i - \bar{\mathbf{p}}_j\|_2^2)$. In the same way, $\bar{\mathbf{f}}_j^{2D}$ $(j = 1,...,B)$ are vertices in $\mathcal{G}_I$, its adjacent matrix $\mathbf{A}_{2D}$ is constructed by $\exp(-\|\bar{\mathbf{q}}_i - \bar{\mathbf{q}}_j\|_2^2)$. Due to the setting of SAM, the patch number is no more than 128, so that the graph construction step can be done in real-time.

Correction-based topology reasoning (CTR). In the next, we illustrate the structure of $\phi_{\text{top}}(\mathcal{G}_I, \mathcal{G}_P)$. Due to the modality difference of $\mathcal{I}$ and $\mathcal{P}$, $\mathcal{G}_I$ and $\mathcal{G}_P$ are heterogeneous graphs, which brings the challenge of $\hat{\mathbf{A}}_{\text{top}}$ prediction. To deal with this issue, we design a simple yet effective CTR module in Fig. 4. Graph convolution network (GCN) layers [Kipf and Welling, 2017] are used to extract topology features $\mathbf{x} = \text{GCN}_{2d}(\mathcal{G}_I) \in \mathbb{R}^{B\times c}$ and $\mathbf{y} = \text{GCN}_{3d}(\mathcal{G}_P) \in \mathbb{R}^{A\times c}$. To eliminate the modality difference, we attempt to correct features by estimating feature shift $\Delta\mathbf{x}$ and $\Delta\mathbf{y}$:

$$\Delta\mathbf{x} = mlp_{2d}\left(\frac{\mathbf{x}^T\mathbf{y}}{\sqrt{c}}\mathbf{y}\right), \ \Delta\mathbf{y} = mlp_{3d}\left(\frac{\mathbf{y}^T\mathbf{x}}{\sqrt{c}}\mathbf{x}\right) \tag{12}$$

where mlp_{2d} and mlp_{3d} are two MLP layers. Then, topology features are updated as $\mathbf{x}' = \mathbf{x} + \Delta\mathbf{x}$ and $\mathbf{y}' = \mathbf{y} + \Delta\mathbf{y}$, respectively. $\hat{\mathbf{A}}_{\text{top}}$ is computed using a normalized cosine distance from $\mathbf{x}'$ to $\mathbf{y}'$. We find that most elements in the boolean matrix $\mathbf{A}_{\text{top}}$ are zero. So, we exploit a differentiable Sinkhorn layer [Sinkhorn, 1967] to regularize $\hat{\mathbf{A}}_{\text{top}}$. And loss in CTR module is designed as [Sarlin *et al.*, 2020]:

$$L_{\text{Topo}} = \frac{1}{N_{mask}}\|\mathbf{A}_{\text{top}}[mask] - \hat{\mathbf{A}}_{\text{top}}[mask]\|_2^2 \tag{13}$$

where $mask$ is the indices in $\mathbf{A}_{\text{top}}$ with ones, N_{mask} is the number of elements in $\mathbf{A}_{\text{top}}$ that is one.

Fast topology feature interaction (FTI). The topology feature interaction in Eqs. (7) and (9) cannot meet the fast I2P registration demand, although it greatly reduced the computation burden. The reason why Eqs. (7) and (9) are inefficient lies in pixel-wise and point-wise computation. To speed up this procedure, we design the FTI module. The core idea behind FTI is to replace $\mathbf{f}_q^{2D}$ and $\mathbf{f}_p^{3D}$ as $\bar{\mathbf{f}}_j^{2D}$ and $\bar{\mathbf{f}}_i^{3D}$ in Eqs. (7) and (9). Then, feature interaction is rewritten in a matrix operation manner with MLP layers:

$$\begin{pmatrix} \bar{\mathbf{f}}_{1,\text{new}}^{2D} \\ \vdots \\ \bar{\mathbf{f}}_{B,\text{new}}^{2D} \end{pmatrix} = mlp_a\left\{\alpha\begin{pmatrix} \bar{\mathbf{f}}_1^{2D} \\ \vdots \\ \bar{\mathbf{f}}_B^{2D} \end{pmatrix} + (1-\alpha)\hat{\mathbf{A}}_{\text{top,row}}^T\begin{pmatrix} \bar{\mathbf{f}}_1^{3D} \\ \vdots \\ \bar{\mathbf{f}}_A^{3D} \end{pmatrix}\right\} \tag{14}$$

