

CMR-Agent: Learning a Cross-Modal Agent for Iterative Image-to-Point Cloud Registration

Gongxin Yao¹, Yixin Xuan², Xinyang Li² and Yu Pan^{1,*}

Abstract—Image-to-point cloud registration aims to determine the relative camera pose of an RGB image with respect to a point cloud. It plays an important role in camera localization within pre-built LiDAR maps. Despite the modality gaps, most learning-based methods establish 2D-3D point correspondences in feature space without any feedback mechanism for iterative optimization, resulting in poor accuracy and interpretability. In this paper, we propose to reformulate the registration procedure as an iterative Markov decision process, allowing for incremental adjustments to the camera pose based on each intermediate state. To achieve this, we employ reinforcement learning to develop a cross-modal registration agent (CMR-Agent), and use imitation learning to initialize its registration policy for stability and quick-start of the training. According to the cross-modal observations, we propose a 2D-3D hybrid state representation that fully exploits the fine-grained features of RGB images while reducing the useless neutral states caused by the spatial truncation of camera frustum. Additionally, the overall framework is well-designed to efficiently reuse one-shot cross-modal embeddings, avoiding repetitive and time-consuming feature extraction. Extensive experiments on the KITTI-Odometry and NuScenes datasets demonstrate that CMR-Agent achieves competitive accuracy and efficiency in registration. Once the one-shot embeddings are completed, each iteration only takes a few milliseconds. [code³].

I. INTRODUCTION

Accurate localization mechanisms are crucial for autonomous driving systems. Although the Global Positioning System (GPS) [1] is widely used, obstacles in real-world environments frequently disrupt the reception of satellite signals, affecting the precision and reliability of localization. To address these challenges, some methods eschew satellite communication and instead use sensors like LiDAR [2] and camera [3] to localize by comparing online sensor measurements with reference maps. Specifically, LiDAR-based methods [4] estimate relative position and rotation by registering online 3D scans with a pre-built point cloud map. They have exhibited superior performance, primarily due to the accurate distance measurements and the immunity to variations in lighting and viewing angles. However, the substantial cost of LiDAR sensors restrict their deployment on micro platforms. Camera-based methods [5] achieve localization by matching and aligning real-time RGB images with previously captured photographs. Despite the economic

This research is supported by the National Natural Science Foundation of China under Grants No. U22A20102.

¹Gongxin Yao and Yu Pan are with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, 310027, China. (Yu Pan* is the corresponding author, email: ypan@zju.edu.cn)

²Yixin Xuan and Xinyang Li are with the Polytechnic Institute, Zhejiang University, Hangzhou, 310015, China.

³<https://github.com/y2w-oc/CMR-Agent>

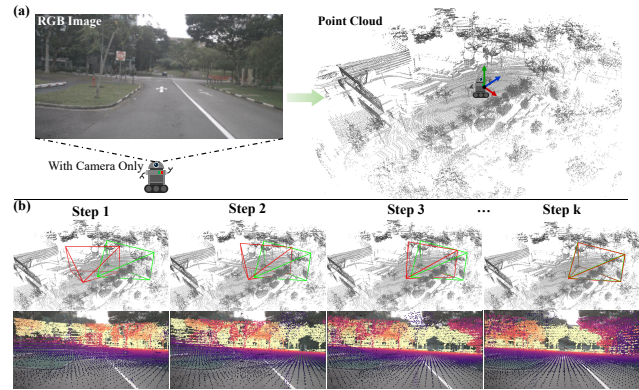


Fig. 1. Illustrations of image-to-point cloud registration. (a) A lightweight system equipped solely with a camera can localize itself by registering the online RGB image and the pre-built LiDAR map. (b) Our CMR-Agent takes the cross-modal observations as inputs, predicting actions (i.e., relative rigid transformations) to iteratively improve registration. The red and green frustums represent the estimated and actual camera poses, respectively.

and lightweight advantages of cameras, these methods are notably susceptible to variations of environmental conditions, such as lighting, weather, and season. Therefore, a question naturally arises: Is it possible to combine the strengths of both LiDAR-based and camera-based techniques by establishing direct connections between RGB images and point cloud maps? As shown in Fig.1 (a), the autonomous driving system can be equipped solely with a low-cost camera for efficient localization in pre-built LiDAR maps, reducing the negative impact of environmental changes on visual-based techniques, since LiDAR maps remain unaffected by such visual discrepancies.

The core of cross-modal localization is to estimate the relative camera pose between an RGB image and a point cloud, commonly termed as image-to-point cloud registration [6]. Most current approaches [7], [8], [9], [10] focus on establishing direct 2D-3D correspondences in feature space, where the feature matching process is conducted only once. These methods are categorized as one-shot registration due to their lack of a feedback mechanism for iterative optimization. Motivated by these concerns, we argue that mimicking the instinctive human process of aligning two objects may enhance the performance of image-to-point cloud registration. It involves iterative observation and adjustment, allowing us to continuously fine-tune the camera pose based on immediate feedback. Accordingly, we reformulate the registration problem as a Markov decision process with multi-step iterations. While reinforcement learning [11] offers an appropriate model to depict this process, migrating it to the

cross-modal context remains a complex task.

