— Supplemental Material —
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

We provide additional details on the dataset, implementation, comparisons, and limitation analysis of the proposed Hg-I2P in this supplemental material.

A. More Details of the Self-Collected Dataset

To evaluate the generalization ability of I2P registration, we collected an I2P dataset on a campus building. Given that I2P registration is mainly used in AR/VR applications, this dataset was collected in representative indoor scenes, including tables, living rooms, paintings, books, and toys. RGB-D data was captured using an Intel RealSense depth camera `D415`, and images were resized to 640×480 resolution. The point clouds contain inherent noise caused by depth measurement errors, which is common in the real-world I2P dataset. Visualizations of this dataset are shown in Fig. A1. Ground truth (GT) camera poses are estimated using 3D Iterative Closest Point (ICP). For training, we also use RGB-D frames as samples, where the GT camera poses are set to the identity matrix. The depth range in this dataset varies from 1.0 m to 9.0 m. Because the dataset contains only 0.1K samples, it is used exclusively for zero-shot I2P registration.

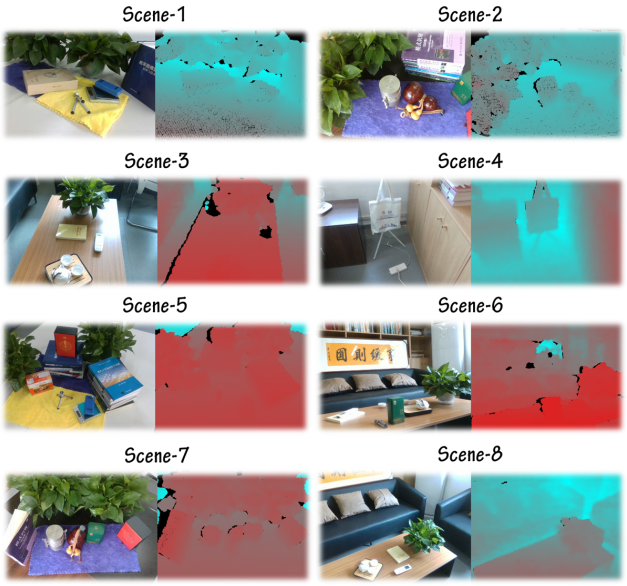

Figure A1. Examples from the self-collected indoor I2P dataset. Representative RGB and depth scenes captured using an Intel RealSense D415 in indoor environments such as tables, living rooms, and shelf-like spaces. The dataset includes realistic depth noise and fine-grained geometric variation and is used exclusively for zero-shot evaluation of I2P registration.

B. More Implementation Details of Hg-I2P

We provide additional technical details and motivations behind the implementation of Hg-I2P.

Segmentation. To construct the heterogeneous graph, the first step is segmenting RGB images and 3D point clouds. A naive strategy is to use pre-trained semantic segmentation models. However, real-world indoor scenes often contain unseen or long-tail objects that such models fail to segment reliably. Therefore, we adopt the Segment Anything Model (SAM) [9] to obtain fine-grained segmentation on RGB images. Because standard SAM requires nearly 1 second per image, we instead employ the pre-trained FastSAM [17] for faster inference.

To segment 3D point clouds, we follow Yang et al. [15], which generates a segmented point-cloud map by fusing a sequence of SAM-segmented RGB-D frames. In I2P registration scenarios, point clouds are typically pre-built from RGB-D videos, making this fusion-based approach [15] especially suitable for 3D segmentation. Since both images and point clouds are segmented by SAM, 2D and 3D regions maintain strong shape consistency, which simplifies the prediction of heterogeneous edges. More visualizations

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

Figure A2. 2D and 3D SAM-based segmentation results. The resulting 2D and 3D region sets exhibit strong shape and boundary consistency, enabling reliable heterogeneous graph construction for cross-modal representation learning.

of 2D and 3D segments are shown in Fig. A2.

Heterogeneous graph initialization. We now describe the graph initialization process. Given an image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ and a point cloud $\mathcal{P} \in \mathbb{R}^{N\times 3}$, we extract a 2D feature map $\mathbf{F}_I \in \mathbb{R}^{H\times W\times C}$ using a ResNet-based U-Net following [10]. The i-th vertex feature $\mathbf{v}_i^I$ is computed as:

$$\mathbf{v}_i^I = \texttt{Torch.mean}(\mathbf{F}_I[\texttt{Mask}_i^I]) \tag{A1}$$

Similarly, we extract 3D point features $\mathbf{F}_P \in \mathbb{R}^{N\times C}$ from $\mathcal{P}$ using a KPConv-based U-Net following [10], and compute the j-th vertex feature $\mathbf{v}_j^P$ as:

$$\mathbf{v}_j^P = \texttt{Torch.mean}(\mathbf{F}_P[\texttt{Mask}_j^P]) \tag{A2}$$

We employ the same ResNet and KPConv backbones as in [10] because of their efficiency and robustness, achieving a runtime of about 3 - 4 ms on an NVIDIA GTX 3080 GPU.