$$\begin{pmatrix} \bar{\mathbf{f}}_{1,\text{new}}^{3D} \\ \vdots \\ \bar{\mathbf{f}}_{A,\text{new}}^{3D} \end{pmatrix} = mlp_b\left\{\alpha\begin{pmatrix} \bar{\mathbf{f}}_1^{3D} \\ \vdots \\ \bar{\mathbf{f}}_A^{3D} \end{pmatrix} + (1-\alpha)\hat{\mathbf{A}}_{\text{top,col}}\begin{pmatrix} \bar{\mathbf{f}}_1^{2D} \\ \vdots \\ \bar{\mathbf{f}}_B^{2D} \end{pmatrix}\right\} \tag{15}$$

where $\hat{\mathbf{A}}_{\text{top,col}}$ and $\hat{\mathbf{A}}_{\text{top,row}}$ are the column and row normalized matrices from $\hat{\mathbf{A}}_{\text{top}}$. It means that FTI focuses on interaction with average features. To improve the efficiency of feature interaction, FTR can be stacked in Top-I2P where the stack number is 3. After feature interaction, we update $\mathbf{f}_q^{2D}$ by adding shift $(\bar{\mathbf{f}}_{j,\text{new}}^{2D} - \bar{\mathbf{f}}_j^{2D})$, update $\mathbf{f}_p^{3D}$ by adding shift $(\bar{\mathbf{f}}_{i,\text{new}}^{3D} - \bar{\mathbf{f}}_i^{3D})$.

Training scheme. Single-stage training is unstable primarily due to the inaccurate predictions of topological relationships. To address this issue, we utilize a two-stage training scheme. First, we train backbone networks without CTR and FTI modules. The loss function is based on the I2P registration loss as defined in 2D3D-MATR [Li *et al.*, 2023]. Second, based on the pre-trained backbone networks, we begin to learn the topological relationship prediction task and train both the CTR and FTI modules. It is achieved by training the whole Top-I2P architecture by adding L_{Topo} loss defined in Eq. (13).

4 Experiments and Discussions

4.1 Configurations

Since limited research exists on open-domain I2P registration, we configure this task setting independently. First, we evaluate model generalization ability on the 7-Scenes dataset [Glocker *et al.*, 2013]. One scene is chosen for training, while the remaining six scenes are used for testing. Image and point cloud pairs are created from the RGB-D data. Then, we evaluate the generalization ability on the RGBD v2 [Lai *et al.*, 2014], ScanNet [Dai *et al.*, 2017], and self-collected datasets.

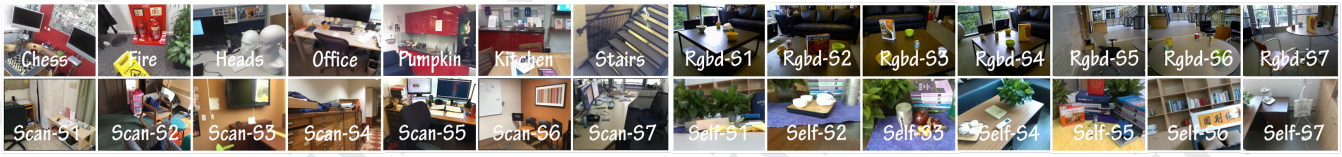

Figure 5: Examples scenes from the 7-Scenes, RGBD-V2, ScanNet, and self-collected (referred to as *RGBD*, *Scan*, *Self*) datasets.