In this paper, we employ reinforcement learning to develop a **Cross-Modal Registration Agent** (CMR-Agent) that operates as illustrated in Fig. 1 (b). Our primary focuses are three-fold: 1) How to construct cross-modal state representations? 2) How to guide the training of CMR-Agent? 3) How to improve the efficiency in light of the high time complexity due to repeated iterations? For the first problem, we leverage the current camera pose to project point clouds into 2D grids for pixel-by-pixel feature comparison with RGB images. Since the camera’s field of view is a truncated segment of the 3D space, projections of point clouds from certain poses may not contain any useful information for decision (these states within 2D space are termed as neutral states in Section III-C). To supplement more information, we predict the actual camera frustum in 3D space by point classification and compare it with the camera frustum of the current pose. The final state representation integrates the information from both 2D and 3D space. For the second problem, we use the ground-truth camera pose to back-project 2D pixels into 3D points and design a point-to-point alignment reward function. The total reinforcement loss follows the well-known proximal policy optimization (PPO) algorithm [12], and imitation learning [13] is utilized to initialize the registration policy for quick-start of the training. For the last problem, we observe that a smaller relative pose does not lead to better feature embeddings for point clouds and images. Consequently, the overall framework of CMR-Agent is meticulously crafted to perform one-shot cross-modal embeddings, which are efficiently reused during iteration. It circumvents the repetitive and time-consuming feature extraction. In conclusion, the main contributions are:

- We reformulate image-to-point cloud registration as a Markov Decision process and develop a fully cross-modal registration agent using reinforcement learning and imitation learning.
- We propose a 2D-3D hybrid state representation that efficiently exploits the fine-grained features of images while reducing the useless neutral states.
- We propose a point-to-point alignment reward between the 3D points and the back-projected 2D pixels to guide the training of the agent.
- We design an efficient framework for the agent to reuse the one-shot cross-modal embeddings, decreasing the time complexity of repeated iterations.

Extensive experiments on the KITTI-Odometry [14] and NuScenes [15] datasets show that CMR-Agent outperforms the state-of-the-art methods in registration accuracy. Despite being an iterative method, CMR-Agent demonstrates higher efficiency than some one-shot algorithms, executing 10 iterations in less than 68 milliseconds on a 3090 GPU.

II. RELATED WORKS

A. Registration between Same Modal Data

Registration plays an important role in camera-based and LiDAR-based localization methods. Image-to-image registration tackles multi-view geometry by estimating the relative

pose between two images. The absence of depth information makes the epipolar constraint [16] a classic family of methods for delineating the geometric relationship between pixels across images. Besides, Optical flow [17], [18] and direct feature matching [5], [19] constitute another classic family of methods for visual pose estimation. Point cloud registration seeks to determine the rigid transformation between two point clouds, a process simplified by the presence of 3D geometric information. Methods like Iterative Closest Point (ICP) [20] and its variants [21] are the seminal works to solve this problem. In addition, learning-based feature matching [4], [22] combined with SVD decomposition [23] is also a standard paradigm for point cloud registration. While image-to-image and point cloud registration techniques have achieved satisfactory performance, LiDAR devices are not suitable for micro platforms, and cameras are not robust to changes in environmental conditions.

B. Cross-Modal Registration

Unlike high-precision camera-LiDAR extrinsic calibration [24], [25], image-to-point-cloud registration can adapt to a broader range of scenarios for localization. One of the most popular pipelines is to get 2D-3D point correspondences and then estimate the camera pose using a RANSAC-based PnP solver. 2D3D-MatchNet [7], Corri2P [8], and CFi2P [9] employ pseudo-siamese neural networks to establish the cross-modal correspondences from 3D points to 2D pixels. In addition, there are some matching-free algorithms. HyperMap [26] converts point clouds into depth maps and proposes a late projection strategy to precompute and compress the 3D map features offline. DeepI2P[6] frames the registration as a two-stage process of point-wise classification and custom pose optimization. EFGHNet[27] utilizes a divide-and-conquer approach, estimating rotation and translation separately by aligning RGB images with 360° panoramic depth maps. However, most of these methods are one-shot techniques with limited accuracy. The design of an iterative paradigm has been seldom explored.

C. Reinforcement Learning in Registration

Reinforcement learning [11] provides a framework to describe the iterative decision-making process. Recent visual pose estimation algorithms [28], [29] leverage reinforcement learning to develop agents that can register two RGB images. Specifically, these image-oriented agents employ a fully convolutional architecture for feature extraction and state representation, outputting actions to iteratively adjust the relative camera pose. As for reinforcement learning-powered point cloud registration, ReAgent [30] employs the PointNet architecture to represent states, progressively moving the source towards the target point clouds. In this study, our goal is to extend reinforcement learning to cross-modal registration between images and point clouds. Direct application of the previous single-modal state representations is not feasible for this task. Additionally, developing training strategies for the agent and improving efficiency pose greater challenges in cross-modal contexts.

