MinCD-PnP: Learning 2D-3D Correspondences with Approximate Blind PnP

Pei An[1]*, Jiaqi Yang[2]*, Muyao Peng[1], You Yang[1]†, Qiong Liu[1], Xiaolin Wu[3], Liangliang Nan[4]
[1]Huazhong University of Science and Technology, China
[2]Northwestern Polytechnical University, China
[3]Southwest Jiaotong University, China [4]Delft University of Technology, Netherlands

Abstract

Image-to-point-cloud (I2P) registration is a fundamental problem in computer vision, focusing on establishing 2D-3D correspondences between an image and a point cloud. Recently, the differentiable perspective-n-point (PnP) has been widely used to supervise I2P registration networks by enforcing projective constraints on 2D-3D correspondences. However, differentiable PnP is highly sensitive to noise and outliers in the predicted correspondences, which hinders the effectiveness of correspondence learning. Inspired by the robustness of blind PnP to noise and outliers in correspondences, we propose an approximate blind PnP-based correspondence learning approach. To mitigate the high computational cost of blind PnP, we reformulate it as a more tractable problem: minimizing the Chamfer distance between learned 2D and 3D keypoints, referred to as MinCD-PnP. To effectively solve MinCD-PnP, we introduce a lightweight multi-task learning module, MinCD-Net, which can be easily integrated into the existing I2P registration architectures. Extensive experiments on 7-Scenes, RGBD-V2, ScanNet, and self-collected datasets demonstrate that MinCD-Net outperforms state-of-the-art methods and achieves higher inlier ratio and registration recall in both cross-scene and cross-dataset settings. The source code: `https://github.com/anpei96/mincd-pnp-demo`.

1. Introduction

Image-to-point-cloud (I2P) registration [12] is a fundamental task in computer vision [2], aiming to establish 2D-3D correspondences between images and point clouds [35]. These correspondences are used to estimate the six-degree-of-freedom (6 DoF) camera pose with the perspective-n-point (PnP) algorithm [23], enabling I2P registration by aligning images with point clouds. Thus, I2P registration is widely used in visual localization, navigation, visual odometry, 3D reconstruction, and so on [1, 13, 21, 26, 34].

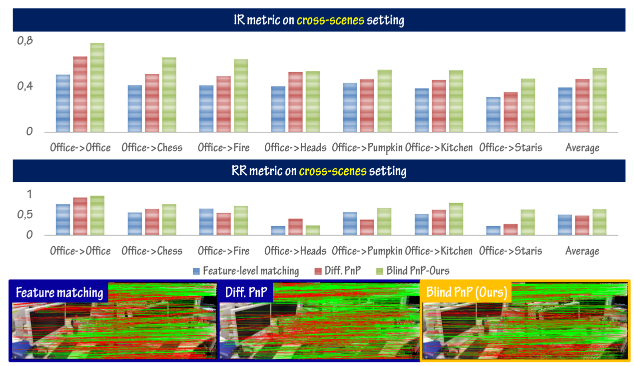

Figure 1. To overcome the limitation of **feature-level matching**, **differentiable PnP** employs the projective constraints of 2D-3D correspondences but is highly sensitive to correspondence quality. In this paper, we incorporate **blind PnP** to enhance I2P registration and achieve a salient improvement compared to other methods.

Learning-based approaches have gained significant attention in I2P registration [1, 35]. Deep neural networks (DNNs) help bridge the modality gap between images and point clouds [2, 25] by estimating 2D-3D correspondences through pixel-to-point feature-level matching (i.e., comparing feature distances between each 2D pixel and 3D point) [12]. However, feature-level matching struggles to remove outliers, as it ignores the projective constraints inherent in 2D-3D correspondences, as shown in Fig. 1.

To utilize the constraints of projective geometry in learning 2D-3D correspondences, the mainstream technique leverages differentiable perspective-n-point (PnP) [4, 6, 42]. The objective is to refine camera pose estimation via differentiable PnP, thereby improving the accuracy of global projective correspondences. However, differentiable PnP is highly sensitive to noise and outliers in the predicted correspondences [37]. This issue makes the estimated camera pose unreliable, thus hindering the effectiveness of differentiable PnP on correspondence learning.

*Equal contribution
†Corresponding author: yangyou@hust.edu.cn

To overcome the limitations of differentiable PnP, inspired by the robustness of blind PnP against noise and outliers in correspondences [5], we propose an approximate blind PnP based 2D-3D correspondence learning approach. Since blind PnP is computationally expensive [5], we reformulate it as a task of **min**imizing **C**hamfer **d**istance between the learned 2D and 3D keypoints, called **MinCD-PnP** in the sequel. MinCD-PnP ensures the feasibility of learning correspondence with blind PnP and retains the robustness of blind PnP to noise and outliers in correspondences. To effectively solve MinCD-PnP, we propose a lightweight multi-task learning module, denoted by **MinCD-Net**. Operationally, MinCD-Net can be seamlessly integrated into the existing I2P registration architectures and jointly optimized in an end-to-end manner. Extensive experiments on the 7-Scenes [14], RGBD-V2 [22], ScanNet [8], and self-collected datasets show that MinCD-Net achieves a higher inlier ratio (IR) and registration recall (RR) than state-of-the-art methods in both cross-scene and cross-dataset settings. Our core contributions are:

- We introduce MinCD-PnP, a formulation that simplifies blind PnP into a more tractable task of minimizing the Chamfer distance between learned 2D and 3D keypoints.
- We design a lightweight, multi-task learning module, MinCD-Net, to effectively solve MinCD-PnP. It can be easily integrated into existing I2P registration pipelines.
- MinCD-Net achieves superior performance across five datasets, outperforming state-of-the-art methods in both cross-scene and cross-dataset generalization.

2. Related Work

I2P registration. Most I2P registration methods rely on deep learning, as DNNs help bridge the modality gap between images and point clouds. Feng *et al.* designed the first deep learning based method for I2P registration, training a DNN to learn 3D keypoints descriptors [12]. Li and Lee [24] developed DeepI2P, which enhances the feature representation through global feature interaction. Ren *et al.* [29] further refined this approach in 2023. Building on the image registration method D2-Net [10], Wang *et al.* [35] developed P2-Net, which jointly learns 2D-3D keypoints and their descriptors. Circle loss [32] was used to alleviate the extreme imbalance between inliers and outliers. Li *et al.* [25] followed the point cloud registration architecture GeoTrans [28] to develop 2D3D-MATR, which outperformed P2-Net [35]. This work was further improved by Wu *et al.* [38] in 2024 by integrating a diffusion model [17] to iteratively denoise correspondence matrix. In 2024, An *et al.* [2] introduced Proj-ICP, a non-learning algorithm to estimate camera pose by minimizing the 2D-3D contour distances. They also surveyed to summarize the I2P registration methods for LiDAR-camera extrinsic calibration [1]. Wang *et al.* [36] designed an architecture, FreeReg which utilized the pre-trained vision fundamental models to minimize the modality difference between images and point clouds. Based on the above discussions, most current methods **follow a pixel-to-point feature-matching paradigm** to establish correspondences.