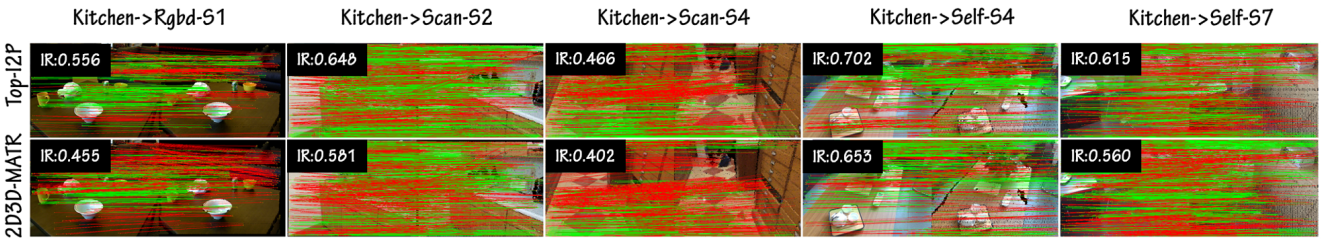

Figure 6: Qualitative comparison of Top-I2P and 2D3D-MATR on unseen scenes. Top-I2P estimates inliers more than 2D3D-MATR.

Metric-IR	Chess→Chess	Chess→Fire	Chess→Heads	Chess→Office	Chess→Pumpkin	Chess→Kitchen	Chess→Stairs	Average†
P2-Net	0.516	0.436	0.330	0.414	0.421	0.405	0.251	0.376
2D3D-MATR	0.761	0.455	0.359	0.420	0.411	0.390	0.288	0.387
Top-I2P	**0.767**	**0.491**	**0.427**	**0.455**	**0.461**	**0.437**	**0.327**	**0.433**
Metric-RR	Chess→Chess	Chess→Fire	Chess→Heads	Chess→Office	Chess→Pumpkin	Chess→Kitchen	Chess→Stairs	Average†
P2-Net	0.875	0.536	0.162	0.672	0.561	0.563	0.293	0.464
2D3D-MATR	**1.000**	0.537	0.167	0.759	0.581	0.612	0.214	0.478
Top-I2P	**1.000**	**0.611**	**0.583**	**0.843**	**0.558**	**0.678**	**0.500**	**0.628**
Metric-IR	Office→Office	Office→Chess	Office→Fire	Office→Heads	Office→Pumpkin	Office→Kitchen	Office→Stairs	Average†
P2-Net	0.506	0.416	0.413	0.403	0.434	0.386	0.308	0.393
2D3D-MATR	0.645	0.498	0.491	0.521	0.442	0.448	0.338	0.456
Top-I2P	**0.667**	**0.512**	**0.494**	**0.532**	**0.466**	**0.462**	**0.353**	**0.469**
Metric-RR	Office→Office	Office→Chess	Office→Fire	Office→Heads	Office→Pumpkin	Office→Kitchen	Office→Stairs	Average†
P2-Net	0.769	0.566	0.661	0.232	0.577	0.532	0.234	0.510
2D3D-MATR	**0.940**	0.660	0.556	0.417	0.395	0.636	0.286	0.491
Top-I2P	0.880	**0.702**	**0.759**	**0.423**	**0.605**	**0.645**	**0.293**	**0.571**
Metric-IR	Kitchen→Kitchen	Kitchen→Chess	Kitchen→Fire	Kitchen→Office	Kitchen→Heads	Kitchen→Pumpkin	Kitchen→Stairs	Average†
P2-Net	0.678	0.516	0.512	0.504	0.506	0.555	0.358	0.491
2D3D-MATR	**0.717**	0.571	0.594	0.537	0.538	0.612	0.370	0.537
Top-I2P	0.697	**0.626**	**0.619**	**0.627**	**0.631**	**0.643**	**0.433**	**0.596**
Metric-RR	Kitchen→Kitchen	Kitchen→Chess	Kitchen→Fire	Kitchen→Office	Kitchen→Heads	Kitchen→Pumpkin	Kitchen→Stairs	Average†
P2-Net	0.851	0.857	0.583	0.250	0.769	0.611	0.429	0.621
2D3D-MATR	0.901	0.872	0.778	0.667	0.723	0.698	0.500	0.706
Top-I2P	**0.926**	**0.936**	**0.792**	**0.722**	**0.831**	**0.860**	**0.571**	**0.785**

Table 1: Open-domain I2P registration performance on the 7-Scenes datasets. Here † represents the average metrics on the unseen scenes.

