This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Hg-I2P: Bridging Modalities for Generalizable Image-to-Point-Cloud Registration via Heterogeneous Graphs

Pei An[1], Junfeng Ding[1], Jiaqi Yang[2*], Yulong Wang[3], Jie Ma[1], Liangliang Nan[4*]
[1]Huazhong University of Science and Technology, China
[2]Northwestern Polytechnical University, China
[3]Huazhong Agricultural University, China [4]Delft University of Technology, Netherlands

Abstract

Image-to-point-cloud (I2P) registration aims to align 2D images with 3D point clouds by establishing reliable 2D-3D correspondences. The drastic modality gap between images and point clouds makes it challenging to learn features that are both discriminative and generalizable, leading to severe performance drops in unseen scenarios.

We address this challenge by introducing a heterogeneous graph that enables refining both cross-modal features and correspondences within a unified architecture. The proposed graph represents a mapping between segmented 2D and 3D regions, which enhances cross-modal feature interaction and thus improves feature discriminability. In addition, modeling the consistency among vertices and edges within the graph enables pruning of unreliable correspondences. Building on these insights, we propose a heterogeneous graph embedded I2P registration method, termed Hg-I2P. It learns a heterogeneous graph by mining multi-path feature relationships, adapts features under the guidance of heterogeneous edges, and prunes correspondences using graph-based projection consistency. Experiments on six indoor and outdoor benchmarks under cross-domain setups demonstrate that Hg-I2P significantly outperforms existing methods in both generalization and accuracy. Code is released on https://github.com/anpei96/hg-i2p-demo.

1. Introduction

Image-to-point-cloud (I2P) registration aims to establish pixel-to-point correspondences between an image and a 3D point cloud [11, 23], serving as a cornerstone for tasks such as visual localization [17], navigation [38], and 3D reconstruction [7, 12, 26]. Recent learning-based methods approach this task by extracting features for both modalities and identifying correspondences through feature matching [3, 11, 23, 35]. While progress has been made through improved backbones [36], matching strategies [5, 37], and loss functions [24], these methods still struggle to generalize to unseen environments.

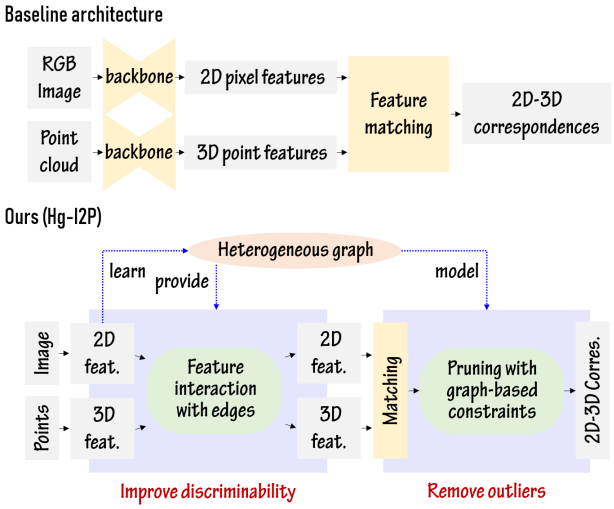


Figure 1. **Motivation of Hg-I2P**. To achieve generalizable I2P registration, we reformulate the baseline architecture using a heterogeneous graph. Its heterogeneous edges enhance cross-modal feature interaction, improving feature discriminability. Projection constraints within the graph enable consistency-based pruning of mismatched correspondences, enhancing robustness. The resulting framework, Hg-I2P, jointly refines both features and matches, achieving strong generalization across unseen scenes.

The main reason for this limitation lies in the significant modality gap between images and point clouds. The appearance-driven 2D features and geometry-driven 3D features follow fundamentally different distributions. Consequently, even the correct correspondences may exhibit low feature similarity, making it difficult for neural networks to distinguish true correspondences [9]. Existing

*Corresponding authors: Jiaqi Yang (jqyang@nwpu.edu.cn) and Liangliang Nan (liangliang.nan@tudelft.nl)

methods attempt to alleviate this problem either by refining cross-model features [2] or pruning incorrect correspondences [5], yet each only tackles one aspect in isolation. Furthermore, feature refinement methods [2, 5, 9] often lack explicit cross-modal reasoning, while correspondence pruning typically relies on depth prediction or handcrafted heuristics. This fragmented treatment leaves the generalization problem largely unsolved.

To address these issues, we argue that a unified structure capable of jointly reasoning 2D and 3D relationships is essential for robust and generalizable registration. To this end, we introduce a heterogeneous graph for I2P data. It captures a mapping between 2D and 3D regions, thereby strengthening cross-modal feature interaction and enriching representations with both geometric and visual cues. The consistency of vertices and edges within the graph provides a principled means to prune incorrect correspondences.

Building on these insights, we develop a **H**eterogeneous **g**raph embedded **I2P** registration method (**Hg-I2P**). It incorporates three core components: **(1) MP-mining**, which learns heterogeneous edges by mining multi-path adjacency relations on a graph; **(2) HG-adapting**, which refines I2P features by coarse-to-fine cross-modal message passing along heterogeneous edges; and **(3) HC-pruning**, which enforces graph-aware projection consistency to effectively filter out incorrect correspondences. Together, these components enable a joint refinement of both features and correspondences on the graph, significantly enhancing generalization across domains. Our contributions are as follows:

- We identify **a unified perspective for generalizable I2P** registration, where a heterogeneous graph simultaneously refines both cross-modal features and correspondences.
- We propose Hg-I2P, **a heterogeneous graph-embedded registration architecture** that incorporates graph learning, feature adaptation, and correspondence pruning.
- We conduct **extensive cross-domain experiments** on six indoor and outdoor datasets, demonstrating strong generalization and state-of-the-art performance.

2. Related Works

We first review the architectures of mainstream learning-based I2P registration methods. Then, we discuss the representative works that aim to improve the generalization ability of I2P registration.

Architectures of learning-based I2P registration. The design of I2P registration architectures usually draws inspiration from learning-based image or point cloud registration methods [22, 23]. In 2019, Feng et al. [11] learned a modality-invariant descriptor for both 2D and 3D pre-detected keypoints. But, pre-detected keypoints often contain few inliers and negatively affect learning-based I2P registration performance [22]. To avoid this issue, researchers began to jointly learn both keypoints and descriptors, leading to progress in I2P architecture design. Wang et al. [35] introduced a mechanism to learn keypoints by incorporating spatial-wise and channel-wise maxima on I2P feature maps. Li et al. [22] and Ren et al. [29] proposed cross-modal global attention modules for more discriminative I2P feature learning. Building on this, Zhou et al. [47] refined loss functions by embedding adaptive-weight optimization into the circle loss [32], dynamically adjusting correspondence weights. They also incorporated differentiable perspective-n-point (PnP) [8] to further suppress outliers. Some researchers designed I2P registration models inspired by point cloud registration architectures [28], such as MATR [23], CoFiI2P [16], and PAPI-Reg [43]. These methods predict 2D-3D correspondences in a coarse-to-fine manner (i.e., first matching 2D-3D patches, then generating correspondences guided by the matched regions), thereby improving accuracy on public benchmarks.

Improving generalization in I2P registration. As more researchers have observed unstable I2P performance in unseen environments [1, 2, 9, 36], recent works have focused on improving feature generalization across domains. Vision foundation models, such as DepthAnything [39], Segment Anything Model (SAM) [18], and object detectors [42] are increasingly leveraged to bridge modality gaps in I2P data. For example, Wang et al. [36] used a pre-trained ControlNet [44] and depth prediction model to realign cross-modal features, though their approach requires over $\geq$ 10 seconds per image. An et al. [1] employed 3D object detectors to align objects between images and point clouds, designing Proj-I2P to extract correspondences from aligned object pairs. Wu et al. [37] utilized the DINO model [6] to obtain spatially-aware RGB features, while Bie et al. [5] leveraged DepthAnything [39] to convert I2P registration into a 3D-3D point cloud registration task. Because the predicted depths lack real-world scale, they proposed a distribution alignment module to minimize the difference between LiDAR and back-projected point clouds. Murai et al. [26] applied the multi-view reconstruction model MASt3R [21] to generate point clouds from images and then used a gradient-based ICP scheme in $\mathbf{Sim}(3)$ space to estimate camera poses. An et al. [2] employed both the 2D and 3D SAM models [18, 48] to segment images and point clouds into regions. They predict topological relationships between these segments to assist cross-modal alignment.