Heterogeneous edge prediction. As discussed in Sec. 5.1, I2P registration is primarily used for visual localization in the pre-built point cloud map, where a pose prior is available [4]. This allows segmenting the relevant sub-map for I2P registration, and implies that the GT transform matrix $\mathbf{T}_{\text{gt}}$ is close to the identity matrix $\mathbf{I}$. In MP-mining, to enforce edge sparsity, we apply a mask $\mathbf{M}_{\text{I2P}} \in \mathbb{R}^{M\times N}$ to $\hat{\mathbf{E}}_{\text{I2P}}$ to remove incorrect edges based on a reprojection threshold τ_{2d}. $\hat{\mathbf{E}}_{\text{I2P}}$ is computed as:

$$\hat{\mathbf{E}}_{\text{I2P}} \leftarrow \hat{\mathbf{E}}_{\text{I2P}} \odot \mathbf{M}_{\text{I2P}}, (\mathbf{M}_{\text{I2P}})_{ij} = \mathbf{1}(\|\bar{\mathbf{p}}_i^I - \bar{\mathbf{p}}_j^P\|_2 \leq \tau_{2d}) \tag{A3}$$

where $\mathbf{1}(\cdot)$ is an indicator function. $\odot$ denotes an element-wise product. $\bar{\mathbf{p}}_i^I$ is the average pixel coordinate of the i-th image segment. $\bar{\mathbf{p}}_j^P$ is the projection of the center point of the j-th segment of the point cloud using the identity projection. τ_{2d} is empirically set to 16.

Criteria in HC-Pruning. We next provide more details on HC-Pruning. To estimate the initial pose $\tilde{\mathbf{T}}$, we use a PnP-based cost over the centers of matched 2D-3D regions, i.e., $L_{\text{pnp}}(\tilde{\mathbf{T}}) = \sum_{i,k\in\mathcal{E}_{\text{I2P}}(i|I)} \|\bar{\mathbf{p}}_i^I - \pi(\bar{\mathbf{P}}_k^P|\tilde{\mathbf{T}})\|_2$, where $\bar{\mathbf{p}}_i^I$ and $\bar{\mathbf{P}}_k^P$ denote the respective centers of $\mathcal{I}_i$ and $\mathcal{P}_k$. We then prune based on graph-aware criteria. Because the heterogeneous edge relations are effective at filtering outliers, our **first pruning criterion** identifies an inlier if at least one of the three geometric consistency conditions holds, i.e., $\langle \mathbf{p}_i^c, \mathbf{P}_i^c \rangle$ *is an inlier if at least one of the conditions is satisfied: (i)* $\mathbf{p}_i^c \in \mathcal{I}_i$, $\mathbf{P}_i^c \in \mathcal{P}_k, \forall k \in \mathcal{E}_{I2P}(i|I)$*; (ii)* $\mathbf{P}_i^c \in \mathcal{P}_i$, $\mathbf{p}_i^c \in \mathcal{I}_k, \forall k \in \mathcal{E}_{I2P}(i|P)$*; (iii)* $\|\mathbf{p}_i^c - \pi(\mathbf{P}_i^c|\tilde{\mathbf{T}})\|_2 \leq \delta_{rej}$.

Afterward, we construct vectors $\mathbf{s}_i \in \mathbb{R}^M$ and $\mathbf{t}_i \in \mathbb{R}^N$ to describe each correspondence's relative position to graph vertices:

$$\begin{aligned} \mathbf{s}_i &= (\mathbf{p}_i^c - \bar{\mathbf{p}}_1^I, ..., \mathbf{p}_i^c - \bar{\mathbf{p}}_M^I)^T, \\ \mathbf{t}_i &= (\pi(\mathbf{P}_i^c|\tilde{\mathbf{T}}) - \bar{\mathbf{q}}_1^I, ..., \pi(\mathbf{P}_i^c|\tilde{\mathbf{T}}) - \bar{\mathbf{q}}_M^I)^T \end{aligned} \tag{A4}$$