The self-collected dataset was collected using an Intel RealSense D351 depth camera. We obtained the GT pose between images and point clouds based on an RGB-D-based 3D reconstruction toolbox[2]. In this case, all compared methods are trained on the 7-Scenes dataset [Glocker *et al.*, 2013]. From Figs. 1 and 5, scenes in the different datasets are largely distinct, making them suitable for open-domain evaluation. Point clouds include RGB features.

To evaluate I2P registration performance, IR and RR are used as the main metrics. The root mean square error (RMSE) threshold is set to 0.025m and 0.1m for the strict I2P registration performance testing, while other metric thresholds follow the defaults in the literature [Li *et al.*, 2023]. Subsequently, we discuss the details of Top-I2P. α is set to 0.86. The stack number of the message passing submodule in FTI is 2. The first training stage consists of 5 epochs, while the second stage consists of 20 epochs. To generate the ground truth (GT) $\mathbf{A}_{\text{top}}$, we set an intersection-over-union (IoU) threshold $\gamma_{IoU} = 0.1$. $\mathbf{A}_{\text{top},ij}$ is set as 1 only if $IoU(\pi(\mathcal{P}_i) \cap \mathcal{I}_j) \geq \gamma_{IoU}$, and this procedure filters out several unimportant topology relations. The learning rate and optimizer of Top-I2P are the same in the literature [Li *et al.*, 2023].

[2]www.open3d.org

4.2 Comparison results

Current I2P registration methods, P2-Net [Wang *et al.*, 2021] and 2D3D-MATR [Li *et al.*, 2023], are used for comparison with Top-I2P. The notation $A \rightarrow B$ indicates that the model is trained on scene A but tested on scene B. Results on the 7-Scenes dataset [Glocker *et al.*, 2013] are provided in Table 1, and results on RGBD-V2 [Lai *et al.*, 2014], ScanNet [Dai

Metric-IR	Kitchen→Rgbd-S1	Kitchen→Rgbd-S2	Kitchen→Rgbd-S3	Kitchen→Rgbd-S4	Kitchen→Rgbd-S5	Kitchen→Rgbd-S6	Kitchen→Rgbd-S7	Average
P2-Net	0.090	0.259	0.268	0.263	0.098	0.106	0.094	0.168
2D3D-MATR	0.351	0.353	0.336	0.316	0.250	0.209	0.222	0.291
Top-I2P	**0.397**	**0.413**	**0.389**	**0.398**	**0.252**	**0.247**	**0.253**	**0.335**
Metric-RR@0.1	Kitchen→Rgbd-S1	Kitchen→Rgbd-S2	Kitchen→Rgbd-S3	Kitchen→Rgbd-S4	Kitchen→Rgbd-S5	Kitchen→Rgbd-S6	Kitchen→Rgbd-S7	Average
P2-Net	0.234	0.861	0.845	0.921	0.280	0.232	0.167	0.505
2D3D-MATR	0.970	**0.880**	0.871	0.741	0.480	0.449	0.458	0.692
Top-I2P	**0.974**	0.840	**0.986**	**0.963**	**0.600**	**0.710**	**0.479**	**0.793**
Metric-IR	Kitchen→Scan-S1	Kitchen→Scan-S2	Kitchen→Scan-S3	Kitchen→Scan-S4	Kitchen→Scan-S5	Kitchen→Scan-S6	Kitchen→Scan-S7	Average
P2-Net	0.359	0.413	0.286	0.190	0.370	0.312	0.195	0.303
2D3D-MATR	0.495	0.550	0.424	0.337	0.507	0.434	0.414	0.451
Top-I2P	**0.550**	**0.587**	**0.472**	**0.340**	**0.547**	**0.485**	**0.475**	**0.493**
Metric-RR@0.05	Kitchen→Scan-S1	Kitchen→Scan-S2	Kitchen→Scan-S3	Kitchen→Scan-S4	Kitchen→Scan-S5	Kitchen→Scan-S6	Kitchen→Scan-S7	Average
P2-Net	0.778	0.750	0.846	0.273	0.893	0.929	0.510	0.711
2D3D-MATR	**0.956**	**0.954**	0.974	0.433	0.923	0.909	0.750	0.842
Top-I2P	0.889	0.902	**0.976**	**0.620**	**0.962**	**0.981**	**0.976**	**0.901**
Metric-IR	Kitchen→Self-S1	Kitchen→Self-S2	Kitchen→Self-S3	Kitchen→Self-S4	Kitchen→Self-S5	Kitchen→Self-S6	Kitchen→Self-S7	Average
P2-Net	0.336	0.299	0.270	0.429	0.304	0.443	0.235	0.331
2D3D-MATR	0.497	0.462	0.426	**0.618**	**0.507**	**0.619**	0.412	0.506
Top-I2P	**0.506**	**0.499**	**0.476**	0.617	0.473	0.615	**0.459**	**0.521**
Metric-RR@0.05	Kitchen→Self-S1	Kitchen→Self-S2	Kitchen→Self-S3	Kitchen→Self-S4	Kitchen→Self-S5	Kitchen→Self-S6	Kitchen→Self-S7	Average
P2-Net	0.333	0.224	0.056	0.833	0.222	0.942	0.052	0.380
2D3D-MATR	0.556	0.389	0.333	**0.976**	0.532	**0.964**	0.278	0.575
Top-I2P	**0.722**	**0.444**	**0.452**	0.944	**0.611**	0.952	**0.333**	**0.636**