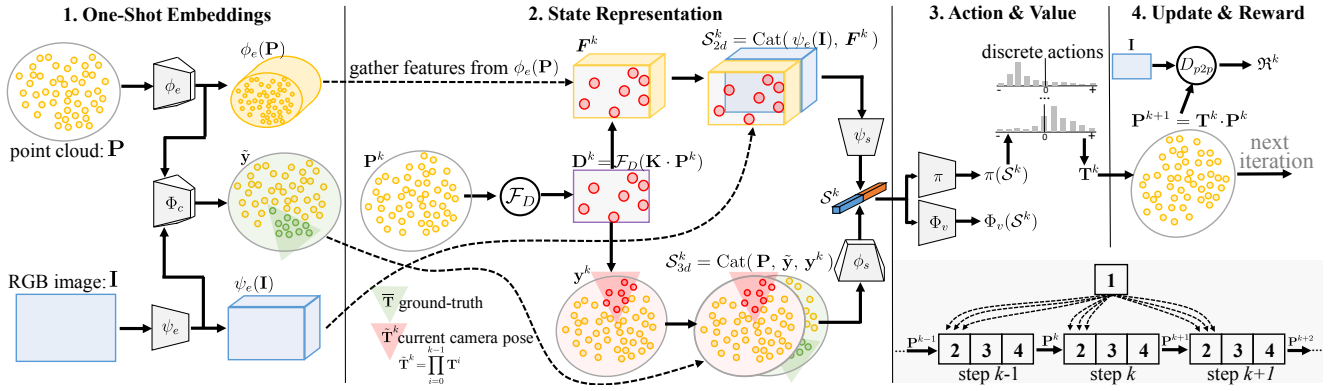


Fig. 2. The overall framework of our cross-modal registration agent (CMR-Agent). Parts 2 to 4 on the right constitute the iterative body. The dashed lines indicate the data flow for reusing one-shot embeddings. The bottom right corner is a schematic overview of the entire iterative process.

III. METHODOLOGY

A. Overview

Problem Formulation: In the context of image-to-point cloud registration, denote the scene point cloud as $\mathbf{P} \in \mathbb{R}^{N \times 3}$ with N points and the RGB image as $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ that is taken from the same scene, our goal is to estimate the camera pose $\bar{\mathbf{T}} \in SE(3)$ when the image was taken. Formally, given the camera intrinsic matrix $\mathbf{K} \in \mathbb{R}^{3 \times 3}$ and $\bar{\mathbf{T}}$, we project \mathbf{P} to image plane and get a depth map $\bar{\mathbf{D}} \in \mathbb{R}^{H \times W \times 1}$ as:

$$\begin{aligned} [x_i, y_i, z_i]^T &= \mathbf{K} \cdot \bar{\mathbf{T}} \cdot \mathbf{p}_i, \\ \mathbf{u}_i &= [x_i/z_i, y_i/z_i]^T, \\ \bar{\mathbf{D}}_{\mathbf{u}_i} &= z_i, \end{aligned} \quad (1)$$

where $\mathbf{p}_i \in \mathbf{P}$ is a 3D point and \mathbf{u}_i represents a 2D pixel. We encapsulate the entire process of deriving a depth map from the point cloud \mathbf{P} into a single function $\mathcal{F}_D(\cdot)$ as:

$$\bar{\mathbf{D}} = \mathcal{F}_D(\mathbf{K} \cdot \bar{\mathbf{T}} \cdot \mathbf{P}). \quad (2)$$

Although $\bar{\mathbf{D}}$ and \mathbf{I} represent the scene using depth and color, respectively, they are aligned to each other pixel by pixel. Besides, we encapsulate the first two steps in Eq. (1) into another function $\mathcal{F}_u(\cdot)$ for point-to-pixel projection as:

$$\mathbf{u}_i = \mathcal{F}_u(\mathbf{K} \cdot \bar{\mathbf{T}} \cdot \mathbf{p}_i). \quad (3)$$

Markov Modeling: Previous registration methods take \mathbf{P} and \mathbf{I} as the observations to derive a camera pose $\mathbf{T} \in SE(3)$ in one step. Inevitably, \mathbf{T} shows great deviation from $\bar{\mathbf{T}}$. In this study, we propose to reformulate the one-shot process as a Markov decision process, thereby progressively improving the estimated camera pose. To achieve this, we employ reinforcement learning to train a cross-modal registration agent $\mathcal{A}(\cdot)$. For every step $k \geq 0$, it takes \mathbf{P}^k and \mathbf{I} as the observations to predict a rigid transformation $\mathbf{T}^k \in SE(3)$, and \mathbf{T}^k is then applied to \mathbf{P}^k to obtain \mathbf{P}^{k+1} as:

$$\begin{aligned} \mathbf{T}^k &= \mathcal{A}(\mathbf{P}^k, \mathbf{I}), \\ \mathbf{P}^{k+1} &= \mathbf{T}^k \cdot \mathbf{P}^k, \end{aligned} \quad (4)$$

where $\mathbf{P}^0 = \mathbf{P}$. These two steps are executed alternately. The accumulation of rigid transformations before the k -th step is

the current camera pose $\tilde{\mathbf{T}}^k$, which we aim to progressively converge towards $\bar{\mathbf{T}}$ as:

$$\lim_{k \rightarrow \infty} \tilde{\mathbf{T}}^k = \lim_{k \rightarrow \infty} \prod_{i=0}^{k-1} \mathbf{T}^i = \bar{\mathbf{T}}. \quad (5)$$

Note that each \mathbf{T}^k is termed as an action at the k -th step, and the calculation of $\tilde{\mathbf{T}}^k$ follows the method of disentangled transformation [31] to avoid rotation-induced translation.