Some works improve I2P feature generalization without relying on vision foundation models. Cheng et al. [9] modeled the matching uncertainty between 2D and 3D patches. Yao et al. [41] formulated correspondence generation as a Markov decision process and developed a reinforcement learning based agent for I2P registration. Li et al. [24] proposed an implicit correspondence learning module to predict 2D and 3D keypoints jointly.

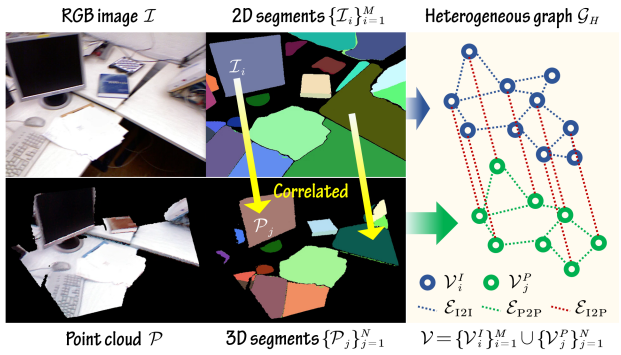


Figure 2. Constructing a heterogeneous graph ($\mathcal{G}_H$) for cross-modal reasoning. This graph models the relationship between 2D and 3D regions. By explicitly linking visual and geometric entities through heterogeneous edges, it captures structured 2D-3D dependencies essential for joint reasoning.

Discussions. To improve I2P registration generalization, leveraging vision foundation models has become a mainstream strategy, as these models possess strong generalization capabilities in unseen domains [25]. In this paper, we utilize 2D and 3D SAM models to construct a heterogeneous graph as a bridge for cross-modal feature matching. Compared with the prior work [2] that also employed SAM, our method has three key advantages: (1) We define a heterogeneous graph that systematically models relationships between 2D and 3D regions; (2) We exploit both edge information and projection constraints in this graph to refine both I2P features and correspondences; (3) Hg-I2P achieves stronger generalization and higher accuracy than state-of-the-art methods [2, 3, 5, 9, 23] across multiple datasets.

3. Heterogeneous Graph Representation

Before illustrating the effect of the heterogeneous graph on I2P registration, we define this structure formally. Given an RGB image and a colored point cloud, we use 2D and 3D SAM [18, 40, 48] to decompose the image and point cloud into multiple 2D and 3D regions, and then define a heterogeneous weighted graph $\mathcal{G}_H = (\mathcal{V}_H, \mathcal{E}_H)$ to systematically model relationships between these regions (see Fig. 2).

Definition. In $\mathcal{G}_H$, the vertices and edges are defined as $\mathcal{V}_H = \mathcal{V}_I \cup \mathcal{V}_P = \{\mathcal{V}_i^I\}_{i=1}^M \cup \{\mathcal{V}_j^P\}_{j=1}^N$ and $\mathcal{E}_H = \mathcal{E}_{\text{I2I}} \cup \mathcal{E}_{\text{P2P}} \cup \mathcal{E}_{\text{I2P}}$. We assign each vertex $\mathcal{V}_i^I$ (or $\mathcal{V}_j^P$) with a c-dimensional feature vector $\mathbf{v}_i^I \in \mathbb{R}^c$ (or $\mathbf{v}_j^P$). In practice, we use 2D and 3D SAM [18, 48] to segment $\mathcal{I}$ and $\mathcal{P}$ into non-overlapping regions $\{\mathcal{I}_i\}_{i=1}^M$ and $\{\mathcal{P}_j\}_{j=1}^N$:

$$\mathcal{I} = \cup_i^M \mathcal{I}_i, \ \mathcal{P} = \cup_j^N \mathcal{P}_j, \tag{1}$$

where each region $\mathcal{I}_i$ (or $\mathcal{P}_j$) corresponds to a vertex $\mathcal{V}_i^I$ (or $\mathcal{V}_j^P$). In $\mathcal{G}_H$, $\mathcal{E}_{\text{I2I}}$ (or $\mathcal{E}_{\text{P2P}}$) are **homogeneous** edge set describing intra-modal relationship between $\mathcal{I}_i$ and $\mathcal{I}_j$ (or $\mathcal{P}_i$ and $\mathcal{P}_j$). $\mathcal{E}_{\text{I2P}}$ is the **heterogeneous** edge set describing cross-modal relationships between $\{\mathcal{I}_i\}_{i=1}^M$ and $\{\mathcal{P}_j\}_{j=1}^N$. $\mathcal{E}_{\text{I2I}}$, $\mathcal{E}_{\text{P2P}}$, and $\mathcal{E}_{\text{I2P}}$ are represented by three adjacent matrices $\mathbf{E}_{\text{I2I}} \in \mathbb{R}^{M\times M}$, $\mathbf{E}_{\text{P2P}} \in \mathbb{R}^{N\times N}$, and $\mathbf{E}_{\text{I2P}} \in \mathbb{R}^{M\times N}$, respectively. As $\mathcal{G}_H$ is a weighted graph, the homogeneous edges can be defined based on feature distance:

$$(\mathbf{E}_{\text{I2I}})_{ij} = e^{-\alpha\|\mathbf{v}_i^I - \mathbf{v}_j^I\|_2^2}, (\mathbf{E}_{\text{P2P}})_{ij} = e^{-\alpha\|\mathbf{v}_i^P - \mathbf{v}_j^P\|_2^2} \tag{2}$$

where α is a distance scaling factor, and $(\cdot)_{ij}$ queries a value from a matrix at index (i, j).

Heterogeneous edges. In general, $\mathcal{E}_{\text{I2P}}$ reflects the correlation of 2D-3D regions. Extending from work [2], we model this correlation through the overlap between a 2D region $\mathcal{I}_i$ and the projection of a 3D region $\mathcal{V}_j$, quantified using the interaction-over-union (IoU). Thus,

$$(\mathbf{E}_{\text{I2P}})_{ij} = \text{IoU}(\mathcal{I}_i, \pi(\mathcal{P}_j|\mathbf{T}_{gt})) \tag{3}$$

where $\pi(\cdot|\mathbf{T})$ denotes the camera projection under a pose $\mathbf{T} \in \mathbf{SE}(3)$ [45], and $\mathbf{T}_{gt}$ is the ground truth (GT) pose. If no overlap exists (i.e., $\mathcal{I}_i \cap \pi(\mathcal{P}_j|\mathbf{T}_{gt}) = \emptyset$), $(\mathbf{E}_{\text{I2P}})_{ij} = 0$, making $\mathbf{E}_{\text{I2P}}$ a sparse matrix. This construction provides a **coarse mapping between 2D and 3D regions**, enabling the refinement of both I2P features and correspondences.

Initialization. Using 2D, 3D backbone networks [15, 33], we extract feature maps $\mathbf{F}_I \in \mathbb{R}^{H\times W\times c}$ and $\mathbf{F}_P \in \mathbb{R}^{A\times c}$ from the image $\mathcal{I}$ and point cloud $\mathcal{P}$. Here, H and W represent the image height and width, respectively. A is the points number, and c is feature dimension. In practice, we set the vertex feature $\mathbf{v}_i^I$ (or $\mathbf{v}_j^P$) as $\bar{\mathbf{f}}_i^I$ (or $\bar{\mathbf{f}}_j^P$), which is average pooled from $\mathbf{F}_I$ (or $\mathbf{F}_P$) on region $\mathcal{I}_i$ (or $\mathcal{P}_j$). More detail is shown in the supplemental materials. Since $\mathbf{T}_{gt}$ is unavailable in I2P registration, $\mathcal{E}_{\text{I2P}}$ cannot be directly computed. We learn $\mathcal{E}_{\text{I2P}}$ in the Hg-I2P architecture.