Learning correspondences with PnP. Recent research has highlighted the absence of the geometrical constraint in I2P registration, leading to the development of differentiable PnP for improved correspondence learning. In 2023, Zhou *et al.* [42] explored the effect of end-to-end probabilistic PnP (EPro-PnP) [6] on the 2D-3D correspondence learning task. Although EPro-PnP is robust to correspondence noise, its performance becomes unstable in the presence of excessive outliers. In 2024, Wu *et al.* [38] regarded correspondence learning as a denoising procedure and combined the diffusion model with differentiable PnP to refine 2D-3D correspondences. To make differentiable PnP more robust to correspondence noise and outliers, Campbell *et al.* [4] were the first to study blind PnP and designed a weighted differentiable blind PnP layer based on a declarative network [15]. In their work [4], RANSAC-based PnP [9] filters correspondences with large noises, and the declarative network computes the loss backward gradient of RANSAC-based PnP layer. Although work [4] is robust to correspondence noise and outliers, the loss gradient from filtered correspondences provides limited benefits to the overall I2P architecture. Thus, **an effective differentiable PnP for I2P registration** is still an open problem.

3. Problem Formulation and Analysis

In this section, we revisit I2P registration from an optimization perspective and analyze the bottleneck of 2D-3D correspondence learning (as illustrated in Fig. 2). For a given pixel $q \in \mathcal{I}$ and a point $p \in \mathcal{P}$, their correspondence $\langle q, p\rangle$ is determined using feature-level matching [25, 35, 36]:

$$d(\mathbf{f}_q^{\mathrm{2D}}, \mathbf{f}_p^{\mathrm{3D}}) \leq \delta \Rightarrow \langle q, p\rangle \text{ is a correspondence} \quad (1)$$

$$\mathbf{F}_{\mathrm{I}}, \mathbf{F}_{\mathrm{P}} = \varphi(\mathcal{I}, \mathcal{P}) \quad (2)$$

where $d(\cdot,\cdot)$ represents the per-feature normalized L_2 distance, and δ is a predefined threshold. The features $\mathbf{f}_q^{\mathrm{2D}}$ and $\mathbf{f}_p^{\mathrm{3D}}$ on q and p are extracted from $\mathbf{F}_{\mathrm{I}}$ and $\mathbf{F}_{\mathrm{P}}$, respectively, and φ denotes the neural network used for I2P registration. It is learned by the following optimization problem:

$$\varphi^{\star} = \arg\min_{\varphi} \sum_{p,q} L_{\mathrm{corr}}(\mathbf{f}_q^{\mathrm{2D}}, \mathbf{f}_p^{\mathrm{3D}}) \quad (3)$$

where p, q are pixel-to-point pair that satisfies Eq. (1). L_{corr} is the common correspondence loss, such as circle loss [32], because it helps mitigate the severe imbalance between inliers and outliers [25, 35].

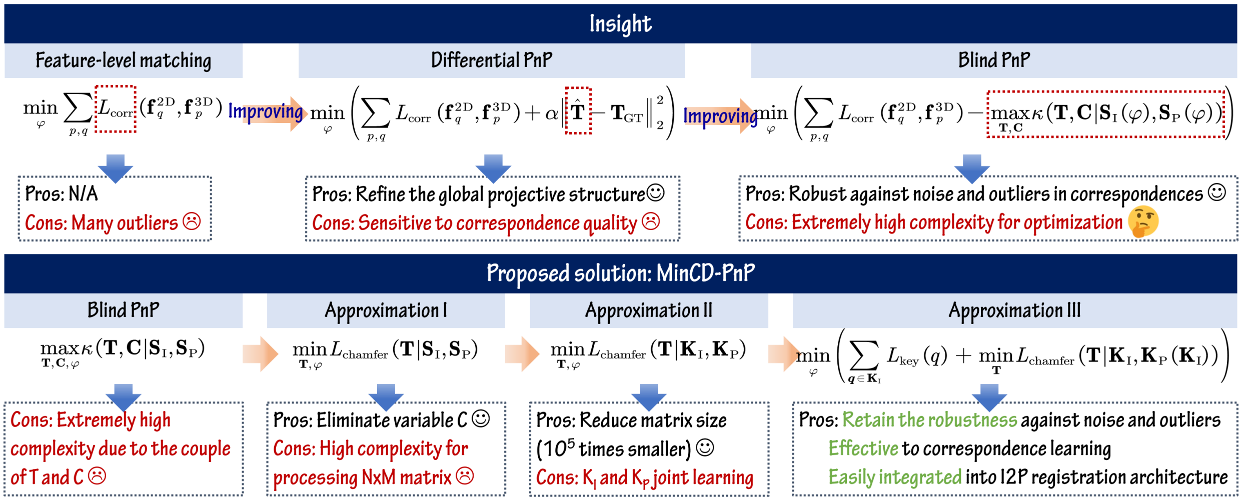

Figure 2. Motivation for the proposed MinCD-PnP. First, we analyze correspondence learning from an optimization perspective and observe that blind PnP is robust to the correspondence quality. To mitigate the complexity of blind PnP, we simplify blind PnP as a new task, MinCD-PnP, using a triple approximation strategy.

The optimization in Eq. (3) is suboptimal, as it ignores the projective constraint of $\langle q,p\rangle$. A valid correspondence $\langle q,p\rangle$ must satisfy $q=\pi(\mathbf{T}p)$, where $\pi(\cdot)$ represents the camera projection operator [40]. $\mathbf{T}$ is the transformation from the point cloud to the camera coordinate system. Differentiable PnP based methods incorporate projective constraints [38, 42] by refining Eq. (3) as:

$$\min_{\varphi}\left(\sum_{p,q} L_{\text{corr}}(\mathbf{f}_q^{\text{2D}},\mathbf{f}_p^{\text{3D}})+\alpha\|\hat{\mathbf{T}}-\mathbf{T}\|_2^2\right) \tag{4}$$

$$\hat{\mathbf{T}}=\arg\min_{\mathbf{T}}\sum_{p,q}\mathbb{I}(d(\mathbf{f}_q^{\text{2D}},\mathbf{f}_p^{\text{3D}})\le\delta)\cdot\|q-\pi(\mathbf{T}p)\|_2^2 \tag{5}$$

where $\|q-\pi(\mathbf{T}p)\|_2$ measures the reprojection error of correspondence $\langle q,p\rangle$. $\mathbb{I}(x)$ is an indicator function that outputs 1 if x is true, or 0 otherwise. Equations (4) and (5) form a coupled optimization problem, with α controlling the weight of the pose loss. In Eq. (4), the term $\|\hat{\mathbf{T}}-\mathbf{T}\|_2^2$ enforces global geometric consistency and improves the accuracy of estimated correspondences. However, solving Eq. (5) is highly sensitive to noise and outliers in the correspondences [37]. Moreover, since φ inevitably predicts outliers and noisy inliers, existing differentiable PnP methods struggle to improve correspondence learning effectively.