where $(\bar{\mathbf{q}}_1^I, ..., \bar{\mathbf{q}}_M^I)^T = \hat{\mathbf{E}}_{\text{I2P}} \cdot (\pi(\bar{\mathbf{P}}_1^P|\tilde{\mathbf{T}}), ..., \pi(\bar{\mathbf{P}}_N^P|\tilde{\mathbf{T}}))^T$. In the ideal case (i.e., $\tilde{\mathbf{T}}$ and $\hat{\mathbf{E}}_{\text{I2P}}$ are correct), $\bar{\mathbf{q}}_i^I \approx \bar{\mathbf{p}}_i^I$ is satisfied. For an inlier, the cosine distance of $\mathbf{s}_i$ and $\mathbf{t}_i$ is close to 1. Thus, the **second pruning criterion** identifies an inlier if this cosine distance meets a threshold, i.e., $\langle \mathbf{p}_i^c, \mathbf{P}_i^c \rangle$ *is an inlier if the cosine distance of* $\mathbf{s}_i$ *and* $\mathbf{t}_i$ *is greater than* τ_{rej}. In practice, due to errors in $\tilde{\mathbf{T}}$ and $\hat{\mathbf{E}}_{\text{I2P}}$, we adopt a relaxed rule and treat a correspondence as an inlier if either criterion is satisfied.

Implementation details. The model is trained using the Adam optimizer for 30 epochs with a batch size of 1, a learning rate of 10^{-4}, and a weight decay of 10^{-6}. All training and testing are conducted on a single NVIDIA GTX 3080 GPU. Detailed parameters will be provided in the open-source release. Training from scratch requires approximately 6 hours, while fine-tuning requires 2-3 hours.

C. More Discussions of Method Comparisons

We provide additional analysis of method comparisons.

Table A1. Mean RTE and RRE of MATR, Top-I2P, MinCD, and Hg-I2P in the cross-scene setting on 7-Scenes. The variant Hg-I2P† excludes HC-pruning. The results highlight the impact of heterogeneous graph modeling and pruning on registration accuracy.

Methods	Venue	RTE/m	RRE/deg
MATR	ICCV 2023	0.029	0.955
Top-I2P	IJCAI 2025	0.030	0.963
MinCD	ICCV 2025	0.027	0.937
Hg-I2P		0.028	0.954
Hg-I2P†		**0.026**	**0.903**

Selection of compared methods. Top-I2P [1] (IJCAI'25) and MinCD [2] (ICCV'25) are representative recent works focusing on I2P registration in open-domain scenarios. They outperform previous methods such as MATR [10], P2-Net [13], FreeReg [14], and Bridge [5]. Therefore, we focus our comparison on Top-I2P, MinCD, and the proposed Hg-I2P.

In-depth analysis in indoor scenes. Top-I2P [1] and Hg-I2P share a common trait: *both leverage SAM-segmented 2D and 3D regions for I2P registration*. However, Hg-I2P yields significantly better performance. For example, in cross-scene testing (training on the Kitchen scene), Hg-I2P improves IR and RR by $\mathbf{7.9}\%$ and $\mathbf{11.9}\%$ over Top-I2P [1]. Table A1 shows that Hg-I2P achieves more accurate RTE and RRE metrics. Hg-I2P's superiority arises from three key factors: (i) a heterogeneous graph that systematically encodes 2D-2D, 3D-3D, and 2D-3D relations, whereas Top-I2P does not; (ii) multi-path relation mining, rather than the trivial GCN for 2D-3D region matching [1]; (iii) a complete cross-modal feature adaptation and refinement pipeline, compared to Top-I2P's simple interaction module.

MinCD [2] enhances I2P registration by improving 2D-3D correspondence learning. However, in cross-scene and cross-dataset experiments, Hg-I2P consistently achieves higher IR and RR, indicating that a comprehensive cross-modal feature adaptation is more effective than refining only the correspondence loss. Overall, these results demonstrate that Hg-I2P achieves stronger generalization in indoor scenes.

In-depth analysis on driving scenes. Hg-I2P was initially designed for indoor scenes and cannot be applied directly to outdoor scenarios because the overlap between LiDAR point clouds and RGB images is limited. To address this, we used pre-trained CorrI2P [12] to extract the overlapped LiDAR region corresponding to the RGB image. Depth ranges also differ dramatically between indoor and outdoor scenes, making KPConv's fixed radius unsuitable. To avoid radius retuning, we scale the LiDAR point cloud such that its maximum depth is below $5.0m$, and incorporate this scaling in the RTE computation.