Framework: We implement $\mathcal{A}(\cdot)$ as a policy network, which uses state representation to predict the distribution of actions. $\mathcal{A}(\cdot)$ is more complex than the agents for same-modal registration [29], [30] because \mathbf{P}^k and \mathbf{I} can't share neural network weights for processing. Although projecting \mathbf{P}^k as a depth map $\mathbf{D}^k = \mathcal{F}_D(\mathbf{K} \cdot \mathbf{P}^k)$ can unify the data structure [32], the disparate visual attributes between \mathbf{D}^k and \mathbf{I} require the network be deeper to achieve sufficient expressive capability, leading to high time complexity when iterating it. To this end, we propose an efficient framework for $\mathcal{A}(\cdot)$, as illustrated in Fig.1. We first build a pseudo-siamese neural network to generate one-shot embeddings $\phi_e(\mathbf{P}) \in \mathbb{R}^{N \times f}$ and $\psi_e(\mathbf{I}) \in \mathbb{R}^{H \times W \times f}$ for \mathbf{P} and \mathbf{I} , where f is feature dimension. Instead of using \mathbf{P}^k and \mathbf{I} as the state representation \mathcal{S}^k at the k -th step, we construct \mathcal{S}^k by gathering features from $\phi_e(\mathbf{P})$ and $\psi_e(\mathbf{I})$ based on the current camera pose $\tilde{\mathbf{T}}^k$. This design reuses $\phi_e(\mathbf{P})$ and $\psi_e(\mathbf{I})$ during iterations, avoiding repetitive and time-consuming feature extraction. After that, a policy head predicts the probability distribution of actions as $\pi(\mathcal{S}^k)$ for decision, and a value head predicts a value as $\Phi_v(\mathcal{S}^k)$ to guide the policy optimization. Once the current camera pose is updated, the agent receives a reward \mathfrak{R}^k to judge how well it performs.

B. One-Shot Cross-Modal Embeddings

The goal of the networks $\phi_e(\cdot)$ and $\psi_e(\cdot)$ is to map \mathbf{P} and \mathbf{I} into a unified embedding space while making the correlated 2D pixels and 3D points closer. For this purpose, we adopt circle loss [33] to guide the training. We initially sample n 3D points that fall in the camera frustum of pose $\bar{\mathbf{T}}$. Denote the points as $\mathbf{P}^s \in \mathbb{R}^{n \times 3}$ and $\mathbf{P}^s \subset \mathbf{P}$, the corresponding

2D pixels is $\mathbf{U}^s = \{\mathbf{u}_i | \mathbf{u}_i = \mathcal{F}_u(\mathbf{K} \cdot \bar{\mathbf{T}} \cdot \mathbf{p}_i), \mathbf{p}_i \in \mathbf{P}^s\}$. For any 3D point \mathbf{p}_i , the set of positive 2D pixels is defined as $\mathcal{E}_+^i = \{\mathbf{u}_j | \mathbf{u}_j \in \mathbf{U}^s, \|\mathbf{u}_i - \mathbf{u}_j\|_2 \leq r\}$, where r is a radius. Conversely, $\mathcal{E}_-^i = \{\mathbf{u}_j | \mathbf{u}_j \in \mathbf{U}^s, \|\mathbf{u}_i - \mathbf{u}_j\|_2 > r\}$ is the negative set. Then, the circle loss anchored in 3D points is formulated as:

$$\mathcal{L}_c^{3d} = \frac{1}{n} \sum_{i=1}^n \log \left[1 + \sum_{j \in \mathcal{E}_+^i} e^{\theta_+^j (d_i^j - \Delta_+)} \cdot \sum_{k \in \mathcal{E}_-^i} e^{\theta_-^k (\Delta_- - d_i^k)} \right], \quad (6)$$

where d_i^j is the distance between \mathbf{p}_i and \mathbf{u}_j in the embedding space, and Δ_+ , Δ_- represent the positive and negative margins, respectively. The positive and negative weights are computed adaptively for each samples with $\theta_+^j = \lambda(d_i^j - \Delta_+)$ and $\theta_-^k = \lambda(\Delta_- - d_i^k)$. λ is a hyper-parameter. Similarly, we get the reverse circle loss \mathcal{L}_c^{2d} in the same way, and the total circle loss is defined as $\mathcal{L}_c = \mathcal{L}_c^{2d} + \mathcal{L}_c^{3d}$.