4. Proposed Method

4.1. Motivation

We aim to enhance 2D-3D correspondence learning by leveraging blind PnP. An overview of the proposed approach is illustrated in Fig. 2. With the blind PnP cost function [5], we revise Eq. (3) as:

$$\min_{\varphi}\left(\sum_{p,q} L_{\text{corr}}(\mathbf{f}_q^{\text{2D}},\mathbf{f}_p^{\text{3D}})-\max_{\mathbf{T},\mathbf{C}}\kappa(\mathbf{T},\mathbf{C}|\mathbf{S}_{\text{I}}(\varphi),\mathbf{S}_{\text{P}}(\varphi))\right)$$
$$\text{s.t. } \mathbf{T}\in\mathbf{SE}(3),\mathbf{C}\in\mathbb{B}^{M\times N},\mathbf{S}_{\text{I}}=\{q_i\}_{i=1}^{M},\mathbf{S}_{\text{P}}=\{p_i\}_{i=1}^{N} \tag{6}$$

$$\kappa(\mathbf{T},\mathbf{C}|\mathbf{S}_{\text{I}},\mathbf{S}_{\text{P}})=\sum_{\langle q,p\rangle\in\mathbf{C}}\mathbb{I}(\|q-\pi(\mathbf{T}p)\|_2^2\le\tau) \tag{7}$$

where $\mathbf{S}_{\text{I}}(\varphi)$ and $\mathbf{S}_{\text{P}}(\varphi)$ are pixel and point sets of the candidate correspondences, sampled from $\mathbf{F}_{\text{I}}$ and $\mathbf{F}_{\text{P}}$ via Eq. (1). As $\mathbf{F}_{\text{I}}$ and $\mathbf{F}_{\text{P}}$ are learned from φ, $\mathbf{S}_{\text{I}}(\varphi)$ and $\mathbf{S}_{\text{P}}(\varphi)$ can be regarded as functions of φ. For the discussion simplicity, $\mathbf{S}_{\text{I}}(\varphi)$ and $\mathbf{S}_{\text{P}}(\varphi)$ are simplified as $\mathbf{S}_{\text{I}}$ and $\mathbf{S}_{\text{P}}$. $\mathbf{C}$ is a boolean $M\times N$ matrix to denote the correspondences between $\mathbf{S}_{\text{I}}$ and $\mathbf{S}_{\text{P}}$. $\kappa(\mathbf{T},\mathbf{C}|\mathbf{S}_{\text{I}},\mathbf{S}_{\text{P}})$ denotes the inlier number, and τ is a pixel threshold to determine whether the correspondence is an inlier. Blind PnP is robust to noise and outliers in correspondences by jointly optimizing $\mathbf{T}$ and $\mathbf{C}$. However, optimizing $\kappa(\mathbf{T},\mathbf{C}|\mathbf{S}_{\text{I}},\mathbf{S}_{\text{P}})$ is computationally intractable due to its high complexity [31], so blind PnP cannot be directly used for correspondence learning.

4.2. MinCD-PnP formulation

To address this challenge, we propose MinCD-PnP by simplifying blind PnP with a triple approximation strategy.

4.2.1. Approximation I: from inlier maximization to Chamfer distance minimization

First, we approximate the inlier maximization cost function $\kappa(\mathbf{T}, \mathbf{C}|\mathbf{S}_\mathrm{I}, \mathbf{S}_\mathrm{P})$ as a lightweight Chamfer distance minimization. To reach this goal, we study an inequality:

$$\begin{aligned}\max_{\mathbf{T},\mathbf{C}} \kappa(\mathbf{T}, \mathbf{C}|\mathbf{S}_\mathrm{I}, \mathbf{S}_\mathrm{P}) &\leq \max_{\mathbf{T}} \kappa(\mathbf{T}, \mathbf{C}^\star|\mathbf{S}_\mathrm{I}, \mathbf{S}_\mathrm{P}) \\ &\leq \max_{\mathbf{T}} \kappa^\star(\mathbf{T}^\star|\mathbf{S}_\mathrm{I}, \mathbf{S}_\mathrm{P})\end{aligned} \tag{8}$$

$$\begin{aligned}\kappa^\star(\mathbf{T}|\mathbf{S}_\mathrm{I}, \mathbf{S}_\mathrm{P}) = &\sum_{q\in\mathbf{S}_\mathrm{I}} \mathbb{I}(\min_{p\in\mathbf{S}_\mathrm{P}} \|q - \pi(\mathbf{T}p)\|_2^2 \leq \tau) \\ &+ \sum_{p\in\mathbf{S}_\mathrm{P}} \mathbb{I}(\min_{q\in\mathbf{S}_\mathrm{I}} \|q - \pi(\mathbf{T}p)\|_2^2 \leq \tau)\end{aligned} \tag{9}$$

where $\mathbf{C}^\star$ is the optimal correspondence matrix and $\kappa(\mathbf{T}, \mathbf{C}^\star|\mathbf{S}_\mathrm{I}, \mathbf{S}_\mathrm{P}) \geq \kappa(\mathbf{T}, \mathbf{C}|\mathbf{S}_\mathrm{I}, \mathbf{S}_\mathrm{P})$. We explain the last term in inequality (8). For a correspondence $\langle q, p\rangle \in \mathbf{C}^\star$, based on the above assumption, we both have $q = \arg\min_{q\in\mathbf{S}_\mathrm{I}} \|q - \pi(\mathbf{T}p)\|_2^2$ and $p = \arg\min_{p\in\mathbf{S}_\mathrm{P}} \|q - \pi(\mathbf{T}p)\|_2^2$. And $2\kappa(\mathbf{T}^\star, \mathbf{C}^\star|\mathbf{S}_\mathrm{I}, \mathbf{S}_\mathrm{P}) = \kappa^\star(\mathbf{T}^\star|\mathbf{S}_\mathrm{I}, \mathbf{S}_\mathrm{P}) = 2N$, where $\mathbf{T}^\star$ is the optimal pose. It leads to the last term in inequality (8). Based on inequality (8), we reformulate the inlier maximization objective in Eq. (6) as a Chamfer distance minimization cost function:

$$\min_{\varphi}\left(\sum_{p,q} L_\text{corr}(\mathbf{f}_q^\text{2D}, \mathbf{f}_p^\text{3D}) + \min_{\mathbf{T}} L_\text{Chamfer}(\mathbf{T}|\mathbf{S}_\mathrm{I}, \mathbf{S}_\mathrm{P})\right) \tag{10}$$

$$\begin{aligned}L_\text{Chamfer}(\mathbf{T}|\mathbf{S}_\mathrm{I}, \mathbf{S}_\mathrm{P}) = &\sum_{q\in\mathbf{S}_\mathrm{I}} \min_{p\in\mathbf{S}_\mathrm{P}} \|q - \pi(\mathbf{T}p)\|_2^2 \\ &+ \sum_{p\in\mathbf{S}_\mathrm{P}} \min_{q\in\mathbf{S}_\mathrm{I}} \|q - \pi(\mathbf{T}p)\|_2^2\end{aligned} \tag{11}$$

Eq. (10) **eliminates** the $M \times N$ boolean matrix $\mathbf{C}$ in Eq. (6), which significantly reduces computation complexity