Compared with outdoor-focused approaches such as CoFiI2P [8], CMR-Agent [16], GraphI2P [3], and ImCorr [11], Hg-I2P achieves more accurate performance because of : (i) fine-grained cross-modal feature interaction via the heterogeneous graph; (ii) pruning mechanism that reduces invalid correspondences; (iii) leveraging the overlapped point cloud to simplify I2P registration.

Visualizations of Hg-I2P and MATR [10] on the KITTI dataset [7] are presented in Fig. A3. It shows that the fine-tuned Hg-I2P predicts the high-quality correspondences across diverse traffic scenes. We also provide the SAM results on the KITTI dataset [7] in Fig. A4, which shows that SAM performs reliably outdoors, contributing to the strong RTE and RRE metrics on the KITTI dataset [7].

D. More Qualitative Comparisons

Beyond the KITTI visualizations in Fig. A3, we provide additional qualitative results. Comparisons among Top-I2P [1], MinCD [2], and Hg-I2P are provided in Fig. A5. We observe that Hg-I2P maintains stable performance across varied indoor scenes as it produces significantly fewer outliers. We also evaluate HC-pruning, and the qualitative comparisons are shown in Fig. A6, which further reduces outliers and improves IR.

Failure case analysis. Despite outperforming state-of-the-art methods on many public datasets, Hg-I2P still fails in certain cases, as shown in Fig. A7. In complex scenes, SAM may produce over-segmentation, leading to overly complex graph connectivity and inaccurate heterogeneous graph edge prediction. These inaccuracies can distort feature adaptation and produce correspondences with a larger number of outliers. Additionally, the pruning criteria are sensitive to their thresholds: tight thresholds remove true inliers, while loose thresholds fail to remove outliers. Future work will explore improved heterogeneous edge prediction and correspondence pruning strategies.

E. More Hyperparameter Experiments

The main hyper-parameters in Hg-I2P are α, δ_{rej}, τ_{rej}, and λ_1. We discuss their selections in this section.

(1) α controls the adjacency structure of 2D-2D and 3D-3D regions. If α approaches zero, the adjacency matrices lose sparsity; if too large, they become overly sparse. Both extremes degrade heterogeneous edge prediction. We set $\alpha = 1.6$, and the performance is stable when $\alpha \in [0.5, 2]$.

(2) δ_{rej} is the threshold in criterion I of HC-pruning. Because $\tilde{\mathbf{T}}$ is not perfectly accurate, a small threshold incorrectly removes many inliers (see Table A2). Experiments show that $\delta_{\text{rej}} = 15$ yields the best RR.

(3) τ_{rej} is the threshold in criterion II for HC-pruning. We find that an adaptive, top-K strategy improves IR and RR.