4. Proposed Method

4.1. Overview of Hg-I2P

To effectively utilize the heterogeneous graph for generalized I2P registration, we propose a complete pipeline called Hg-I2P. As shown in Fig. 3, it consists of three core components: (1) **Heterogeneous edges learning**, which predicts $\mathcal{E}_{\text{I2P}}$ and construct $\mathcal{G}_H$; (2) **Feature adaptation on graph**, which refines the cross-modal features under the guidance of $\mathcal{E}_{\text{I2P}}$; (3) **Match refinement with graph**, which enforces projection consistency in $\mathcal{G}_H$ to prune outliers. Overall, by leveraging $\mathcal{G}_H$, the proposed feature adaptation and match refinement steps enhance both feature discriminability and correspondence accuracy, thereby improving the generalization capability of I2P registration.

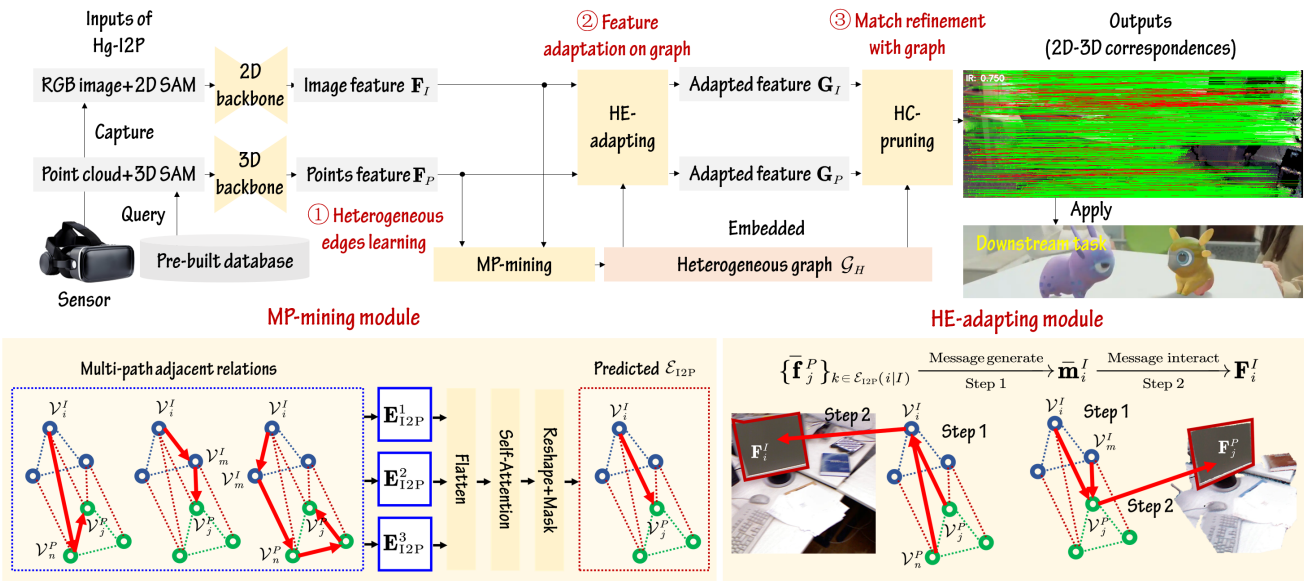


Figure 3. **Pipeline of Hg-I2P**. MP-mining learns heterogeneous edges $\mathcal{E}_{\text{I2P}}$ and constructs the graph $\mathcal{G}_H$. HE-adapting refines cross-modal features via message passing along heterogeneous edges, improving cross-modal alignment. HC-pruning enforces projection consistency within $\mathcal{G}_H$ to remove outliers. Together, these components refine both features and correspondences for robust I2P registration.

4.2. Learning heterogeneous edges with multi-path adjacent relations mining

To construct $\mathcal{G}_H$, accurately learning $\mathcal{E}_{\text{I2P}}$ from I2P data is a crucial step. To fully exploit $\mathcal{E}_{\text{I2I}}$ and $\mathcal{E}_{\text{P2P}}$ for this purpose, we **mine** **m**ulti-**p**ath (MP) adjacent relations and introduce an **MP-mining** module to predict $\mathcal{E}_{\text{I2I}}$.

To incorporate $\mathcal{E}_{\text{I2I}}$ and $\mathcal{E}_{\text{P2P}}$, we model the adjacency between $\langle \mathcal{I}_i, \mathcal{P}_j \rangle$ by traversing multiple paths from $\mathcal{V}_i^I$ to $\mathcal{V}_j^P$. For example, path $\mathcal{V}_i^I \rightarrow \mathcal{V}_m^I \rightarrow \mathcal{V}_j^P$ is one feasible route on $\mathcal{G}_H$ $(m = 1, ..., M)$. In this case, the adjacency relation is modeled as $(\mathbf{E}_{\text{I2P}})_{ij} \leftarrow \sum_{m=1}^{M} (\mathbf{E}_{\text{I2I}})_{im} (\mathbf{E}_{\text{I2P}})_{mj}$. Similarly, other paths such as $\mathcal{V}_i^I \rightarrow \mathcal{V}_n^P \rightarrow \mathcal{V}_j^P$ and $\mathcal{V}_i^I \rightarrow \mathcal{V}_m^I \rightarrow \mathcal{V}_n^P \rightarrow \mathcal{V}_j^P$ are also considered. We initialize the heterogeneous edges as $\tilde{\mathbf{E}}_{\text{I2P}}$, where $(\tilde{\mathbf{E}}_{\text{I2P}})_{ij} = e^{-\alpha \|\mathbf{v}_i^I - \mathbf{v}_j^P\|_2^2}$ and model the adjacency relations as three matrices: $\mathbf{E}_{\text{I2P}}^1 = \mathbf{E}_{\text{I2I}} \tilde{\mathbf{E}}_{\text{I2P}}$, $\mathbf{E}_{\text{I2P}}^2 = \tilde{\mathbf{E}}_{\text{I2P}} \mathbf{E}_{\text{P2P}}$, $\mathbf{E}_{\text{I2P}}^3 = \mathbf{E}_{\text{I2I}} \tilde{\mathbf{E}}_{\text{I2P}} \mathbf{E}_{\text{P2P}}$.

From a **Bayesian inference** perspective[1], $\{\mathbf{E}_{\text{I2P}}^k\}_{k=1}^3$ capture the significant casual relations between $\mathcal{I}_i$ and $\mathcal{P}_j$. Based on this insight, MP-mining predicts $\mathbf{E}_{\text{I2P}}$ by mining correlation features from these matrices:

$$\begin{aligned} \hat{\mathbf{E}}_{\text{I2P}} &= \text{RS}(\text{Attention}(\mathbf{F}_{\text{I2P}} \mathbf{W}_{\text{MP}}^Q, \mathbf{F}_{\text{I2P}} \mathbf{W}_{\text{MP}}^K, \mathbf{F}_{\text{I2P}} \mathbf{W}_{\text{MP}}^V)) \\ \mathbf{F}_{\text{I2P}} &= \mathbf{W}_{\text{MP}}^F \cdot \text{Flatten}([\mathbf{E}_{\text{I2P}}^1, \mathbf{E}_{\text{I2P}}^2, \mathbf{E}_{\text{I2P}}^3]) \end{aligned} \tag{4}$$

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\mathbf{Q}\mathbf{K}^T / \sqrt{d_k})\mathbf{V} \tag{5}$$

where $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ is an attention layer [34] and $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V} \in \mathbb{R}^{d_k}$ denote the query, key, and value vectors, respectively. $\text{RS}(\cdot)$ is the reshape operation that converts an $M \times N$-dimensional vector into an $M \times N$ matrix. $\mathbf{W}_{\text{MP}}^F \in \mathbb{R}^{3MN \times MN}$, $\mathbf{W}_{\text{MP}}^Q$, $\mathbf{W}_{\text{MP}}^K$, and $\mathbf{W}_{\text{MP}}^V \in \mathbb{R}^{MN \times MN}$ are the learnable matrices. $\text{Flatten}(\cdot)$ is an operation to flatten a tensor as vector. In practice, $\hat{\mathbf{E}}_{\text{I2P}}$ is $L1$-normalized row-wise before used. Together with $\mathcal{E}_{\text{I2I}}$, $\mathcal{E}_{\text{P2P}}$, and the learned $\mathcal{E}_{\text{I2P}}$, the final heterogeneous graph $\mathcal{G}_H$ can be constructed.