Additionally, we follow the previous work [6] to perform point-wise binary classification on \mathbf{P} as:

$$\tilde{\mathbf{y}} = \Phi_c(\phi_e(\mathbf{P}), \psi_e(\mathbf{I})), \quad (7)$$

where $\Phi_c(\cdot)$ is the classification head, $\tilde{\mathbf{y}} \in \mathbb{R}^{N \times 1}$ and any \tilde{y}_i indicates whether the point \mathbf{p}_i is within the camera frustum of ground-truth pose $\bar{\mathbf{T}}$. In 3D state representation, $\tilde{\mathbf{y}}$ will be used to assess the degree of frustum alignment. Note that the prediction of $\tilde{\mathbf{y}}$ is also a one-shot step. Given $\bar{\mathbf{T}}$ as a prior, we obtain the training labels via camera projection and checking whether a 3D point falls into the frustum. We adopt weighted binary cross-entropy as the loss to train $\Phi_c(\cdot)$.

C. State Representation

At every step k , we first represent the state between \mathbf{P}^k and \mathbf{I} individually in image plane and point cloud space. Then, we merge the two into a hybrid state representation.

2D State Representation: We project the point cloud \mathbf{P}^k onto the image plane as $\mathbf{D}^k = \mathcal{F}_D(\mathbf{K} \cdot \mathbf{P}^k)$, and assess the alignment between \mathbf{D}^k and \mathbf{I} . To improve the efficiency of assessment, we reuse the one-shot embeddings $\phi_e(\mathbf{P})$ and $\psi_e(\mathbf{I})$ here. Specifically, we first gather the 3D features from $\phi_e(\mathbf{P})$ to fill the grid of \mathbf{D}^k and obtain $\mathbf{F}^k \in \mathbb{R}^{H \times W \times f}$. For every pixel \mathbf{u} , we calculate $\mathbf{F}_\mathbf{u}^k \in \mathbb{R}^f$ in \mathbf{F}^k as:

$$\begin{aligned} G_\mathbf{u}^k &= \{\mathbf{p}_i \in \mathbf{P}^k | \mathbf{u} = \mathcal{F}_u(\mathbf{K} \cdot \mathbf{p}_i)\}, \\ \mathbf{F}_\mathbf{u}^k &= \sum_{\mathbf{p}_i \in G_\mathbf{u}^k} \frac{\phi_e(\mathbf{p}_i)}{|G_\mathbf{u}^k|}, \end{aligned} \quad (8)$$

where $\phi_e(\mathbf{p}_i) \in \mathbb{R}^f$ is the feature of point \mathbf{p}_i from $\phi_e(\mathbf{P})$ and $|\cdot|$ represents set cardinality. If $|G_\mathbf{u}^k| = 0$, we initialize $\mathbf{F}_\mathbf{u}^k$ as a zero vector. Then, the 2D state $\mathcal{S}_{2d}^k \in \mathbb{R}^{H \times W \times 2f}$ is represented as:

$$\mathcal{S}_{2d}^k = \text{Cat}(\psi_e(\mathbf{I}), \mathbf{F}^k), \quad (9)$$

where $\text{Cat}(\cdot)$ refers to the concatenate operation.

3D State Representation: We found that the 2D state representation is flawed. Because the camera frustum is a truncated segment of the 3D space, there are always certain

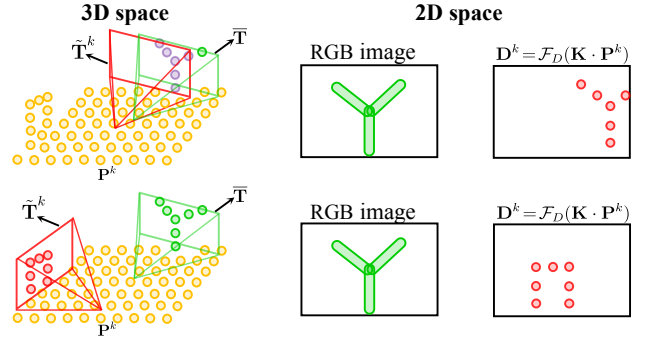


Fig. 3. Illustrations of the useful states (top) and neutral states (bottom). The top row shows a fine case where the camera frustums of $\bar{\mathbf{T}}$ and $\tilde{\mathbf{T}}^k$ overlap with each other. After projection, the geometric features in the depth map \mathbf{D}^k can indicate actions to adjust the camera pose to align with the RGB image. In the bottom row, \mathbf{D}^k does not contain any geometric information of the RGB image, which is useless for action prediction.

poses that result in bad cases, i.e., no 3D point associated with the scene in \mathbf{I} is contained in the current camera frustum, as illustrated in Fig. 3. In such cases, the 2D states fail to provide any useful information for predicting the action policy, so we term these states as **neutral states**. To supplement more information, we assess the alignment between the camera frustums of $\tilde{\mathbf{T}}^k$ and $\bar{\mathbf{T}}$ in 3D space, as illustrated in Fig. 2. The frustum of camera pose $\bar{\mathbf{T}}$ is predicted in Eq. (7). Similarly, we also represent the current camera frustum of $\tilde{\mathbf{T}}^k$ as binary classification results $\mathbf{y}^k \in \mathbb{R}^{N \times 1}$, where every y_i^k can be determined by camera projection and checking whether $\mathbf{p}_i \in \mathbf{P}^k$ falls into the current frustum. Then, the 3D state $\mathcal{S}_{3d}^k \in \mathbb{R}^{N \times 5}$ is represented as:

$$\mathcal{S}_{3d}^k = \text{Cat}(\mathbf{P}, \tilde{\mathbf{y}}, \mathbf{y}^k), \quad (10)$$