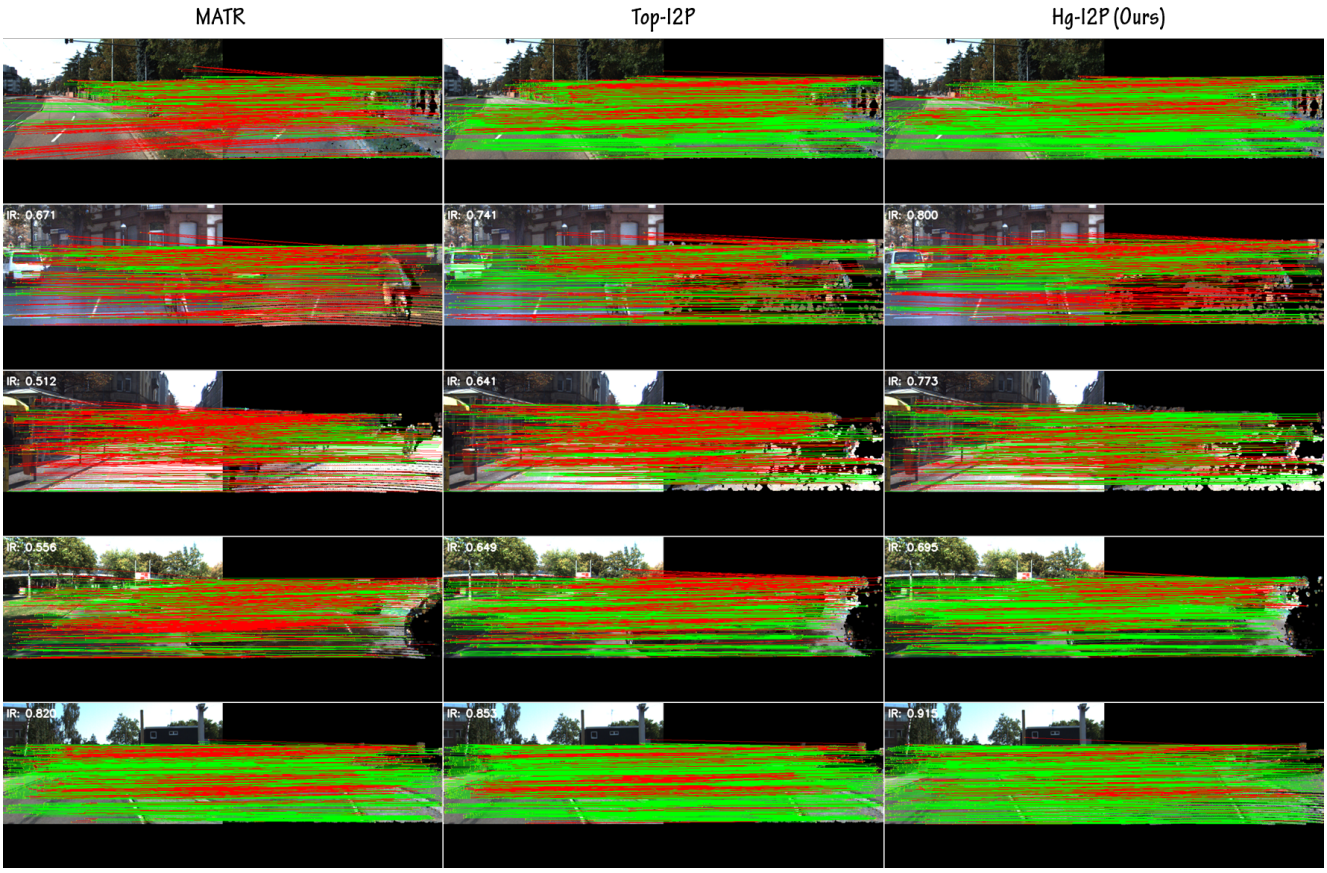

Figure A3. Qualitative I2P registration results of MATR [10] (left), Top-I2P [1] (middle), and the proposed Hg-I2P (right) on the KITTI dataset [7]. Hg-I2P produces more accurate and stable 2D-3D correspondences across diverse traffic scenes, benefiting from heterogeneous graph modeling and HC-pruning. Green and red lines denote correct and incorrect correspondences, respectively.

Figure A4. Examples of SAM segmentation results on outdoor images from KITTI [7]. Despite the presence of objects with irregular structures and large depth variability, SAM reliably identifies fine-grained regions, contributing to the robustness of Hg-I2P in outdoor environments after overlap extraction.

Specifically, we sort correspondences in descending order by cosine similarity and retain the top 85%. The ablation results are provided in Table A3.

(4) λ_1 controls the weight of heterogeneous edge prediction in the total loss. Because L_{corr} dominates cross-modal learning, λ_1 must remain below a certain value to ensure

$$\lambda_1 \leq L_{\text{corr}} / \|\hat{\mathbf{E}}_{\text{I2P}}[\text{mask}] - \mathbf{E}_{\text{I2P}}[\text{mask}]\|_2^2 \qquad \text{(A5)}$$

Guided by the inequality in (A5), we empirically set $\lambda_1 \leq 0.1$. The ablation result of λ_1 is provided in Table A4. If too small, heterogeneous edge prediction deteriorates; if too large, the model overemphasizes it at the expense of registration. The optimal value is set to 0.064.

We also evaluate the runtime of Hg-I2P, and the result is provided in Table A5. Hg-I2P runs at 5.6 FPS, compared to 12.8 FPS for MATR [10]. Most time is spent on