4.2.2. Approximation II: reducing complexity in Chamfer distance optimization with keypoints

In the second stage, we introduce further refinements to improve the optimization efficiency of Eq. (10). Since images typically contain 10^6 pixels and point clouds 10^5 points, $M \times N$ can exceed 10^{11}, leading to a prohibitively expensive Chamfer distance computation. To address this problem, we sample the representative keypoints $\mathbf{K}_\mathrm{I} = \{q_i\}_{i=1}^{M_0}$ and $\mathbf{K}_\mathrm{P} = \{p_i\}_{i=1}^{N_0}$ * from $\mathbf{S}_\mathrm{I}$ and $\mathbf{S}_\mathrm{P}$, and revise Eq. (10) as:

$$\min_{\varphi}\left(\sum_{p,q} L_\text{corr}(\mathbf{f}_q^\text{2D}, \mathbf{f}_p^\text{3D}) + \min_{\mathbf{T}} L_\text{Chamfer}(\mathbf{T}|\mathbf{K}_\mathrm{I}, \mathbf{K}_\mathrm{P})\right) \tag{12}$$

A key advantage of Eq. (12) is the reduction of the Chamfer distance matrix from $M \times N$ to $M_0 \times N_0$. As 2D and 3D keypoints number is nearly 10^3, the matrix size is **smaller than** 10^5 **times**. Although Eq. (12) improves optimization efficiency, a key challenge remains: how to effectively learn the representative $\mathbf{K}_\mathrm{I}$ and $\mathbf{K}_\mathrm{P}$? To ensure that $L_\text{Chamfer}(\mathbf{T}|\mathbf{K}_\mathrm{I}, \mathbf{K}_\mathrm{P})$ contributes effectively to φ, $\mathbf{K}_\mathrm{I}$ and $\mathbf{K}_\mathrm{P}$ should sufficiently represent 2D and 3D spaces.

4.2.3. Approximation III: learning 3D keypoints with the guidance of 2D keypoints

To deal with the above learning problem of $\mathbf{K}_\mathrm{I}$ and $\mathbf{K}_\mathrm{P}$, we design the third approximation that approximates joint 2D and 3D keypoints learning as a single learning task. We aim to learn 3D keypoints that **mimic the 2D keypoints distribution**, since jointly learning both with sufficient inliers is a challenging task [25]. In this scheme, $\mathbf{K}_\mathrm{I}$ is pre-detected using a pre-trained model or a non-learning algorithm. Existing 2D keypoint detectors ensure that $\mathbf{K}_\mathrm{I}$ captures representative structures in the image. Then, we design a 2D keypoints guided 3D keypoints learning scheme:

$$\begin{aligned}&\min_{\varphi} \sum_{q\in\mathbf{K}_\mathrm{I}} \|q - \pi(\mathbf{T}p)\|_2 \\ &\text{s.t. } p = \arg\min_{p\in\mathcal{P}} d(\mathbf{f}_q^\text{2D}, \mathbf{f}_p^\text{3D}),\ q \in \mathbf{K}_\mathrm{I}\end{aligned} \tag{13}$$

However, directly learning with Eq. (13) is unreliable, as some 2D keypoints lack salient features, making it difficult to identify corresponding 3D points. It makes the loss in Eq. (13) unstable. So, we approximate Eq. (13) as:

$$\begin{aligned}&\min_{\varphi} \sum_{q\in\mathbf{K}_\mathrm{I}} L_\text{key}(q) \\ &= \sum_{q\in\mathbf{K}_\mathrm{I}} -\mathbb{I}(\|q - \pi(\mathbf{T}_\text{gt}p_q^\star)\|_2^2 \leq \tau) \cdot \mathbb{I}(s_q^\star \leq s_\text{th})\end{aligned} \tag{14}$$

$$\begin{aligned}p_q^\star &= \arg\min_{p\in\mathcal{P}}\{d(\mathbf{f}_q^\text{2D}, \mathbf{f}_1^\text{3D}), ...d(\mathbf{f}_q^\text{2D}, \mathbf{f}_p^\text{3D})..., d(\mathbf{f}_q^\text{2D}, \mathbf{f}_{N_0}^\text{3D})\} \\ s_q^\star &= \min_{p\in\mathcal{P}}\{d(\mathbf{f}_q^\text{2D}, \mathbf{f}_1^\text{3D}), ...d(\mathbf{f}_q^\text{2D}, \mathbf{f}_p^\text{3D})..., d(\mathbf{f}_q^\text{2D}, \mathbf{f}_{N_0}^\text{3D})\}\end{aligned} \tag{15}$$

where the term $\min\{d(\mathbf{f}_q^\text{2D}, \mathbf{f}_1^\text{3D}), ..., d(\mathbf{f}_q^\text{2D}, \mathbf{f}_{N_0}^\text{3D})\}$ approximates $d(\mathbf{f}_q^\text{2D}, \mathbf{f}_p^\text{3D})$. This implies that Eq. (15) aims to learn a set of 3D keypoints that best **approximate** the detected 2D keypoints in an error bound of τ. s_th is a threshold used to filter out low-confidence 3D keypoints.

*As shown in Eq. (6), $\mathbf{S}_\mathrm{I}$ and $\mathbf{S}_\mathrm{P}$ are functions of φ, so that $\mathbf{K}_\mathrm{I}$ and $\mathbf{K}_\mathrm{P}$ are also functions of φ. It means that $\mathbf{K}_\mathrm{I}$ and $\mathbf{K}_\mathrm{P}$ are learned from φ.

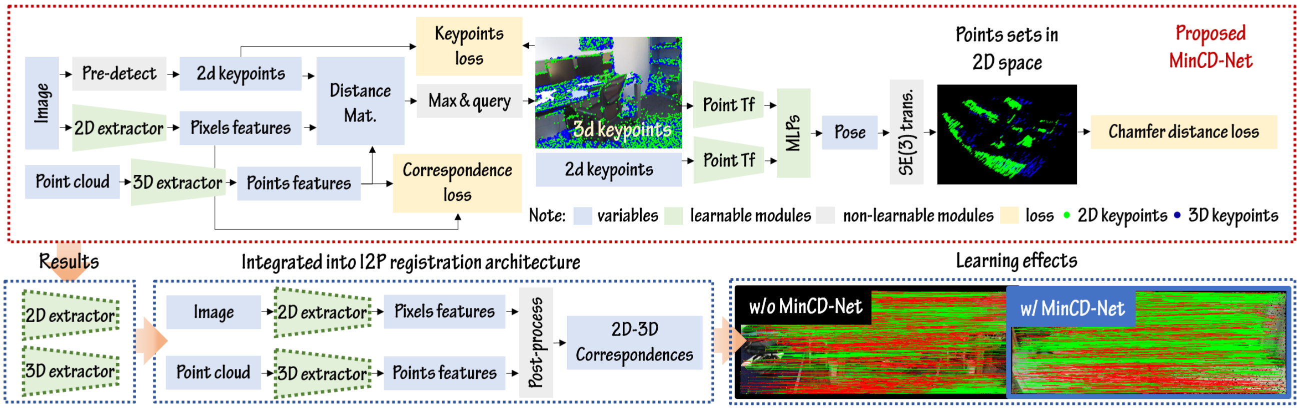

Figure 3. Proposed 2D-3D correspondence learning module, MinCD-Net. It converts the optimization in Eq. (16) into a **multi-task learning mechanism**. MinCD-Net can be integrated into existing I2P registration frameworks.