4.3. Heterogeneous edges guided feature adaptation

As discussed in Sec. 3, $\mathcal{E}_{\text{I2P}}$ defines a mapping between 2D and 3D segmented regions. To leverage this mapping for cross-modal feature learning, we propose a **heterogeneous edge** guided **adaptation** module, termed **HE-adapting**.

Given $\{\mathcal{I}_i\}_{i=1}^M$ and $\{\mathcal{P}_j\}_{j=1}^N$, we decompose $\mathbf{F}_I$ and $\mathbf{F}_P$ into feature groups $\{\mathbf{F}_i^I\}_{i=1}^M$ and $\{\mathbf{F}_j^P\}_{j=1}^N$[2]. Using $\mathcal{E}_{\text{I2P}}$, region $\mathcal{I}_i$ (or $\mathcal{P}_j$) is associated with its neighboring regions $\{\mathcal{P}_k\}_{k \in \mathcal{E}_{\text{I2P}}(i|I)}$ (or $\{\mathcal{I}_k\}_{k \in \mathcal{E}_{\text{I2P}}(j|P)}$). We adopt a lightweight *message generation-interaction* scheme for feature refinement:

$$\{\bar{\mathbf{f}}_k^P\}_{k \in \mathcal{E}_{\text{I2P}}(i|I)} \xrightarrow[\text{Step 1}]{\text{Message generate}} \bar{\mathbf{m}}_i^I \xrightarrow[\text{Step 2}]{\text{Message interact}} \mathbf{F}_i^I$$

$$\{\bar{\mathbf{f}}_k^I\}_{k \in \mathcal{E}_{\text{I2P}}(j|P)} \xrightarrow[\text{Step 1}]{\text{Message generate}} \bar{\mathbf{m}}_j^P \xrightarrow[\text{Step 2}]{\text{Message interact}} \mathbf{F}_j^P$$

In this process, cross-modal message feature $\bar{\mathbf{m}}_i^I$ (or $\bar{\mathbf{m}}_j^P$)

[1] As $(\mathbf{E}_{\text{I2I}})_{ij} \in (0, 1)$, it can be regarded as a conditional probability $P(\mathcal{V}_i^I | \mathcal{V}_j^P)$. This way, $\mathbf{E}_{\text{I2P}}^3$ (as an example) can be explained as: $(\mathbf{E}_{\text{I2P}}^3)_{ij}$ is $P(\mathcal{V}_i^I | \mathcal{V}_j^P) = \sum_m \sum_n P(\mathcal{V}_i^I | \mathcal{V}_m^I) P(\mathcal{V}_m^I | \mathcal{V}_n^P) P(\mathcal{V}_n^P | \mathcal{V}_j^P)$.

[2] $\mathbf{F}_i^I \in \mathbb{R}^{m_i \times c}$ and $\mathbf{F}_j^P \in \mathbb{R}^{n_j \times c}$. The m_i (or n_j) denotes the number of pixels (or points) in the region $\mathcal{I}_i$ (or $\mathcal{P}_j$).

is first derived from vertex features $\{\bar{\mathbf{f}}_k^P\}_{k\in\mathcal{E}_{\text{I2P}}(i|I)}$ (or $\{\bar{\mathbf{f}}_k^I\}_{k\in\mathcal{E}_{\text{I2P}}(j|P)}$). Subsequently, each region feature $\mathbf{F}_i^I$ (or $\mathbf{F}_j^P$) is refined through interaction with the corresponding message feature $\bar{\mathbf{m}}_i^I$ (or $\bar{\mathbf{m}}_j^P$). The procedure is detailed below:

Step 1: Message generation. The message $\bar{\mathbf{m}}_i^I$ (or $\bar{\mathbf{m}}_j^P$) is encoded from $\{\bar{\mathbf{f}}_j^P\}_{k\in\mathcal{E}_{\text{I2P}}(i|I)}$ (or $\{\bar{\mathbf{f}}_i^I\}_{k\in\mathcal{E}_{\text{I2P}}(j|P)}$):

$$\begin{aligned}\bar{\mathbf{m}}_i^I &= \text{Attention}(\bar{\mathbf{f}}_i^I\mathbf{W}_v^I, \bar{\mathbf{g}}_i^P\mathbf{W}_v^P, \bar{\mathbf{g}}_i^P\mathbf{W}_v^P) \in \mathbb{R}^c,\\ \bar{\mathbf{m}}_j^P &= \text{Attention}(\bar{\mathbf{f}}_j^P\mathbf{W}_v^P, \bar{\mathbf{g}}_j^I\mathbf{W}_v^I, \bar{\mathbf{g}}_j^I\mathbf{W}_v^I) \in \mathbb{R}^c\end{aligned} \tag{6}$$

$$\bar{\mathbf{g}}_i^P = \frac{\sum_{j=1}^N(\hat{\mathbf{E}}_{\text{I2P}})_{ij}\bar{\mathbf{f}}_j^P}{\sum_{j=1}^N(\hat{\mathbf{E}}_{\text{I2P}})_{ij}},\ \bar{\mathbf{g}}_j^I = \frac{\sum_{i=1}^M(\hat{\mathbf{E}}_{\text{I2P}})_{ij}\bar{\mathbf{f}}_i^I}{\sum_{i=1}^M(\hat{\mathbf{E}}_{\text{I2P}})_{ij}} \tag{7}$$

where $\bar{\mathbf{g}}_i^P$ (or $\bar{\mathbf{g}}_j^I$) denotes the average-pooled feature from $\mathcal{V}_P$ (or $\mathcal{V}_I$) that shares edges with $\mathcal{V}_i^I$ (or $\mathcal{V}_j^P$). Through the cross-attention in Eq. 6, the message feature $\bar{\mathbf{m}}_i^I$ is obtained by learning the correlation between $\bar{\mathbf{g}}_i^P$ and $\bar{\mathbf{f}}_i^I$. In Eq. 6, $\mathbf{W}_v^I$ and $\mathbf{W}_v^P \in \mathbb{R}^{c\times c}$ are learnable matrices.