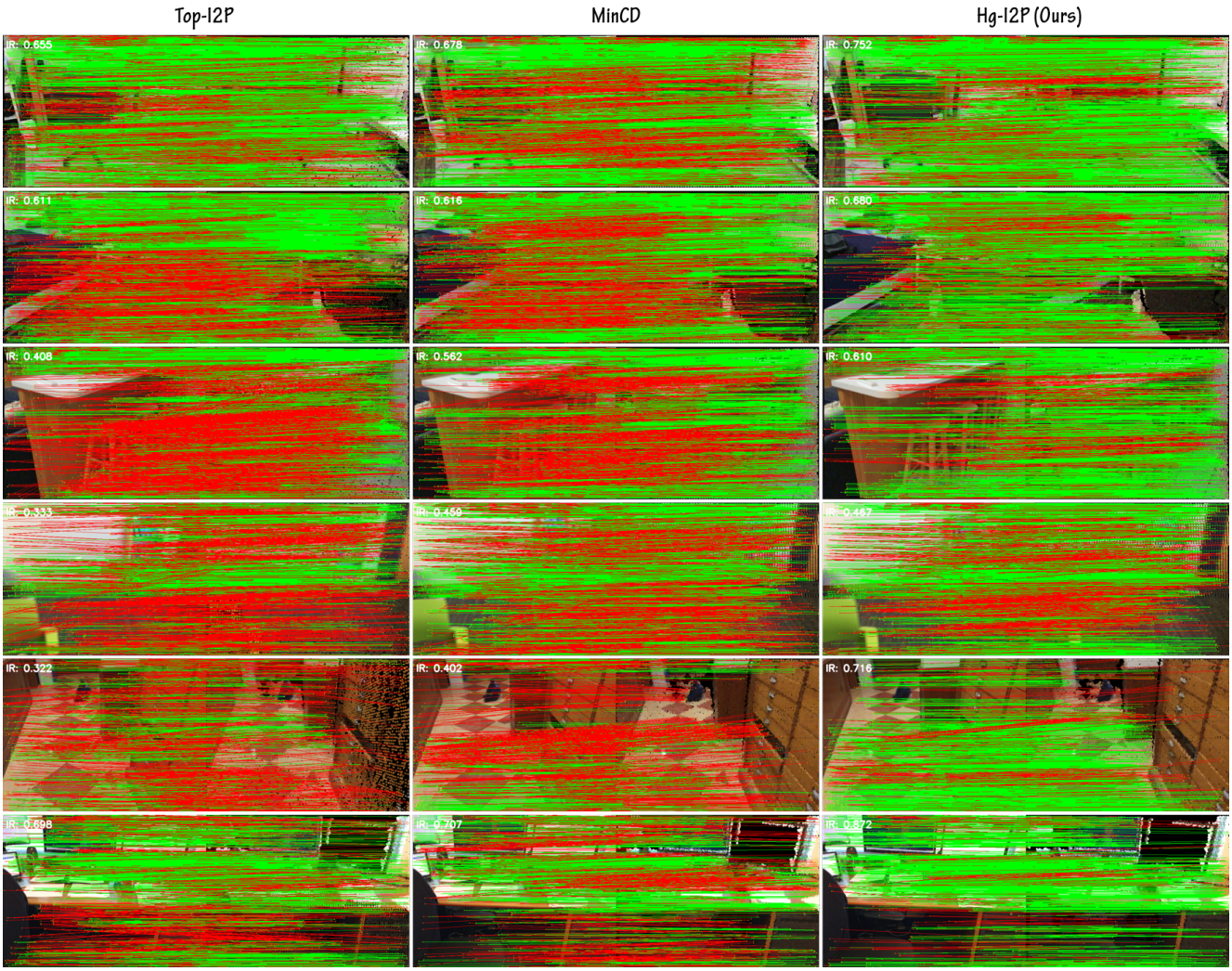

Figure A5. Qualitative comparison of Top-I2P [1] (left), MinCD [2], and Hg-I2P (right, ours) on ScanNet [6]. Hg-I2P consistently yields denser correct matches and significantly fewer outliers because of its heterogeneous graph design and multi-path relation mining. Green and red lines denote the correct and incorrect correspondences, respectively.

Table A2. Ablation results showing how the geometric reprojection threshold δ_{rej} influences registration recall. Moderate thresholds (e.g., $\delta_{\text{rej}} = 15$) achieve the best balance between retaining inliers and rejecting misaligned correspondences.

δ_{rej}	5	10	15	20	25
Registration recall (RR)	0.587	0.652	**0.682**	0.670	0.662

Table A3. Comparison of fixed and top-K adaptive cosine-similarity threshold τ_{rej} in HC-pruning. Adaptive selection improves both inlier ratio and registration recall.

Scheme of τ_{rej}	Fixed	Adaptive (Top-K selection)
Inlier ratio (IR)	0.537	**0.558**
Registration recall (RR)	0.685	**0.690**

2D SAM and graph initialization. Although slower, Hg-I2P offers stronger generalization.