2D-3D Hybrid State Representation: To fully exploit the fine-grained features of \mathcal{S}_{2d}^k while reducing the neutral states, we propose to merge both \mathcal{S}_{2d}^k and \mathcal{S}_{3d}^k into a hybrid state representation $\mathcal{S}^k \in \mathbb{R}^{2df'}$ as:

$$\mathcal{S}^k = \text{Cat}(\psi_s(\mathcal{S}_{2d}^k), \phi_s(\mathcal{S}_{3d}^k)), \quad (11)$$

where $\psi_s(\mathcal{S}_{2d}^k) \in \mathbb{R}^{f'}$ and $\phi_s(\mathcal{S}_{3d}^k) \in \mathbb{R}^{f'}$. $\phi_s(\cdot)$ and $\psi_s(\cdot)$ are two networks that map the original 2D and 3D states to d dimensional state space.

D. Action Policy

The continuous action space couples $n_r + n_t$ degrees of freedom, e.g., $n_r = 3$ for rotation and $n_t = 3$ for translation. Following the previous method [29], we decompose the action space into $n_r + n_t$ orthogonal subspace, and predict the action in each subspace with discrete and limited step size. Taking the i -th subspace as an example, our agent predicts the optimal action from a set of candidate actions $\mathbf{A}_i = [-\Delta_j, -\Delta_{j-1}, \dots, 0, \dots, +\Delta_{j-1}, +\Delta_j]$, where $N_a = |\mathbf{A}_i|$ (i.e., N_a is $2j + 1$ here). Note that the rotation actions are expressed by the angles around different axes. The signs $+$ and $-$ indicate the direction of the action. Thus, given the state \mathcal{S}^k , our agent will predict the action policy as $\pi(\mathcal{S}^k) \in \mathbb{R}^{(n_r+n_t) \times N_a}$, which are the probability distributions

of the candidate actions in all subspace. In the i -th subspace, the final action \mathbf{a}_i^k is calculated as:

$$\mathbf{a}_i^k = \arg \max_{\mathbf{a} \in \mathbf{A}_i} \pi(\mathbf{a} | \mathcal{S}^k), \quad (12)$$

where $\pi(\mathbf{a} | \mathcal{S}^k)$ represents the probability of action \mathbf{a} from $\pi(\mathcal{S}^k)$. Finally, we adopt the disentangled transformation method [31] to convert the $n_r + n_t$ actions into \mathbf{T}^k , thereby updating the current camera pose.

E. Policy Optimization

Reward: The goal of the cross-modal registration agent is to align the image and the point cloud as described in Section III-A, therefore, the degree of alignment can serve as a criterion for evaluating the action quality. To quantify it, an intuitive approach is to calculate the average distance between the corresponding pixels in $\mathbf{D}^k = \mathcal{F}_D(\mathbf{K} \cdot \mathbf{P}^k)$ and \mathbf{I} , however, it does not work in the neutral states. Instead, we calculate an average distance in 3D space, as illustrated in Fig. 4. Denote the 3D points that fall in the camera frustum of pose $\bar{\mathbf{T}}$ as $\bar{\mathbf{P}} \in \mathbb{R}^{m \times 3}$ and $\bar{\mathbf{P}} \subset \mathbf{P}$, they are transformed as $\check{\mathbf{P}}^k = \check{\mathbf{T}}^k \cdot \bar{\mathbf{P}}$ at step k . The corresponding 2D pixels are defined as $\bar{\mathbf{U}} = \{\mathbf{u}_i | \mathbf{u}_i = \mathcal{F}_u(\mathbf{K} \cdot \bar{\mathbf{T}} \cdot \mathbf{p}_i), \mathbf{p}_i \in \bar{\mathbf{P}}\}$. We back-project $\bar{\mathbf{U}}$ to 3D space as $\check{\mathbf{P}}$ through the reverse process of Eq. (1). Then, the average distance between $\check{\mathbf{P}}$ and $\bar{\mathbf{P}}^k$ is:

$$D_{p2p}(\check{\mathbf{P}}, \bar{\mathbf{P}}^k) = \sum_{i=1}^m \frac{\|\check{\mathbf{P}}_i - \bar{\mathbf{P}}_i^k\|_2}{m}. \quad (13)$$

During the process of iterative registration, this distance will become smaller until 0. Inspired by the previous work [30], the step-wise reward is defined as:

$$\mathfrak{R}^k = \begin{cases} \varepsilon^- & \text{if } D_{p2p}(\check{\mathbf{P}}, \bar{\mathbf{P}}^{k+1}) > D_{p2p}(\check{\mathbf{P}}, \bar{\mathbf{P}}^k), \\ \varepsilon^o & \text{if } D_{p2p}(\check{\mathbf{P}}, \bar{\mathbf{P}}^{k+1}) = D_{p2p}(\check{\mathbf{P}}, \bar{\mathbf{P}}^k), \\ \varepsilon^+ & \text{if } D_{p2p}(\check{\mathbf{P}}, \bar{\mathbf{P}}^{k+1}) < D_{p2p}(\check{\mathbf{P}}, \bar{\mathbf{P}}^k), \end{cases} \quad (14)$$

where $\varepsilon^- \leq 0$ and $\varepsilon^o \leq 0$ penalize pausing and divergent actions, and $\varepsilon^+ > 0$ encourages actions that promote alignment.