Step 2: Message interaction. The encoded messages $\{\bar{\mathbf{m}}_i^I\}_{i=1}^M$ (or $\{\bar{\mathbf{m}}_j^P\}_{j=1}^N$) are used to adapt the region-wise features $\{\mathbf{F}_i^I\}_{i=1}^M$ (or $\{\mathbf{F}_j^P\}_{j=1}^N$):

$$\begin{aligned}\mathbf{G}_i^I &= (1-\beta)\mathbf{F}_i^I\\ &\quad + \beta\cdot\text{Attention}(\mathbf{H}_i^I\mathbf{W}_I^Q, \mathbf{H}_i^I\mathbf{W}_I^K, \mathbf{H}_i^I\mathbf{W}_I^V),\\ \mathbf{G}_j^P &= (1-\beta)\mathbf{F}_j^P\\ &\quad + \beta\cdot\text{Attention}(\mathbf{H}_j^P\mathbf{W}_P^Q, \mathbf{H}_j^P\mathbf{W}_P^K, \mathbf{H}_j^P\mathbf{W}_P^V)\end{aligned} \tag{8}$$

$$\begin{aligned}\mathbf{H}_i^I &= \text{Concat}(\mathbf{F}_i^I, \text{Repeat}(\bar{\mathbf{m}}_i^I)) \in \mathbb{R}^{m_i\times 2c},\\ \mathbf{H}_j^P &= \text{Concat}(\mathbf{F}_j^P, \text{Repeat}(\bar{\mathbf{m}}_j^P)) \in \mathbb{R}^{n_j\times 2c}\end{aligned} \tag{9}$$

where $\beta \in [0,1]$ is a hyperparameter balancing the original and adapted features. $\mathbf{W}_I^Q$, $\mathbf{W}_I^K$, $\mathbf{W}_I^V$, $\mathbf{W}_P^Q$, $\mathbf{W}_P^K$, and $\mathbf{W}_P^V \in \mathbb{R}^{2c\times c}$ are learnable matrices. $\text{Repeat}(\cdot)$ replicates a tensor. $\text{Concat}(\cdot,\cdot)$ concatenates feature vectors along the channel dimension. Through the self-attention in Eq. 8, HE-adapting captures the relations between each pixel (or point) in $\mathcal{I}_i$ (or $\mathcal{P}_j$) and its corresponding message feature $\bar{\mathbf{m}}_i^I$ (or $\bar{\mathbf{m}}_j^P$), yielding the adapted features $\mathbf{G}_i^I$ and $\mathbf{G}_j^P$. The resulting features $\{\mathbf{G}_i^I\}_{i=1}^M$ and $\{\mathbf{G}_j^P\}_{j=1}^N$ are concatenated to form the refined representations $\mathbf{G}_I \in \mathbb{R}^{H\times W\times c}$ and $\mathbf{G}_P \in \mathbb{R}^{A\times c}$, respectively.

Finally, the refined feature sets $\mathbf{G}_I$ and $\mathbf{G}_P$ are utilized for feature-level matching [23] to produce 2D-3D correspondences $\{\langle\mathbf{p}_i^c, \mathbf{P}_i^c\rangle\}_{i=1}^C$, where C denotes the number of matches.

4.4. Match refinement with heterogeneous graph based pruning criteria

Removing outliers from $\{\langle\mathbf{p}_i^c, \mathbf{P}_i^c\rangle\}_{i=1}^C$ is crucial for robust I2P registration. To achieve this, we propose a non-learning module called **HC-pruning**, which refines correspondences based on **heterogeneous** graph-based **pruning criteria**, as shown in Fig. 4.

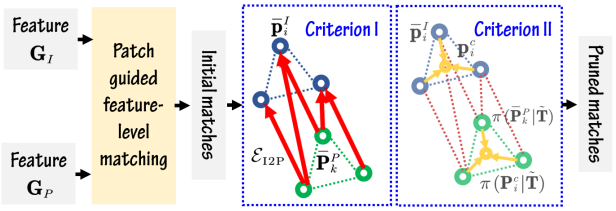


Figure 4. Overview of HC-pruning. It filters incorrect correspondences using two projection consistency criteria derived from $\mathcal{G}_H$.

In HC-pruning, an initial camera pose $\tilde{\mathbf{T}} \in \mathbf{SE}(3)$ is estimated via a RANSAC-based perspective-n-point (PnP) algorithm [20], which leverages the strong shape correlation between 2D and 3D segments [2]. Then, with $\tilde{\mathbf{T}}$, two pruning criteria are applied: (1) based on the adjacency in $\mathcal{E}_{\text{I2P}}$ and reprojection distance δ_{rej}, and (2) based on the cosine similarity between relative position vectors derived from graph-based projections. A correspondence is identified as an inlier if at least one of the criteria is satisfied. The details of these criteria are elaborated in the supplementary materials. This dual-criterion approach effectively removes false matches caused by pose estimation noise or imperfect edge learning.

4.5. Loss functions

To train Hg-I2P, we define the overall loss as:

$$L_{\text{Hg-I2P}} = L_{\text{corr}} + \lambda_1\|\hat{\mathbf{E}}_{\text{I2P}}[\text{mask}] - \mathbf{E}_{\text{I2P}}[\text{mask}]\|_2^2 \tag{10}$$

where L_{corr} is the standard correspondence loss used in I2P registration [23], which contains the circle loss [32] to penalize the correspondences generated from $\mathbf{G}_I$ and $\mathbf{G}_P$ by a patch guided feature-level matching. The second term supervises the learning of $\mathcal{E}_{\text{I2P}}$. The mask denotes the indices of valid (non-zero) elements in $\mathbf{E}_{\text{I2P}}$, and λ_1 is a weighting factor controlling edge supervision strength. This formulation ensures that both cross-modal alignment and edge prediction are jointly optimized, leading to consistent improvements in generalization and registration robustness.

5. Experiments and Discussions

5.1. Configurations

To evaluate the robustness and generalization capability of existing I2P registration methods, we conduct experiments primarily in a **cross-domain** configuration, which includes both cross-scene and cross-dataset evaluations.

Datasets information. Six datasets are employed, including five indoor datasets (i.e., 7-Scene [14], RGBD-V2 [19],

Table 1. Cross-scene I2P registration results on the 7-Scenes dataset [14]. † denotes the Hg-I2P variant without HC-pruning.

IR	C→F	C→H	C→O	C→P	C→K	C→S	AVG
MATR	0.455	0.359	0.420	0.411	0.390	0.288	0.387
Top-I2P	0.491	0.427	0.455	0.461	0.437	0.327	0.433
MinCD	0.542	0.424	0.502	0.408	0.416	0.379	0.445
Hg-I2P†	0.547	0.467	0.492	0.520	0.475	0.332	0.472
Hg-I2P (Ours)	**0.592**	**0.588**	**0.744**	**0.533**	**0.557**	**0.469**	**0.581**
RR	C→F	C→H	C→O	C→P	C→K	C→S	AVG
MATR	0.537	0.167	0.759	0.581	0.612	0.214	0.478
Top-I2P	0.611	0.583	0.843	0.558	**0.678**	0.500	0.628
MinCD	0.671	0.250	0.869	0.574	0.619	0.571	0.592
Hg-I2P†	0.712	**0.585**	0.717	0.617	0.655	**0.572**	0.642
Hg-I2P (Ours)	**0.726**	0.500	**0.949**	**0.681**	0.631	0.512	**0.667**
IR	O→C	O→F	O→H	O→P	O→K	O→S	AVG
MATR	0.498	0.491	0.521	0.442	0.448	0.338	0.456
Top-I2P	0.512	0.494	0.532	0.466	0.462	0.353	0.469
MinCD	**0.660**	**0.642**	0.536	**0.550**	0.546	**0.471**	**0.568**
Hg-I2P†	0.526	0.521	0.537	0.486	0.491	0.408	0.494
Hg-I2P (Ours)	0.601	0.599	**0.585**	0.544	**0.556**	0.465	0.558
RR	O→C	O→F	O→H	O→P	O→K	O→S	AVG
MATR	0.660	0.556	0.417	0.395	0.636	0.286	0.491
Top-I2P	0.702	0.759	0.423	0.605	0.645	0.293	0.571
MinCD	0.769	0.726	0.250	0.681	**0.810**	**0.643**	0.647
Hg-I2P†	0.738	0.630	**0.750**	**0.682**	0.667	0.501	0.661
Hg-I2P (Ours)	**0.846**	**0.781**	0.583	0.638	0.768	0.523	**0.690**
IR	K→C	K→F	K→O	K→H	K→P	K→S	AVG
P2-Net	0.516	0.512	0.504	0.506	0.555	0.358	0.491
MATR	0.571	0.594	0.537	0.538	0.612	0.370	0.537
Top-I2P	0.626	0.619	0.627	0.631	0.643	0.433	0.596
MinCD	0.617	0.598	0.540	0.573	0.636	0.445	0.568
Hg-I2P†	0.645	0.677	0.651	0.602	0.686	0.448	0.618
Hg-I2P (Ours)	**0.728**	**0.755**	**0.705**	**0.675**	**0.753**	**0.511**	**0.688**
RR	K→C	K→F	K→O	K→H	K→P	K→S	AVG
MATR	0.872	0.778	0.667	0.723	0.698	0.500	0.706
Top-I2P	0.936	0.792	0.722	0.831	0.860	0.571	0.785
MinCD	0.846	**0.904**	0.683	0.798	**0.872**	0.786	0.814
Hg-I2P†	0.923	0.849	0.750	0.859	0.787	0.643	0.802
Hg-I2P (Ours)	**0.969**	0.890	**0.833**	**0.869**	0.702	**0.857**	**0.853**

Table 2. Cross-dataset I2P registration performance across multiple datasets, including RGBD-V2 [19], ScanNet [10], and a self-collected dataset.