Finally, we analyze the effect of 2D SAM prompting. From Fig. A8, it is observed that the segmentation number is different with prompts. Using denser grid prompts (e.g., 16×16 grid points) produces more regions and improves registration performance, but increases inference time (see Table A6). Applications may choose prompts based on accuracy and speed requirements.

F. Limitations and Future Works

Hg-I2P primarily depends on pre-trained 2D and 3D SAM. In complex scenes, it may produce over- or

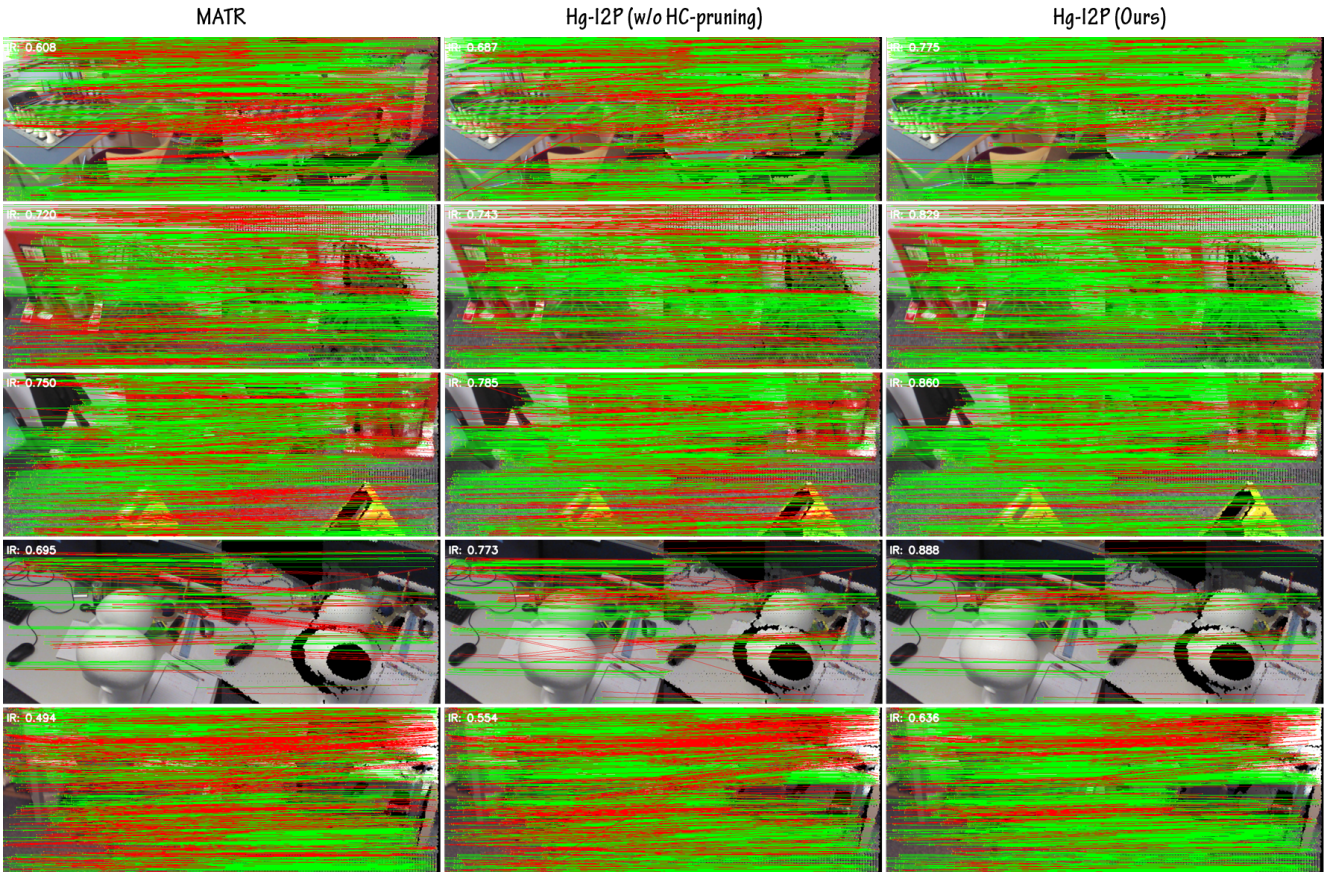

Figure A6. Qualitative comparison among MATR [10] (left), Hg-I2P without HC-pruning (middle), and Hg-I2P (right) on the 7-Scenes dataset [7]. HC-pruning noticeably reduces incorrect correspondences and enhances the inlier ratio by enforcing graph-aware geometric consistency. Green and red lines denote correct and incorrect correspondences, respectively.