Optimization: Our agent is based on actor-critic architecture [11] and we adopt the classic proximal policy optimization (PPO) algorithm [12] to optimize the policy. For this purpose, there is also a value head to predict $\Phi_v(\mathcal{S}^k)$, and the Generalized Advantage Estimation (GAE) [34] is adopted to calculate advantage based on the reward and the value. Following common practices [11], we rollout the trajectories of iterative registration by executing the stochastic policy of the agent and record them in a replay buffer. During training, we gather a batch of random samples from the buffer to update the parameters of agent. The total PPO loss consists of three components: a policy loss for optimizing the policy head $\pi(\cdot)$, a value loss for optimizing the value head $\Phi_v(\cdot)$, and an entropy loss of the policy to encourage exploration. For readers unfamiliar with PPO, we recommend referring to the original paper [12] for details.

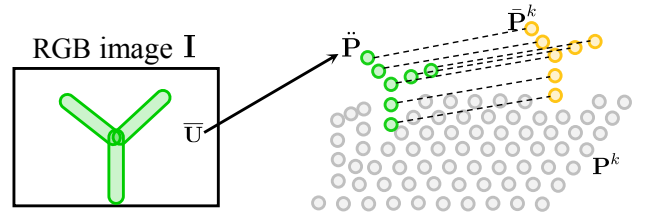


Fig. 4. Illustration of the point-to-point distance between the RGB image and the point cloud.

F. Quick-Start with Imitation Learning

Learning a cross-modal registration agent from scratch with reinforcement learning will take a long time to converge and might settle into a suboptimal policy. To address these problems, we employ imitation learning [13] to initialize the action policy. As the name suggests, the goal of imitation learning is to mimic the behavior of expert, where the expert could be either a human or another algorithmic approach. Here, we adopt the simplest form of imitation learning, Behavioral Cloning, which takes expert actions as training labels to supervise the action policy of agent. Given $\bar{\mathbf{T}}$ as a prior during training, we can easily obtain a greedy expert policy. Specifically, we first calculate the difference between $\bar{\mathbf{T}}$ and the current camera pose $\check{\mathbf{T}}^k$ on the fly as:

$$\Delta_{\mathbf{T}} = \bar{\mathbf{T}} \cdot (\check{\mathbf{T}}^k)^{-1}. \quad (15)$$

After that, we decompose $\Delta_{\mathbf{T}}$ into the optimal actions in $n_r + n_t$ subspace as described in Section III-D. In the i -th subspace, our greedy expert policy selects the action from the discrete set \mathbf{A}_i that is closest to the optimal action mentioned above. Apparently, this is a multi-classification problem, so the imitation learning loss is the cross-entropy between the agent policy $\pi(\mathcal{S}^k)$ and the greedy expert policy.

IV. EXPERIMENTS

A. Experimental Settings

Implementation Details: All the experiments are conducted on an NVIDIA RTX 3090 GPU. We implement our cross-modal agent with PyTorch 1.12, and train it using the Adam optimizer with an initial learning rate of 0.001 and a momentum of 0.98. The dimension of embedding space f is 64. In the one-shot embedding, the positive sample radius r is 1 pixel, the number of sampled points n is 512, λ is 10, Δ_+ is 0.1, and Δ_- is 1.4. The dimension of hybrid state space f' is 128. In action policy, we suggest using actions with exponential step sizes to quickly cover a large space during early iterations while allowing fine-tuning during later iterations. The set of candidate rotation actions is $0.1^\circ \times [0, \pm 1, \pm 5, \pm 25, \pm 125, \pm 625]$, the set of candidate translation actions is $0.1m \times [0, \pm 1, \pm 3, \pm 9, \pm 27, \pm 81]$, and the number of candidates in each subspace N_a is 11. In the reward function, we set $(\varepsilon^-, \varepsilon^o, \varepsilon^+)$ to $(-0.5, 0, 0.5)$ empirically. As for training, since CMR-Agent is agnostic to the one-shot cross-modal embedding, we choose to pre-train the backbone separately. Besides, we follow the work [30] to jointly optimize the PPO and imitation learning loss.

TABLE I

QUANTITATIVE COMPARISON OF REGISTRATION PERFORMANCE. THE BEST RESULTS IN EACH COLUMN ARE HIGHLIGHTED IN BOLD.