Following the triple approximation, we formulate the proposed scheme as a new optimization problem that minimizes Chamfer distance and 3D keypoints learning losses:

$$\varphi^\star = \arg\min_{\varphi} \left(\sum_{p,q} L_{\text{corr}}(\mathbf{f}_q^{2\text{D}}, \mathbf{f}_p^{3\text{D}}) + \lambda_1 \sum_{q \in \mathbf{K}_\text{I}} L_{\text{key}}(q) \right.$$
$$\left. + \lambda_2 \min_{\mathbf{T}} L_{\text{Chamfer}}(\mathbf{T}|\mathbf{K}_\text{I}, \mathbf{K}_\text{P}(\mathbf{K}_\text{I})) \right) \tag{16}$$

where λ_1 and λ_2 are the loss weights. $\mathbf{K}_\text{P}(\mathbf{K}_\text{I})$ denotes that 3D keypoints are learned from 2D keypoints via Eq. (14).

4.3. Correspondence learning with MinCD-PnP

In Sec. 4.2, we have modeled the correspondence learning as a MinCD-PnP problem. To address the MinCD-PnP formulation, we introduce MinCD-Net, a lightweight multi-task learning module, as shown in Fig. 3. Its core is to predict 3D keypoints and compute multi-task losses in Eq. (16).

4.3.1. General architecture of I2P registration

Before detailing the proposed MinCD-Net, we briefly review the architecture of the I2P registration network for clarity. As shown in the left part of Fig. 3, the current network incorporates two feature extractors for learning images and point clouds features $\mathbf{F}_\text{I}$ and $\mathbf{F}_\text{P}$. The previous work [25] employs ResNet [16] and KPConv [33] as feature extractors for the two modalities. A key step in I2P registration is post-processing. Li *et al.* [25] designed a two-stage matching scheme inspired by GeoTrans [28]. In the first stage, 2D and 3D patch features (i.e., obtained from extractors) are used for the 2D-3D patches matching. Then, for every matched patch pair, correspondences are determined using Eq. (1). Overall, φ in Eq. (2) encompasses two feature extractors, and L_{corr} is detailed in literature [25, 35].

4.3.2. Keypoints and Chamfer loss computation

We provide the computational detail of $L_{\text{key}}(q)$ in Eqs. (13-15). By computing the L2 distance between each 2D keypoint and each 3D point feature, we obtain a $M_0 \times N$ distance matrix $\mathbf{D} = (d_{ij})$ with $d_{ij} = d(\mathbf{f}_i^{2\text{D}}, \mathbf{f}_j^{3\text{D}})$. Using the Pytorch API function `min`, elements in Eq. (15) are obtained. To efficiently evaluate $\mathbb{I}(\|q - \pi(\mathbf{T}_{\text{gt}} p_q^\star)\|_2^2 \leq \tau)$, we precompute an overlap mask in $\mathcal{P}$, converting $L_{\text{key}}(q)$ into a loss function based on the intersection of union (IoU) of two sets. We empirically set $s_{\text{th}} = e^{-0.4}$ to achieve optimal performance.

Next, we analyze the Chamfer loss $L_{\text{Chamfer}}(\mathbf{T}|\mathbf{K}_\text{I}, \mathbf{K}_\text{P})$ in Eq. (16). Minimizing this loss during training is computationally expensive. We predict $\mathbf{T}$ from $\mathbf{K}_\text{I}$ and $\mathbf{K}_\text{P}$ in an end-to-end manner, where $L_{\text{Chamfer}}(\mathbf{T}|\mathbf{K}_\text{I}, \mathbf{K}_\text{P})$ serves as a loss function. MinCD-Net employs the point transformer (PointTf) [41] to encode 2D and 3D keypoint features[†], and then compute the global 2D and 3D features. By concatenating these global features, we use a series of multi-layer perceptrons (MLPs) to estimate $\mathbf{T}$ [19]. Using $\mathbf{T}$, we can transform the coordinates of $\mathbf{K}_\text{P}$ and then compute the Chamfer loss $L_{\text{Chamfer}}(\mathbf{T}|\mathbf{K}_\text{I}, \mathbf{K}_\text{P})$.

4.3.3. Summary

We summarize the impact of MinCD-Net on I2P registration below. First, MinCD-Net is **robust to the noise and outliers** in the predicted correspondences, because the proposed loss functions (i.e., $L_{\text{key}}(q)$ and $L_{\text{Chamfer}}(\mathbf{T}|\mathbf{K}_\text{I}, \mathbf{K}_\text{P})$) are only related to $\mathbf{K}_\text{I}$ and $\mathbf{K}_\text{P}$. It addresses the limitations of existing differentiable PnP schemes [4, 38, 42]. Second, MinCD-Net is **effective in learning** φ. Since the pre-detected 2D keypoints can represent the 2D image, $L_{\text{key}}(q)$ ensures that the learned 3D keypoints are

[†]2D features contain pixels' 2D bearing vectors and features obtained from 2D extractor. 3D features contain points' 3D coordinates and features obtained from 3D extractors.

Table 1. I2P registration performance for cross-scene generalization on the 7-Scenes datasets. Here † represents the average metrics across the unseen scenes. MinCD-Net achieves higher IR and RR than other methods in most scenes. Bold indicates the best performance.

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
MATR	0.761	0.455	0.359	0.420	0.411	0.390	0.288	0.387
+Diff. PnP	0.753	0.462	0.364	0.427	0.424	0.402	0.285	0.394
+BPnPNet	0.747	0.492	0.397	0.476	**0.450**	0.365	0.342	0.420
+MinCD-Net	**0.816**	**0.542**	**0.424**	**0.502**	0.408	**0.416**	**0.379**	**0.445**
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
MATR	**1.000**	0.537	0.167	0.759	0.581	0.612	0.214	0.478
+Diff. PnP	**1.000**	0.556	0.184	0.767	0.585	0.622	0.226	0.490
+BPnPNet	**1.000**	0.665	0.224	0.778	**0.660**	0.601	0.142	0.512
+MinCD-Net	0.985	**0.671**	**0.250**	**0.869**	0.574	**0.619**	**0.571**	**0.592**
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
MATR	0.645	0.498	0.491	0.521	0.442	0.448	0.338	0.456
+Diff. PnP	0.653	0.502	0.497	0.532	0.439	0.457	0.351	0.463
+BPnPNet	0.666	0.554	0.565	0.473	0.472	0.454	0.389	0.486
+MinCD-Net	**0.783**	**0.660**	**0.642**	**0.536**	**0.550**	**0.546**	**0.471**	**0.568**
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
MATR	0.940	0.660	0.556	0.417	0.395	0.636	0.286	0.491
+Diff. PnP	0.947	0.672	0.559	**0.422**	0.402	0.648	0.301	0.501
+BPnPNet	0.848	0.708	**0.781**	0.144	0.660	0.750	0.429	0.578
+MinCD-Net	**0.980**	**0.769**	0.726	0.250	**0.681**	**0.810**	**0.643**	**0.647**
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
MATR	0.717	0.571	0.594	0.537	0.538	0.612	0.370	0.537
+Diff. PnP	0.723	0.576	0.602	0.545	0.546	0.627	0.382	0.546
+BPnPNet	0.693	0.562	0.557	0.530	0.562	0.576	0.409	0.532
+MinCD-Net	**0.778**	**0.617**	**0.598**	**0.540**	**0.573**	**0.636**	**0.445**	**0.568**
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
MATR	0.901	0.872	0.778	0.667	0.723	0.698	0.500	0.706
+Diff. PnP	0.918	0.885	0.783	0.685	0.741	0.703	0.532	0.722
+BPnPNet	**0.923**	**0.954**	0.849	0.650	0.717	0.830	0.714	0.785
+MinCD-Net	0.875	0.846	**0.904**	**0.683**	**0.798**	**0.872**	**0.786**	**0.814**