IR	RG-1	RG-2	RG-3	RG-4	RG-5	RG-6	RG-7	AVG
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
MinCD	0.427	0.415	0.405	0.412	0.310	**0.296**	**0.329**	0.371
Hg-I2P (Ours)	**0.503**	**0.499**	**0.444**	**0.456**	**0.325**	0.285	0.314	**0.404**
RR	RG-1	RG-2	RG-3	RG-4	RG-5	RG-6	RG-7	AVG
MATR	0.970	0.880	0.871	0.741	0.480	0.449	0.458	0.692
Top-I2P	0.973	0.840	**0.986**	0.962	0.600	0.710	0.479	0.793
MinCD	**0.974**	**0.985**	0.968	**0.963**	0.707	0.725	0.711	**0.870**
Hg-I2P (Ours)	0.966	0.846	0.972	0.926	**0.840**	**0.754**	**0.771**	0.868
IR	SC-1	SC-2	SC-3	SC-4	SC-5	SC-6	SC-7	AVG
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
MinCD	0.517	0.527	0.460	0.343	0.548	0.456	0.428	0.468
Hg-I2P (Ours)	**0.626**	**0.691**	**0.580**	**0.453**	**0.640**	**0.569**	**0.541**	**0.586**
RR	SC-1	SC-2	SC-3	SC-4	SC-5	SC-6	SC-7	AVG
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
MinCD	0.987	**0.979**	0.905	**0.720**	0.962	0.915	0.821	0.898
Hg-I2P (Ours)	**0.996**	0.944	**0.986**	0.705	**0.968**	0.912	0.944	**0.922**
IR	SE-1	SE-2	SE-3	SE-4	SE-5	SE-6	SE-7	AVG
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
MinCD	0.485	0.470	0.437	0.581	0.522	0.604	0.514	0.516
Hg-I2P (Ours)	**0.702**	**0.615**	**0.649**	0.460	**0.730**	0.366	**0.589**	**0.573**
RR	SE-1	SE-2	SE-3	SE-4	SE-5	SE-6	SE-7	AVG
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
MinCD	0.512	0.405	0.389	**0.984**	0.611	0.952	0.389	0.606
Hg-I2P (Ours)	**0.889**	**0.778**	**0.772**	0.833	**0.778**	0.871	**0.556**	**0.775**

Table 3. Comparisons of I2P registration methods on the ScanNet dataset [10]. † denotes evaluation with an RR threshold of 30 cm, while other methods use 5 cm.

Methods	Venue	IR	RR
LCD†	AAAI 2020	0.307	N/A
Glue†	CVPR 2020	0.183	0.065
P2-Net	ICCV 2021	0.303	0.711
MATR	ICCV 2023	0.451	0.842
FreeReg†	ICLR 2024	0.568	0.780
Top-I2P	IJCAI 2025	0.493	0.901
MinCD	ICCV 2025	0.464	0.898
Hg-I2P (Ours)		**0.586**	**0.922**

ScanNet [10], a self-collected dataset, and TUM [31]), as well as one outdoor dataset, KITTI [13].

As I2P registration is mainly used for fine-grained visual localization within a pre-built point cloud map [17, 35], each image maintains a high degree of overlap with the corresponding submap (i.e., a segmented portion of the map determined by a coarse camera pose). Thus, to simulate the realistic applications, we generate I2P samples with high overlaps. The 3D point clouds are down-sampled using a voxel resolution of 1.5 cm, with a maximum of 3.0×10^5 points. RGB images are resized to 640×480 resolution. Since KITTI images are originally 1242×375, we crop and pad them to 640×480 and adjust the camera intrinsic parameters accordingly. As LiDAR scans are typically sparser than depth camera point clouds, we generate LiDAR point clouds through the multi-scan registration. Further details of the self-collected dataset are provided in the supplemental materials. All point clouds in our experiments include RGB features.

Baseline methods and evaluation metrics. Top-I2P [2] and MinCD [3] represent the current state-of-the-art methods for I2P registration in unseen scenes (more discussion is shown in the supplemental materials). Hg-I2P is compared against these methods under a cross-domain evaluation setting, where all models are trained in one scene and tested on the unseen scenes. Besides, Hg-I2P is also compared with more existing methods [5, 9, 24, 35, 36] in the standard evaluation setting. During the evaluation, inlier ratio (IR) and registration recall (RR) are main metrics [23]. In addition, Average relative translational error (RTE) and average relative rotation error (RRE) [16, 29] are used for further analysis.

Implementation details. Before registration, the pre-built map is segmented using SAM3D [40], while FastSAM [46] is used for image segmentation during inference because of its computational efficiency. Using GT camera poses, we align 2D and 3D segments via Eq. 3 to generate the GT het-

Table 4. Performance comparison on the TUM dataset [31] (RR threshold is 10 cm). Hg-I2P consistently achieves the highest IR and RR, and lowest RTE and RRE across all sequences.

Freiburg_01	IR	RR	RTE/m	RRE/deg
MATR	0.319	0.595	0.037	1.238
Top-I2P	0.306	0.655	0.036	1.179
Hg-I2P (Ours)	**0.380**	**0.681**	**0.035**	**1.147**
Freiburg_02	IR	RR	RTE/m	RRE/deg
MATR	0.223	0.223	0.052	1.469
Top-I2P	0.234	0.310	0.054	1.502
Hg-I2P (Ours)	**0.286**	**0.394**	**0.050**	**1.455**
Freiburg_03	IR	RR	RTE/m	RRE/deg
MATR	0.335	0.548	0.051	1.310
Top-I2P	0.314	0.603	0.050	1.294
Hg-I2P (Ours)	**0.393**	**0.666**	**0.047**	**1.198**

Table 5. Comparisons on the KITTI dataset [13]. † indicates that LiDAR point clouds are constructed from multiple scans, a common setup in visual localization using pre-built maps.

Methods	Venue	RTE/m	RRE/deg
CorrI2P	TCSVT 2022	0.740	2.070
VP2P-match	NeurIPS 2023	0.750	3.290
CoFiI2P	RAL 2024	0.290	1.140
CMR-Agent	IROS 2024	0.195	0.589
GraphI2P	CVPR 2025	0.320	1.650
ImCorr	CVPR 2025	0.200	1.240
Hg-I2P (Ours)		0.172	0.502
Hg-I2P† (Ours)		**0.160**	**0.485**

erogeneous edges $\mathbf{E}_{\text{I2P}}$. The backbone networks in Hg-I2P are adapted from from MATR [23], and we adopt the same learning rate and optimizer as Top-I2P [2]. Training is conducted for 60 epochs to ensure accurate heterogeneous edge learning. The main hyperparameters in Hg-I2P (i.e., α, δ_{rej}, τ_{rej}, and λ_1) are discussed in the supplemental materials.