Table 2: Open-domain I2P registration performance across multiple datasets, including RGBD-V2, ScanNet, and self-collected datasets.

et al., 2017], and self-collected datasets are provided in Table 2.

First, P2-Net suffers from significant overfitting to the training scene due to the absence of feature interaction modules. This limitation causes its backbone networks to generalize poorly to unseen scenes. Second, while 2D3D-MATR incorporates a feature interaction module, it operates only on patch-level features, limiting its ability to learn fine-grained pixel-level cross-modality interactions. This results in 2D3D-MATR being superior to P2-Net while weaker than Top-I2P. Third, Top-I2P achieves a more accurate feature interaction through its topology-based approach, which results in a higher number of inliers than 2D3D-MATR across different datasets (Fig. 6). The RR of Top-I2P is consistently and significantly better than that of 2D3D-MATR on all datasets. It suggests that topology relationship are helpful in improving the localization quality of pixel-to-point correspondences.

Furthermore, we compare the proposed Top-I2P with LCD [Pham *et al.*, 2020], SuperGlue (Glue) [Sarlin *et al.*, 2020], and FreeReg [Wang *et al.*, 2024] on ScanNet [Dai *et al.*, 2017]. The results of these methods are obtained from FreeReg [Wang *et al.*, 2024], where the RMSE threshold is 0.3m. The results are provided in Table 3. Even with a stricter RMSE threshold, Top-I2P achieves an RR 12.1% higher than FreeReg [Wang *et al.*, 2024]. It means that the generalization ability of Top-I2P is higher than that of FreeReg. Additionally, we compare Top-I2P with FCGF [Choy *et al.*, 2019], FreeReg [Wang *et al.*, 2024], and Diff-Reg [Wu *et al.*, 2024b] on the RGBD-V2 dataset [Lai *et al.*, 2014], where all models are trained using the train and test splits in literature [Li *et al.*, 2023]. The results are provided in Table 4, and they show that Top-I2P outperforms state-of-the-art methods. Therefore, the above experiments suggest that Top-I2P achieves the best performance in the open-domain I2P registration task.

4.3 Ablation studies

In this section, we study the effect of the CTR and FTI modules in Top-I2P. First, the ablation study of the CTR module is

Methods	P2-Net†	2D3D-MATR†	LCD	Glue	FreeReg	Top-I2P†
IR	0.303	0.451	0.307	0.184	**0.568**	0.493
RR@0.3	0.711	0.842	N/A	0.065	0.780	**0.901**

Table 3: Comparison results on the ScanNet dataset (RMSE threshold is 0.3m). † indicates that the model was trained on the Kitchen scene with the RMSE threshold of 0.05m, **stricter** than 0.3m.