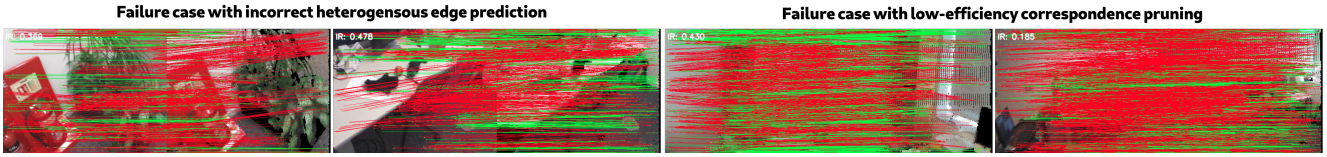

Figure A7. Examples where Hg-I2P fails to generate reliable correspondences. Errors arise from incorrect heterogeneous edge prediction and overly relaxed pruning thresholds, often triggered by SAM over-segmentation or highly cluttered scenes. Green and red lines denote correct and incorrect correspondences, respectively.

Table A4. Ablation study on λ_1 (without HC-pruning), demonstrating that moderate weighting yields optimal registration recall by balancing heterogeneous edge prediction and correspondence learning.

λ_1	0.032	0.048	0.064	0.080	0.096
Registration recall (RR)	0.602	0.638	**0.661**	0.652	0.626

Table A5. Runtime of each pipeline component, including SAM segmentation, graph initialization (including feature extraction), multi-path relation mining, heterogeneous edge adaptation, and HC-pruning. Graph initialization and SAM dominate the total runtime (unit: ms).

Proc.	SAM	Ini.	MP-mining	HE-adapting	HC-pruning	All
Time	47.2	72.4	3.9	34.6	19.7	177.8

under-segmentation, especially for outdoor scenes. Over-segmentation complicates heterogeneous edge prediction and reduces registration stability, as shown in Fig. A7.

Moreover, the inference time for 2D and 3D SAM is substantially higher than that of the other pipeline components.

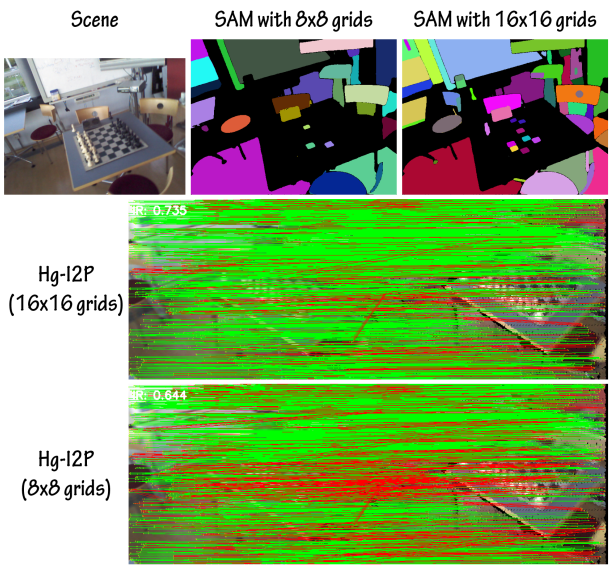

Figure A8. Visualization of how different grid-based prompting schemes for SAM (e.g., 8×8 vs. 16×16) affect segmentation granularity and the resulting 2D–3D correspondences. Denser prompting generally improves alignment but increases inference time. Green and red lines denote correct and incorrect correspondences, respectively.

Table A6. Evaluation of segmentation prompt density (8×8 vs. 16×16 grid points) on registration accuracy and inference time. Denser prompting provides higher registration recall at the cost of additional computation.

Prompt of 2D SAM	RR	Time
8×8 grid points	62.1	**144.2 ms**
16×16 grid points (used in experiments)	**68.2**	177.8 ms

To address these limitations, we plan to explore constructing and applying heterogeneous graphs based on conformal geometry, which may offer a more reliable structure for cross-modal representation.

References

[1] 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. 3, 4, 5

[2] 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. 3, 5

[3] 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. 3

[4] 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. 2

[5] 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. 3

[6] 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. 5

[7] 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. 3, 4, 6

[8] 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. 3

[9] 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. 1

[10] 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. 2, 3, 4, 6

[11] 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. 3

[12] 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. 3

[13] 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. 3

[14] 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. 3

[15] 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. 1

[16] 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. 3

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