5.2. Comparisons

We conduct cross-domain I2P registration evaluations, including both **cross-scene** and **cross-dataset** comparisons.

Cross-scene comparisons. This experiment is conducted on the 7-Scenes dataset [14], which includes the scenes Chess (C), Fire (F), Heads (H), Office (O), Pumpkin (P), Kitchen (K), and Stairs (S). For simplicity, we use C→F to indicate that the model is trained in scene C and evaluated on scene F. Table 1 presents the comparison results with RR threshold as 2.5 cm. Even without the HC-pruning module, the proposed Hg-I2P outperforms MATR [23], Top-I2P [2], and MinCD [3] on both IR and RR metrics across most scenes. Hg-I2P (w/o HC-pruning) shows strong performance because accurate heterogeneous edge prediction enhances cross-modal feature interaction, thereby improving feature discriminability. When HC-pruning is applied, IR and RR increase further, as unreliable correspondences are effectively removed by proposed pruning criteria. These results collectively confirm the effectiveness of Hg-I2P.

Cross-dataset comparisons without fine-tuning. This

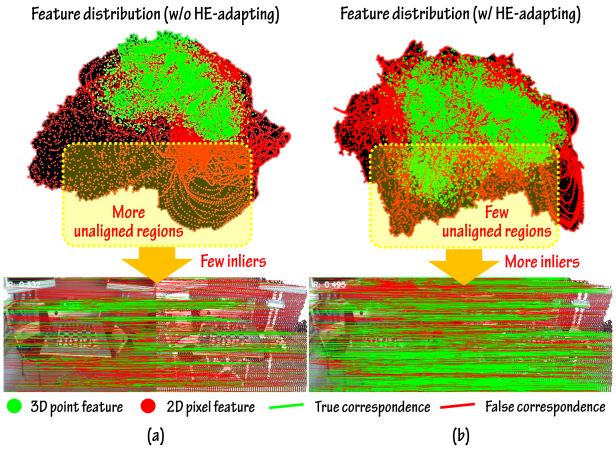


Figure 5. Effect of the proposed HE-adapting module. (a) Baseline without HE-adapting. (b) With HE-adapting, cross-modal features are more accurately aligned, producing a higher number of inliers.

Table 6. Comparisons on the RGBD-V2 dataset [19]. † denotes methods requiring depth input from RGB-D camera.

Methods	Venue	IR	RR
P2-Net	ICCV 2021	0.122	0.384
MATR	ICCV 2023	0.324	0.564
FreeReg	ICLR 2024	0.309	0.573
DiffReg†	ECCV 2024	0.378	**0.874**
Bridge	AAAI 2025	0.351	0.634
DiffI2P	ICCV 2025	0.369	0.605
FlowI2P	IJCV 2025	0.401	0.684
Hg-I2P (Ours)		**0.428**	0.690

evaluation is performed on multiple datasets, including RGBD-V2 [19], ScanNet [10], self-collected, and TUM [31] datasets. Methods are trained on scene K of the 7-Scenes dataset [14] and directly tested on these unseen datasets without fine-tuning. Notations RG-X, SC-X, and SE-X denote the X-th scene from RGBD-V2 [19], ScanNet [10], and the self-collected dataset, respectively. Table 2 summarizes the results (RR threshold is 10 cm for RGBD-V2 and 5 cm for ScanNet and self-collected datasets). Hg-I2P achieves stable and high RR scores across all datasets. Additional comparisons with LCD [27], Glue [30], and FreeReg [36] on ScanNet (Table 3) further demonstrate Hg-I2P's robustness. Results on the TUM dataset (Table 4) show that Hg-I2P achieves RTE below 5 cm, outperforming prior methods [2, 23]. Overall, Hg-I2P consistently delivers accurate registration in unseen domains without any fine-tuning.

Cross-dataset comparisons with fine-tuning. This experiment is conducted on the KITTI dataset [13]. Since LiDAR point clouds exhibit different density and range characteristics from depth-camera point clouds, Hg-I2P is fine-tuned on KITTI. As shown in Table 5, Hg-I2P outperforms existing methods, including CorrI2P [29], VP2P-match [47],

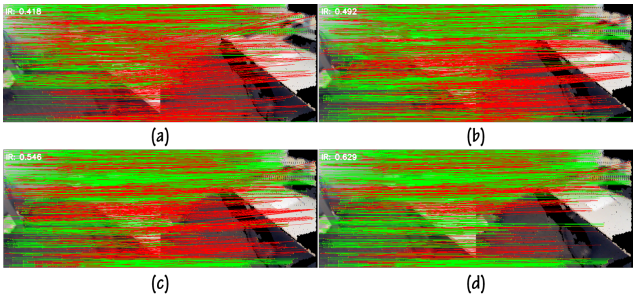


Figure 6. Qualitative comparison of I2P registration results. (a) MATR [23]. (b) Top-I2P [2]. (c) Hg-I2P without HC-pruning. (d) Full Hg-I2P which produces fewer outliers than prior methods.

Table 7. Ablation study of the MP-mining module. $\mathbf{E}^i_{\text{I2P}}$ denotes the inclusion of the i-th feature heterogeneous edge prediction.

$\mathbf{E}^1_{\text{I2P}}$	$\mathbf{E}^2_{\text{I2P}}$	$\mathbf{E}^3_{\text{I2P}}$	Precision	IR	RR
✓			0.826	0.478	0.610
✓	✓		0.874	0.490	0.634
✓	✓	✓	0.893 (**+6.7%**)	0.494 (**+1.6%**)	0.661 (**+5.1%**)

Table 8. Ablation study of the HE-adapting module. Parameter β controls the strength of cross-modal feature interaction through heterogeneous edges.

Parameter β in Eq. (8)	0.0	0.05	0.10	0.15	0.20	0.25
IR	0.462	0.478	**0.494**	0.482	0.454	0.437
RR	0.620	0.648	**0.661**	0.612	0.603	0.582

CoFiI2P [16], CMR-Agent [41], GraphI2P [5], and ImCorr [24], on both RTE and RRE metrics. These results confirm that the proposed MP-mining, HE-adapting, and HC-pruning modules are highly compatible with outdoor scenes and diverse sensor configurations.

Benchmark comparisons. We further evaluate Hg-I2P on the RGBD-V2 dataset [19] following the standard training and testing split [23] (RR threshold is 10 cm). As shown in Table 6, Hg-I2P achieves superior IR and RR performance compared with state-of-the-art works [4, 9, 25, 36, 37]. This demonstrates that Hg-I2P effectively reduces the cross-modality gap between 2D images and 3D point clouds more efficiently than existing methods.

5.3. Ablation studies

We analyze the effectiveness of each proposed module in Hg-I2P through a series of ablation studies on the 7-Scenes dataset [14]. The Office scene is used for training, while the remaining scenes are used for evaluation. Additional results on hyperparameter sensitivity and runtime are provided in the supplemental materials.

Effectiveness of MP-mining. We first verify the MP-mining module in heterogeneous edge prediction. Table 7 shows that multi-path relational features improves edge prediction precision, which in turn leads to higher IR and RR scores. This confirms that MP-mining effectively captures richer inter-region dependencies for cross-modal learning.

Table 9. Ablation study of the HC-pruning module. "Criterion X" indicates the use of the X-th HC-pruning criterion.

Criterion I	Criterion II	IR	RR
		0.494	0.661
✓		0.523	0.682
✓	✓	0.558 (**+6.4%**)	0.690 (**+2.9%**)

Effectiveness of HE-adapting. We next evaluate the HE-adapting module for cross-modal feature adaptation. Results in Table 8 show that when $\beta = 0.1$, both IR and RR reach their peak values, indicating optimal interaction intensity between heterogeneous features. However, when β increases further, over-interaction reduces feature discriminability, highlighting the importance of balanced feature exchange. We also visualize the cross-modal features in 3D space with principal component analysis (PCA), shown in Fig. 5. This PCA visualization confirms that HE-adapting aligns 2D and 3D features more effectively, producing more accurate correspondences.