Methods	P2-Net	2D3D-MATR	FCGF	FreeReg	Diff-Reg	Top-I2P
IR	0.122	0.324	0.081	0.309	N/A	**0.427**
RR@0.1	0.384	0.564	0.304	0.573	0.874	**0.892**

Table 4: Comparison results of current methods on the RGBD-v2 dataset, evaluated with an RMSE threshold of 0.1m.

provided in Table 5. $[m, n]$ denotes two MLP neural network layers with m and n channels. *Empty* indicates that the model is the baseline [Li *et al.*, 2023] if true. Sinkhorn refers to the usage of the Sinkhorn layer [Sinkhorn, 1967]. From Table 5, the Sinkhorn layer plays a dominant role in the CTR module, because it determines the accuracy of topology relationship prediction. This accuracy directly impacts the overall effectiveness of the CTR module. Thus, the performance of CTR depends on the prediction accuracy of $\pi(\mathcal{P}_i) \cap \mathcal{I}_j$.

Second, the ablation study of the FTI module is shown in Table 6. N_S is the stack number of the message passing submodule in FTI, as shown in Fig. 4. We fix N_S and search for the optimal α in the range $[0.80, 0.88]$. The performance is relatively insensitive to α, as the FTI module primarily interacts with average features. We keep the optimal α and search for the best N_S. When the stack number N_S is excessively large, the average features in different patches tend to be similar, which reduces the discriminative ability of cross-modality features. The best α and N_S are 0.86 and 2, respectively.

Third, we investigate the performance of the combination of the CTR and FTI modules. Results are provided in Table 7. Removing CTR (and replacing it with MLPs) causes a significant drop in topology prediction accuracy (Pred. Acc.), leading to degraded IR and RR. Similarly, without FTI, the

Empty	GCN	MLPs	Sinkhorn	IR	RR	Pred. Acc.
Yes	×	×	×	0.376	0.464	N/A
×	[128, 128]	×	×	0.382	0.522	0.354
×	[128, 256]	[256, 128]	×	0.388	0.535	0.375
×	[128, 256]	[256, 128]	Yes	**0.433**	**0.628**	**0.773**

Table 5: Ablation study of the proposed CTR module in Top-I2P on the 7-Scene dataset, where the Chess scene was used for training. Pred. Acc is the accuracy of the predicted topology relationship.

α	0.80	0.82	0.84	0.86	0.88	N_S	1	2	3
IR	0.454	0.455	0.458	**0.469**	0.462	**IR**	0.467	**0.469**	0.465
RR	0.552	0.561	0.566	**0.571**	0.570	**RR**	0.558	**0.571**	0.566

Table 6: Ablation study of hyper-parameters α and N_s of the FTI module on the 7-Scene dataset, where the Office is the training scene.

Methods	Top-I2P (w/o CTR)	Top-I2P (w/o FTI)	Top-I2P
IR	0.392	0.402	**0.433**
RR	0.487	0.558	**0.628**
Pred. Acc.	0.513	0.752	**0.773**

Table 7: Ablation study of the combination of CTR and FTI on the 7-Scene dataset. Pred. Acc. denotes the accuracy of predicted topology relationship.

model struggles to exploit topological cues for cross-modality learning. These results validate the necessity of both CTR and FTI.

Fourth, we evaluate the inference time and GPU memory usage performance of Top-I2P. Actually, there are two schemes to accelerate Top-I2P. First, we can skip 3D SAM during model inference, because 3D point cloud segmentation can be pre-computed in some downstream tasks related to I2P registration (i.e., visual localization). Second, we can use a lightweight 2D SAM during model inference. It can further reduce the runtime and memory usage. The results of the accelerated Top-I2P are shown in Table 8. It is found that the frame per second (FPS) of MATR [Feng *et al.*, 2019] and the proposed Top-I2P are nearly the same.

Besides, we also analyze the failure case of the proposed I2P registration method. Although Top-I2P achieves the best I2P registration performance in the open-domain scenarios, Top-I2P still faces difficulties when input images lack texture or the point clouds have low resolution, as shown in Fig. 7. In such cases, even though the IR is higher than 2D3D-MATR, the 2D-3D correspondences become sparse. Thus, in future work, we attempt to refine the architecture to achieve accurate and real-time I2P registration performance.

5 Conclusions

In this paper, we addressed the challenge of open-domain I2P registration. At first, inspired by the potential connection between topology relationship and cross-modality feature interaction, we developed an I2P registration framework using topology relationship. After that, to efficiently construct and leverage topology relationship from heterogeneous 2D and 3D spaces, we designed Top-I2P with CTR and FTI modules. Extensive experiments on multiple datasets have demonstrated the superior performance of Top-I2P.