close to the pre-detected 2D keypoints. It ensures that the gradient $\nabla_\varphi L_{\text{Chamfer}}(\mathbf{T}|\mathbf{K}_\text{I}, \mathbf{K}_\text{P})$ is closely tied to the pixels and points representing the whole scene. Thus, the backpropagation of $L_{\text{Chamfer}}(\mathbf{T}|\mathbf{K}_\text{I}, \mathbf{K}_\text{P})$ contributes more effectively to φ compared to existing differentiable PnP schemes [4, 38, 42]. Third, MinCD-Net is **easily integrable** with existing I2P registration frameworks, as its inputs are independent of the outputs of I2P registration networks.

5. Experiments and Discussions

5.1. Configurations

To evaluate the performance of the proposed I2P registration method, we conduct experiments on multiple datasets, including RGBD-V2 [22], 7-Scenes [14], ScanNet [8], and the self-collected dataset captured by an Intel RealSense depth camera. The train-test data split for RGBD-V2 and 7-Scenes follows previous work [25], while ScanNet and self-collected datasets are totally utilized for testing. Inlier rate (IR) and registration rate (RR) are the primary evaluation metrics for I2P registration. Definitions of these metrics are provided in the appendices of [25]. The threshold of IR is 0.05m. RR@X represents the RR threshold at X meters, with a default of 0.05m. The implementation details of MinCD-Net are as follows. Its inputs include an RGB image with surface normals and an RGB point cloud with surface normals. Image surface normals are predicted using the pre-trained model DSINE [3]. The extractors in Fig. 3 are ResNet [16] and KPConv [33], where the extractor networks are similar to those in MATR [25]. The threshold s_{th} in Eq. (14) is set to $e^{-0.4}$. Point transformer in Fig. 3 is the single layer of work [41]. Its key, query, and value inputs are the 128 dimensional features which are transformed from pixels and points features. We utilize Shi-Tomasi keypoint detector to extract $\mathbf{K}_\text{I}$ that are uniformly distributed in the image. We train MinCD-Net on a single NVIDIA RTX 3080 GPU for 40 epochs. In the first 20 epochs, λ_1 and λ_2 are set to zero. According to the camera model [40], the criterion of τ is:

$$\tau \leq \left(\frac{\text{Threshold of RR} \cdot \max(f_u, f_v)}{d_{\text{max}}} \right)^2 \tag{17}$$

where f_u and f_v are camera focal lengths, d_{max} is the maximum depth. On 7-Scenes dataset [22], $f_u = f_v = 585.0$

Table 2. I2P registration performance for cross-dataset generalization on the multiple datasets, including RGBD-V2, ScanNet, and self-collected datasets. The proposed MinCD-Net outperforms other methods in most of the scenes.

IR	Kitchen→Rgbd-S1	Kitchen→Rgbd-S2	Kitchen→Rgbd-S3	Kitchen→Rgbd-S4	Kitchen→Rgbd-S5	Kitchen→Rgbd-S6	Kitchen→Rgbd-S7	Average
MATR	0.351	0.353	0.336	0.316	0.250	0.209	0.222	0.291
+Diff. PnP	0.372	0.358	0.352	0.332	0.262	0.214	0.235	0.303
+BPnPNet	0.396	0.378	0.375	0.377	0.230	0.194	0.258	0.315
+MinCD-Net	**0.427**	**0.415**	**0.405**	**0.412**	**0.310**	**0.296**	**0.329**	**0.371**
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
MATR	0.970	0.880	0.871	0.741	0.480	0.449	0.458	0.692
+Diff. PnP	0.972	0.943	0.892	0.750	0.485	0.453	0.464	0.708
+BPnPNet	0.965	0.974	0.954	0.942	0.610	0.507	0.646	0.799
+MinCD-Net	**0.974**	**0.985**	**0.968**	**0.963**	**0.707**	**0.725**	**0.711**	**0.870**
IR	Kitchen→Scan-S1	Kitchen→Scan-S2	Kitchen→Scan-S3	Kitchen→Scan-S4	Kitchen→Scan-S5	Kitchen→Scan-S6	Kitchen→Scan-S7	Average
MATR	0.495	0.550	0.424	0.337	0.507	0.434	0.414	0.451
+Diff. PnP	0.491	**0.552**	0.417	0.339	0.495	0.424	0.408	0.442
+BPnPNet	0.504	0.511	0.426	0.324	0.529	0.427	0.405	0.446
+MinCD-Net	**0.517**	0.527	**0.460**	**0.343**	**0.548**	**0.456**	**0.428**	**0.468**
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
MATR	0.956	0.954	**0.974**	0.433	0.923	0.909	0.750	0.842
+Diff. PnP	0.932	0.927	0.945	0.431	0.947	0.912	0.757	0.836
+BPnPNet	0.929	0.943	0.917	0.455	0.960	**0.923**	0.782	0.844
+MinCD-Net	**0.987**	**0.979**	0.905	**0.720**	**0.962**	0.915	**0.821**	**0.898**
IR	Kitchen→Self-S1	Kitchen→Self-S2	Kitchen→Self-S3	Kitchen→Self-S4	Kitchen→Self-S5	Kitchen→Self-S6	Kitchen→Self-S7	Average
MATR	**0.497**	0.462	0.426	**0.618**	0.507	**0.619**	0.412	0.506
+Diff. PnP	0.473	0.453	0.421	0.592	0.516	0.608	0.438	0.498
+BPnPNet	0.462	0.442	0.415	0.572	0.513	0.598	0.495	0.499
+MinCD-Net	0.485	**0.470**	**0.437**	0.581	**0.522**	0.604	**0.514**	**0.516**
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
MATR	0.556	0.389	0.333	0.976	0.532	0.964	0.278	0.575
+Diff. PnP	**0.564**	0.372	0.345	0.979	0.545	**0.966**	0.306	0.582
+BPnPNet	0.502	0.362	0.352	0.981	0.584	0.948	0.334	0.580
+MinCD-Net	0.512	**0.405**	**0.389**	**0.984**	**0.611**	0.952	**0.389**	**0.606**

Table 3. Comparison results of current methods on the RGBD-v2 dataset, evaluated with an RMSE threshold of 0.1m. † denotes that the proposed method has been pre-trained on several indoor datasets, including 7-Scene and ScanNet.