Effectiveness of HC-pruning. We finally assess the HC-pruning module for correspondence refinement. Table 9 shows that applying the reprojection distance-based criterion brings limited improvement on IR and RR, since $\tilde{\mathbf{T}}$ is still inaccurate. When the relative-coordinate consistency based criterion is also included, false correspondences are more effectively removed, leading to significant gains in IR and RR. As visualized in Fig. 6, HC-pruning reduces false-positive matches and enhances registration robustness.

6. Conclusions

In this paper, we proposed Hg-I2P, a heterogeneous graph-embedded image-to-point-cloud registration framework designed for robust and generalizable cross-modal alignment. The core idea is to introduce a heterogeneous graph that connects segmented 2D and 3D regions, enabling cross-modal feature interaction through heterogeneous edges and outlier suppression through graph-based projection consistency. Extensive experiments across six datasets, including both indoor and outdoor scenarios, demonstrate that Hg-I2P significantly outperforms existing methods in terms of generalization and accuracy.

Acknowledgments

This work is partially supported by the National Key R&D Program of China (Grant ID: 2024YFC3015303), National Natural Science Foundation of China (Grand ID: 62502171), and China Postdoctoral Science Foundation (Grand ID: 2024M761014 and GZC20252285).

References

[1] Pei An, Xuzhong Hu, Junfeng Ding, and et al. Ol-reg: Registration of image and sparse lidar point cloud with object-level dense correspondences. *IEEE Trans. Circuits Syst. Video Technol.*, 34(8):7523–7536, 2024. 2

[2] Pei An, Jiaqi Yang, Muyao Peng, and et al. Top-I2P: Explore open-domain image-to-point cloud registration using topology relationship. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1–9, 2025. 2, 3, 5, 6, 7, 8

[3] Pei An, Jiaqi Yang, Muyao Peng, You Yang, Qiong Liu, Xiaolin Wu, and Liangliang Nan. Mincd-pnp: Learning 2d-3d correspondences with approximate blind pnp. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2025. 1, 3, 6, 7

[4] Pei An, You Yang, Jiaqi Yang, Muyao Peng, Qiong Liu, and Liangliang Nan. Enhance image-to-point-cloud registration with beltrami flow. *Int. J. Comput. Vis.*, 133(12):8589–8616, 2025. 8

[5] Lin Bie, Shouan Pan, Siqi Li, and et al. Graphi2p: Image-to-point cloud registration with exploring pattern of correspondence via graph learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 22161–22171, 2025. 1, 2, 3, 6, 8

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, and et al. Emerging properties in self-supervised vision transformers. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 9630–9640, 2021. 2

[7] Daniele Cattaneo and Abhinav Valada. Cmrnext: Camera to lidar matching in the wild for localization and extrinsic calibration. *IEEE Trans. Robotics*, 41:1995–2013, 2025. 1

[8] Hansheng Chen, Pichao Wang, Fan Wang, and et al. 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*, pages 2771–2780, 2022. 2

[9] Zhixin Cheng, Jiacheng Deng, Xinjun Li, and et al. Bridge 2d-3d: Uncertainty-aware hierarchical registration network with domain alignment. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 2491–2499, 2025. 1, 2, 3, 6, 8

[10] Angela Dai, Angel X. Chang, Manolis Savva, and et al. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2432–2443, 2017. 6, 7

[11] 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. 1, 2

[12] Rao Fu, Jianmin Zheng, and Liang Yu. Consistent normal orientation for 3d point clouds via least squares on delaunay graph. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16932–16942, 2025. 1

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 6, 7

[14] 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. 5, 6, 7, 8

[15] 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. 3

[16] Shuhao Kang, Youqi Liao, Jianping Li, and et al. Cofii2p: Coarse-to-fine correspondences-based image to point cloud registration. *IEEE Robotics Autom. Lett.*, 9(11):10264–10271, 2024. 2, 6, 8

[17] 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. 1, 6

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, and et al. Segment anything. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 3992–4003, 2023. 2, 3

[19] 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. 5, 6, 7, 8

[20] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An accurate $O(n)$ solution to the pnp problem. *Int. J. Comput. Vis.*, 81(2):155–166, 2009. 5

[21] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *Proceedings of European Conference on Computer Vision*, pages 71–91, 2024. 2

[22] 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. 2

[23] Minhao Li, Zheng Qin, Zhirui Gao, and et al. 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. 1, 2, 3, 5, 6, 7, 8

[24] Xinjun Li, Wenfei Yang, Jiacheng Deng, and et al. Implicit correspondence learning for image-to-point cloud registration. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 16922–16931, 2025. 1, 2, 6, 8

[25] Juncheng Mu, Changwei Ren, Weixiang Zhang, and et al. Diff^{2}i2p: Differentiable image-to-point cloud registration with diffusion prior. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2025. 3, 8

[26] Riku Murai, Eric Dexheimer, and Andrew J. Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction

priors. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2025. 1, 2

[27] Quang-Hieu Pham, Mikaela Angelina Uy, Binh-Son Hua, and et al. LCD: learned cross-domain descriptors for 2d-3d matching. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 11856–11864, 2020. 7

[28] Zheng Qin, Hao Yu, Changjian Wang, and et al. 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. 2

[29] 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. 2, 6, 7

[30] 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. 7

[31] Jürgen Sturm, Nikolas Engelhard, Felix Endres, and et al. A benchmark for the evaluation of RGB-D SLAM systems. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 573–580, 2012. 6, 7

[32] Yifan Sun, Changmao Cheng, Yuhan Zhang, and et al. 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. 2, 5

[33] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, and et al. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 6410–6419, 2019. 3

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, and et al. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 4

[35] Bing Wang, Changhao Chen, Zhaopeng Cui, 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. 1, 2, 6

[36] Haiping Wang, Yuan Liu, Bing Wang, and et al. 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. 1, 2, 6, 7, 8

[37] Qianliang Wu, Haobo Jiang, Lei Luo, and et al. Diff-reg: Diffusion model in doubly stochastic matrix space for registration problem. In *Proceedings of IEEE/CVF European Conference on Computer Vision*, pages 160–178, 2024. 1, 2, 8

[38] Fan Yang, Chen Wang, Cesar Cadena, and Marco Hutter. iplanner: Imperative path planning. In *Proceedings of Robotics: Science and Systems*, pages 1–9, 2023. 1

[39] Lihe Yang, Bingyi Kang, Zilong Huang, and et al. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2

[40] Yunhan Yang, Xiaoyang Wu, Tong He, and et al. Sam3d: Segment anything in 3d scenes. In *Proceedings of IEEE/CVF International Conference on Computer Vision Workshops*, pages 1–5, 2023. 3, 6

[41] Gongxin Yao, Yixin Xuan, Xinyang Li, and Yu Pan. 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*, pages 13458–13465, 2024. 2, 8

[42] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 11784–11793, 2021. 2

[43] Yuanchao Yue, Zhengxin Li, Wei Zhang, and Hui Yuan. Papi-reg: Patch-to-pixel solution for efficient cross-modal registration between lidar point cloud and camera image. *CoRR*, abs/2503.15285, 2025. 2

[44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 3813–3824, 2023. 2

[45] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(11): 1330–1334, 2000. 3

[46] Xu Zhao, Wenchao Ding, Yongqi An, and et al. Fast segment anything. *CoRR*, abs/2306.12156:1–11, 2023. 6

[47] Junsheng Zhou, Baorui Ma, Wenyuan Zhang, and et al. 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. 2, 7

[48] Yuchen Zhou, Jiayuan Gu, Tung Yen Chiang, and et al. Point-sam: Promptable 3d segmentation model for point clouds. In *Proceedings of International Conference on Learning Representations*, pages 1–10, 2025. 2, 3