Methods	Inference time/ms	FPS	Memory usage/MB
MATR	127	**7.8**	3114
Top-I2P	45(SAM)+132=177	**5.6**	2432(SAM)+3520=5952

Table 8: Inference time and memory usage of the accelerated Top-I2P.

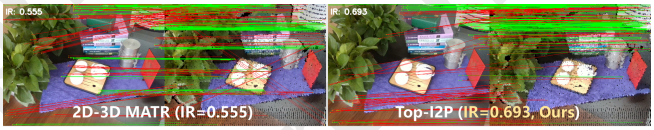

Figure 7: Failure case of Top-I2P in the self-collected dataset.

Limitation and future work. Top-I2P does not fully utilize other topology relationship, such as $\mathcal{A}^\circ \cap \partial\mathcal{B}$ and $\partial\mathcal{A} \cap \partial\mathcal{B}$ in Eq. (11). We plan to refine this framework in the future.

Acknowledgments

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

Contribution Statement

Pei An and Jiaqi Yang have equal contribution in this work. You Yang is the corresponding author in this work.

References

[An *et al.*, 2024a] Pei An, Junfeng Ding, Siwen Quan, Jiaqi Yang, You Yang, Qiong Liu, and Jie Ma. 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, 2024.

[An *et al.*, 2024b] Pei An, Xuzhong Hu, Junfeng Ding, Jun Zhang, Jie Ma, You Yang, and Qiong Liu. 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, 2024.

[Chen *et al.*, 2022] Chaoqi Chen, Luyao Tang, Feng Liu, and et al. Mix and reason: Reasoning over semantic topology with data mixing for domain generalization. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1–15, 2022.

[Cheng *et al.*, 2023a] Keyang Cheng, Qing Liu, Rabia Tahir, Liangmin Wang, and Maozhen Li. Logical topology inference via CPGCN joint optimizing with pedestrian re-id. *IEEE Trans. Neural Networks Learn. Syst.*, 34(8):5099–5111, 2023.

[Cheng *et al.*, 2023b] Yuxin Cheng, Zhiqiang Huang, Siwen Quan, Xinyue Cao, Shikun Zhang, and Jiaqi Yang. Sampling locally, hypothesis globally: accurate 3d point cloud registration with a ransac variant. *Visual Intelligence*, 20:1–15, 2023.

[Choy *et al.*, 2019] Christopher B. Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 8957–8965, 2019.

[Dai *et al.*, 2017] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2432–2443, 2017.

[Dusmanu *et al.*, 2019] Mihai Dusmanu, Ignacio Rocco, Tomás Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable CNN for joint description and detection of local features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 8092–8101, 2019.

[Egenhofer, 1993] Max J. Egenhofer. A model for detailed binary topological relationships. *Geoinformatica*, 47:261–273, 1993.

[Feng *et al.*, 2019] Mengdan Feng, Sixing Hu, Marcelo H. Ang, and Gim Hee Lee. 2D3D-Matchnet: Learning to match keypoints across 2D image and 3D point cloud. In *Proceedings of IEEE International Conference on Robotics and Automation*, pages 4790–4796, 2019.

[Glocker *et al.*, 2013] Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. Real-time RGB-D camera relocalization. In *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, pages 173–179, 2013.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Kim *et al.*, 2023] Minjung Kim, Junseo Koo, and Gunhee Kim. 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*, pages 21470–21480, 2023.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of 5th International Conference on Learning Representations*, pages 1–14, 2017.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, and et al. Segment anything. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 3992–4003, 2023.

[Lai *et al.*, 2014] Kevin Lai, Liefeng Bo, and Dieter Fox. Unsupervised feature learning for 3d scene labeling. In *Proceedings of IEEE International Conference on Robotics and Automation*, pages 3050–3057, 2014.

[Li and Lee, 2021] Jiaxin Li and Gim Hee Lee. DeepI2P: Image-to-point cloud registration via deep classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 15960–15969, 2021.

[Li *et al.*, 2023] Minhao Li, Zheng Qin, Zhirui Gao, Renjiao Yi, Chenyang Zhu, Yulan Guo, and Kai Xu. 2D3D-MATR: 2D-3D matching transformer for detection-free registration between images and point clouds. In *Proceedings of IEEE Conference on Computer Vision*, pages 1–10, 2023.