Methods	P2-Net	MATR	MATR+SN	MATR+D	MATR+Dino	FCGF	Predator	FreeReg+Kabsch	FreeReg+PnP	Diff-Reg	MinCD-Net	MinCD-Net†
IR	0.122	0.324	0.451	0.406	0.434	0.081	0.157	0.309	0.309	0.377	0.472	**0.581**
RR@0.1	0.384	0.564	0.770	0.668	0.744	0.304	0.302	0.341	0.573	0.862	0.823	**0.914**

and $d_{\max} = 10.0m$. For an RR threshold of 0.05m, τ is optimally set to 5. Besides, λ_1 and λ_2 are empirically set to 0.2 and 0.0001 for the best performance.

5.2. Methods Comparisons

Cross-scene generalization. First, we conduct the cross-scene experiment on the 7-scenes dataset [14] that contains seven independent indoor scenes. We use the notation $A \to B$ to denote training on scene A and testing on scene B. As the proposed framework falls into the category of differentiable PnP methods, we mainly compare it with two representative methods: Diff. PnP [6] and BPnPNet [4]. BPnPNet [4] is a previous work that used Blind PnP in correspondence learning. For a fair evaluation, all methods are based on the same baseline, MATR[‡] [25]. Thus, we refer to them as MATR+MinCD-Net (ours), MATR+Diff. PnP, and MATR+BPnPNet, respectively. Another classic method, P2-Net [35] is also used for comparison. The results are shown in Table 1. MATR+MinCD-Net has a significant improvement on both the IR and RR metrics than other methods if the training scene is Office. When the training scene is Chess or Kitchen, MATR+MinCD-Net also outperforms other methods, although the improvement in the IR metric is not significant. So, the proposed MinCD-Net achieves both robust and accurate performance compared to existing differentiable PnP based methods in the cross-scene setting.

Cross-dataset generalization. Next, we evaluate the differentiable PnP based methods in the cross-dataset setting. The results are shown in Table 2. On the RGBD-V2 dataset [22], the IR metric of MinCD-Net outperforms other methods. On the ScanNet dataset [8], all methods exhibit similar performance in the IR metric, but MATR+MinCD-Net learns high-quality correspondences (as seen in the RR metric for Office→Scan-s4). The self-collected dataset is the most challenging dataset, leading to poor RR metrics for all methods. Despite the challenges, our method achieves the highest average IR and RR across datasets, indicating its generalization capability.

Standard comparison. After that, we evaluate the state-of-the-art methods, including P2-Net [35], 2D3D-MATR [25], FCGF [7], Predator [18], FreeReg [36], and Diff-Reg [38] on the RGBD-V2 dataset [22]. Compared models are trained and tested using the same data split of the RGBD-V2

[‡]MATR[25] is a representative baseline for the I2P registration task.

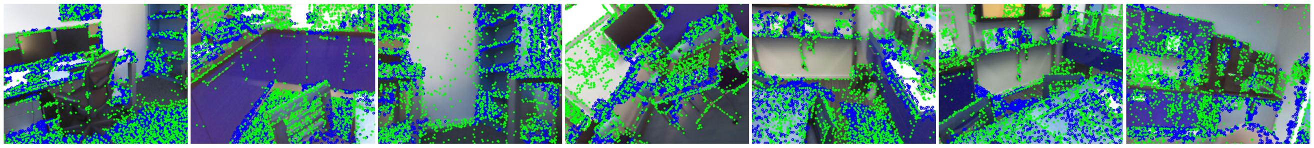

Figure 4. Visualization of pre-detected 2D keypoints (green dots) and learned 3D keypoints (blue dots). With the proposed sub-optimal learning scheme in Sec. 3.2.3, the learned 3D keypoints exhibit a large overlap with the 2D keypoints.

Table 4. Additional comparison results on the ScanNet dataset. † indicates that the model was trained on the Kitchen scene with an RR threshold of 0.05m, **stricter** than 0.3m.

Methods	P2-Net†	MATR†	LCD	Glue	FreeReg	MATR+MinCD-Net†
IR	0.303	0.451	0.307	0.184	**0.568**	0.468
RR@0.3	0.711	0.842	N/A	0.065	0.780	**0.898**

dataset [22]. Results are provided in Table 3. The notations +SN, +D, +Dino indicate the use of surface normals [3], monocular depth maps [39], and the pre-trained Dino v2 backbone [11], respectively. Similarly, +Kabsch and +PnP respectively denote the use of Kabsch [20] and EPnP [23] for outlier removal. Diff-Reg [38] exploits the EPro-PnP [6] in the correspondence learning. MATR+MinCD-Net outperforms existing methods. We include an additional comparison on the ScanNet dataset [8] with other methods, like LCD [27], Superglue (Glue) [30], and FreeReg [36]. Results are shown in Table 4. Even under stricter RR thresholds, MinCD-Net consistently outperforms FreeReg [36].

Results analysis. We analyze why MinCD-Net outperforms Diff. PnP [6] and BPnPNet [4]. Diff. PnP estimates the camera pose from the predicted correspondences. However, pose accuracy is highly sensitive to correspondence quality, making the pose loss less reliable during training. Although the declare network [15] in BPnPNet [4] is an effective module in optimizing blind PnP, it requires an accurate pose prior. In BPnPNet [4], the pose loss computed from the filtered correspondences has a limited impact on the I2P registration architecture. MinCD-Net detects and learns 2D-3D keypoints that are uniformly distributed across 2D and 3D spaces, which achieves a higher learning efficiency and is robust to correspondence quality.

5.3. Ablation studies

We conduct ablation studies to analyze the effects of the hyperparameter s_{th} and the loss functions. We analyze the relationship between s_{th} and the quality of learned 3D keypoints. As presented in Table 5, when $s_{\text{th}} \geq e^{-0.1}$, no 3D keypoints are retained. If s_{th} is set too low, a large number of redundant 3D keypoints are learned that disturb Chamfer distance minimization. To balance precision and recall, s_{th} is best set to $e^{-0.4}$, and the visualization of the learned 3D keypoints is provided in Fig. 4. Using the optimal value of s_{th}, MinCD-Net achieves the top performance across four

Table 5. Recall and precision of the learned 3D keypoints. Precision and recall are computed with respect to the pre-detected 2D keypoints (pixel threshold is 3). Avg. Num represents the average number of learned 3D keypoints.

Parameter s_{th}	$e^{-0.1}$	$e^{-0.2}$	$e^{-0.3}$	$e^{-0.4}$	$e^{-0.5}$
Precision	N/A	0.582	0.531	0.442	0.308
Recall	N/A	0.454	0.562	0.722	0.862
Avg. Num	N/A	≈3.1K	≈5.7K	≈8.4K	≈14.2K

Table 6. Ablation study of different learning schemes. The model was trained on the Office scene and tested on the remaining scenes.