[Li *et al.*, 2024] Jiangmeng Li, Yifan Jin, Hang Gao, Wenwen Qiang, Changwen Zheng, and Fuchun Sun. Hierarchical topology isomorphism expertise embedded graph contrastive learning. In *Proceedings of Thirty-Eighth AAAI Conference on Artificial Intelligence*, pages 13518–13527, 2024.

[Pham *et al.*, 2020] Quang-Hieu Pham, Mikaela Angelina Uy, Binh-Son Hua, Duc Thanh Nguyen, Gemma Roig, and Sai-Kit Yeung. LCD: learned cross-domain descriptors for 2d-3d matching. In *Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 11856–11864, 2020.

[Qin *et al.*, 2022] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11133–11142, 2022.

[Ren *et al.*, 2023] Siyu Ren, Yiming Zeng, Junhui Hou, and Xiaodong Chen. CorrI2P: Deep image-to-point cloud registration via dense correspondence. *IEEE Trans. Circuits Syst. Video Technol.*, 33(3):1198–1208, 2023.

[Sarlin *et al.*, 2020] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4937–4946, 2020.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[Sinkhorn, 1967] R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *Am. Math. Mon.*, 74(4):402–405, 1967.

[Sun *et al.*, 2020] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6397–6406, 2020.

[Thomas *et al.*, 2019] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 6410–6419, 2019.

[van Dijk *et al.*, 2024] Tom van Dijk, Christophe De Wagter, and Guido C. H. E. de Croon. Visual route following for tiny autonomous robots. *Sci. Robotics*, 9(92), 2024.

[Wang *et al.*, 2021] Bing Wang, Changhao Chen, and et al. P2-Net: Joint description and detection of local features for pixel and point matching. In *Proceedings of IEEE International Conference on Computer Vision*, pages 15984–15993, 2021.

[Wang *et al.*, 2023] Huijie Wang, Tianyu Li, Yang Li, and et al. Openlane-v2: A topology reasoning benchmark for unified 3d HD mapping. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1–15, 2023.

[Wang *et al.*, 2024] Haiping Wang, Yuan Liu, Bing Wang, Yujing Sun, Zhen Dong, Wenping Wang, and Bisheng Yang. Freereg: Image-to-point cloud registration leveraging pretrained diffusion models and monocular depth estimators. In *Proceedings of International Conference on Learning Representation*, pages 1–24, 2024.

[Wu *et al.*, 2021] Bingli Wu, Jie Ma, Gaojie Chen, and Pei An. Feature interactive representation for point cloud registration. In *Proceedings IEEE/CVF International Conference on Computer Vision*, pages 5510–5519, 2021.

[Wu *et al.*, 2024a] Dongming Wu, Jiahao Chang, Fan Jia, Yingfei Liu, Tiancai Wang, and Jianbing Shen. Topomlp: A simple yet strong pipeline for driving topology reasoning. In *Proceedings of The Twelfth International Conference on Learning Representations*, pages 1–18, 2024.

[Wu *et al.*, 2024b] Qianliang Wu, Haobo Jiang, Lei Luo, Jun Li, Yaqing Ding, Jin Xie, and Jian Yang. Diff-reg: Diffusion model in doubly stochastic matrix space for registration problem. In *Proceedings of IEEE/CVF European Conference on Computer Vision*, volume 15123, pages 160–178, 2024.

[Zhang *et al.*, 2024] Juexiao Zhang, Gao Zhu, Sihang Li, and et al. Multiview scene graph. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1–15, 2024.

[Zhou *et al.*, 2023] Junsheng Zhou, Baorui Ma, Wenyuan Zhang, Yi Fang, Yu-Shen Liu, and Zhizhong Han. Differentiable registration of images and lidar point clouds with voxelpoint-to-pixel matching. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1–10, 2023.

[Zhou *et al.*, 2024] Yuchen Zhou, Jiayuan Gu, Tung Yen Chiang, Fanbo Xiang, and Hao Su. Point-sam: Promptable 3d segmentation model for point clouds. In *Proceedings of IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2024.