Schemes	L_{corr}	$L_{\text{corr}} + L_{\text{key}}$	$L_{\text{corr}} + L_{\text{key}} + L_{\text{Chamfer}}$	**Gain**
IR	0.473	0.489	**0.567**	↑ **9.4%**
RR	0.502	0.516	**0.646**	↑ **14.4%**

datasets. This suggests that the chosen s_{th} generalizes well across different datasets. Then, we study the different loss functions in Table 6. As expected, using the combined loss $L_{\text{corr}} + L_{\text{key}}$ yields only a marginal improvement over L_{corr}, as L_{key} supervises only 3D keypoints, which are not incorporated into the network's main branch. L_{Chamfer} plays a dominant role, as it acts as a global geometrical constraint.

6. Conclusions

To improve I2P registration accuracy, we incorporate blind PnP into correspondence learning, which is achieved by simplifying blind PnP into MinCD-PnP, a more tractable task of minimizing the Chamfer distance between learned 2D and 3D keypoints. This reformulation enables efficient correspondence learning using blind PnP. To effectively solve MinCD-PnP, we develop MinCD-Net, a lightweight multi-task learning module, which can be seamlessly integrated into I2P registration networks. Extensive experiments on four indoor datasets demonstrate that MinCD-Net achieves superior performance compared to existing methods in both cross-scene and cross-dataset settings.

Acknowledgments

This work is partially supported by the National Key R&D Program of China (2024YFC3015303), National Natural Science Foundation of China (62372377), and China Postdoctoral Science Foundation (2024M761014).

References

[1] 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.*, 25 (11):15342–15366, 2024. 1, 2

[2] 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.*, 34(8): 7523–7536, 2024. 1, 2

[3] Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *Proc. CVPR*, pages 9535–9545, 2024. 6, 8

[4] Dylan Campbell, Liu Liu, and Stephen Gould. Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization. In *Proc. ECCV*, pages 244–261, 2020. 1, 2, 5, 6, 7, 8

[5] Dylan Campbell, Lars Petersson, Laurent Kneip, and Hongdong Li. Globally-optimal inlier set maximisation for camera pose and correspondence estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):328–342, 2020. 2, 3

[6] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *Proc. CVPR*, pages 2771–2780, 2022. 1, 2, 7, 8

[7] Christopher B. Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *Proc. ICCV*, pages 8957–8965, 2019. 7

[8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. CVPR*, pages 2432–2443, 2017. 2, 6, 7, 8

[9] Yaqing Ding, Jian Yang, Viktor Larsson, Carl Olsson, and Kalle Åström. Revisiting the P3P problem. In *Proc. CVPR*, pages 4872–4880, 2023. 2

[10] 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 *Proc. CVPR*, pages 8092–8101, 2019. 2

[11] Maxime Oquab et al. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. 8

[12] 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 *Proc. ICRA*, pages 4790–4796, 2019. 1, 2

[13] Luca Di Giammarino, Boyang Sun, Giorgio Grisetti, Marc Pollefeys, Hermann Blum, and Daniel Barath. Learning where to look: Self-supervised viewpoint selection for active localization using geometrical information. In *Proc. ECCV*, pages 188–205, 2024. 1

[14] Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. Real-time RGB-D camera relocalization. In *Proc. ISMAR*, pages 173–179, 2013. 2, 6, 7

[15] Stephen Gould, Richard I. Hartley, and Dylan Campbell. Deep declarative networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(8):3988–4004, 2022. 2, 8

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016. 5, 6

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. NeurIPS*, 2020. 2

[18] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *Proc. CVPR*, pages 4267–4276, 2021. 7

[19] Ganesh Iyer, Karnik Ram R., J. Krishna Murthy, and K. Madhava Krishna. Calibnet: Geometrically supervised extrinsic calibration using 3d spatial transformer networks. In *Proc. IROS*, pages 1110–1117, 2018. 5

[20] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Sec.A*, 32(5): 922–923, 1972. 8

[21] Minjung Kim, Junseo Koo, and Gunhee Kim. Ep2p-loc: End-to-end 3d point to 2d pixel localization for large-scale visual localization. In *Proc. ICCV*, pages 21470–21480, 2023. 1

[22] Kevin Lai, Liefeng Bo, and Dieter Fox. Unsupervised feature learning for 3d scene labeling. In *Proc. ICRA*, pages 3050–3057, 2014. 2, 6, 7, 8

[23] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep*np*: An accurate $O(n)$ solution to the p*np* problem. *Int. J. Comput. Vis.*, 81(2):155–166, 2009. 1, 8

[24] Jiaxin Li and Gim Hee Lee. DeepI2P: Image-to-point cloud registration via deep classification. In *Proc. CVPR*, pages 15960–15969, 2021. 2

[25] 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 *Proc. ICCV*, pages 1–10, 2023. 1, 2, 4, 5, 6, 7

[26] Yang Miao, Francis Engelmann, Olga Vysotska, Federico Tombari, Marc Pollefeys, and Dániel Béla Baráth. Scenegraphloc: Cross-modal coarse visual localization on 3d scene graphs. In *Proc. ECCV*, pages 127–150, 2024. 1

[27] 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 *Proc. AAAI*, pages 11856–11864, 2020. 8

[28] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proc. CVPR*, pages 11133–11142, 2022. 2, 5

[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

[30] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proc. CVPR*, pages 4937–4946, 2020. 8

[31] Jingnan Shi, Heng Yang, and Luca Carlone. Optimal and robust category-level perception: Object pose and shape estimation from 2-d and 3-d semantic keypoints. *IEEE Trans. Robotics*, 39(5):4131–4151, 2023. 3

[32] 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 *Proc. CVPR*, pages 6397–6406, 2020. 2

[33] 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 *Proc. ICCV*, pages 6410–6419, 2019. 5, 6

[34] 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. 1

[35] Bing Wang, Changhao Chen, Zhaopeng Cui, Jie Qin, Chris Xiaoxuan Lu, Zhengdi Yu, Peijun Zhao, Zhen Dong, Fan Zhu, Niki Trigoni, and Andrew Markham. P2-Net: Joint description and detection of local features for pixel and point matching. In *Proc. ICCV*, pages 15984–15993, 2021. 1, 2, 5, 7

[36] 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 *Proc. ICLR*, pages 1–24, 2024. 2, 7, 8

[37] Jin Wu, Yu Zheng, Zhi Gao, Yi Jiang, Xiangcheng Hu, Yilong Zhu, Jianhao Jiao, and Ming Liu. Quadratic pose estimation problems: Globally optimal solutions, solvability/observability analysis, and uncertainty description. *IEEE Trans. Robotics*, 38(5):3314–3335, 2022. 1, 3

[38] 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 *Proc. ECCV*, pages 160–178, 2024. 2, 3, 5, 6, 7, 8

[39] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proc. CVPR*, pages 10371–10381, 2024. 8

[40] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(11): 1330–1334, 2000. 3, 6

[41] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *Proc. ICCV*, pages 16239–16248, 2021. 5, 6

[42] 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 *Proc. NeurIPS*, pages 1–10, 2023. 1, 2, 3, 5, 6