	KITTI-Odometry-09 (3182 Samples)			KITTI-Odometry-10 (2402 Samples)			NuScenes (16225 Samples)		
	RTE(m)	RRE($^{\circ}$)	RR(%)	RTE(m)	RRE($^{\circ}$)	RR(%)	RTE(m)	RRE($^{\circ}$)	RR(%)
GridCls [6]	1.199 \pm 0.724	6.000 \pm 1.989	76.39	1.063 \pm 0.566	5.744 \pm 2.033	85.22	2.224 \pm 1.094	7.370 \pm 1.647	62.67
DeepI2P(3D) [6]	1.226 \pm 0.757	6.113 \pm 2.279	39.69	1.178 \pm 0.735	6.040 \pm 2.316	36.55	1.730 \pm 0.969	7.058 \pm 2.184	18.82
DeepI2P(2D) [6]	1.435 \pm 0.918	3.805 \pm 2.660	79.73	1.389 \pm 0.862	3.990 \pm 2.727	67.57	1.949 \pm 1.050	3.017 \pm 2.285	92.53
EFGHNet [27]	3.040 \pm 1.269	4.526 \pm 2.347	24.18	3.062 \pm 1.297	4.706 \pm 2.371	23.57	3.918 \pm 1.493	5.742 \pm 3.439	31.25
CorrI2P [8]	0.919 \pm 0.742	2.417 \pm 1.988	92.65	0.878 \pm 0.689	3.126 \pm 2.333	91.59	1.697 \pm 1.030	2.318 \pm 1.868	93.87
CFI2P [9]	0.610 \pm 0.471	1.361 \pm 1.076	99.18	0.574 \pm 0.483	1.438 \pm 1.076	99.63	1.093 \pm 0.668	1.465 \pm 1.230	99.23
CMR-Agent	0.195\pm0.151	0.589\pm0.843	99.94	0.247\pm0.210	0.738\pm1.089	99.63	1.007\pm0.680	0.982\pm1.337	96.96

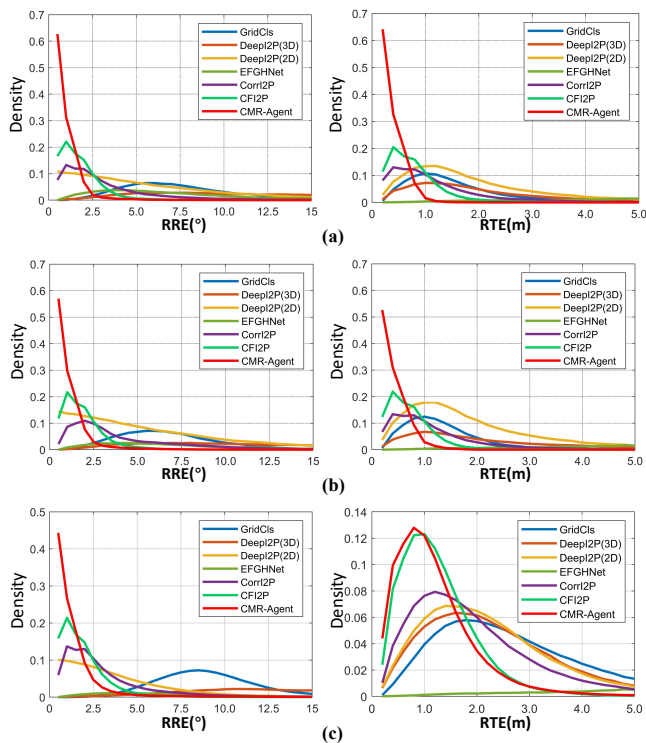


Fig. 5. Error distributions on two datasets. (a) The sequence 09 of KITTI-Odometry. (b) The sequence 10 of KITTI-Odometry. (c) NuScenes.

Dataset: We conducted experiments on KITTI-Odometry [14] and NuScenes [15] datasets. For KITTI-Odometry, sequences 0-8 are used for training, with sequences 9-10 used for testing. Each data pair consists of a single LiDAR scan as the input point cloud and an RGB image sharing the same frame ID. Each point cloud is downsampled to 40,960 points, and the image resolution is cropped to $160 \times 512 \times 3$. For the NuScenes dataset, its official split assigns 850 scenes for training and 150 scenes for testing. Here, multiple LiDAR scans proximate to the frame ID are merged to construct a more comprehensive and uniform point cloud map, alongside selecting the corresponding RGB image. Similarly, this point cloud is sampled to 40,960 points, but the image resolution is cropped to $160 \times 320 \times 3$. For a fair comparison, we follow the previous settings [8] to apply a random transformation to the point cloud, consisting of a rotation around the up-axis of up to 360° and a translation on the ground within a 10-meter range. It means that n_r and n_t are 1 and 2, respectively. Note

TABLE II

COMPUTATIONAL RESOURCES: PARAMETERS, RUNNING MEMORY, AND INFERENCE TIME. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

	Param.(MB)	Mem.(GB)	Time(s)
GridCls [6]	100.75	2.41	0.049
DeepI2P(3D) [6]	100.12	2.01	16.584
DeepI2P(2D) [6]	100.12	2.01	9.381
EFGHNet [27]	574.32	2.76	0.319
CorrI2P [8]	141.07	2.88	0.088
CFI2P [9]	34.74	2.18	0.084
CMR-Agent	34.65	1.87	0.068

that the degrees of freedom can be easily extended to 6D by adjusting the action prediction head.

Evaluation Metrics: We follow the previous work [8] to quantify the performance of image-to-point cloud registration with three metrics: 1) Relative Rotation Error (RRE), 2) Relative Translation Error (RTE), and 3) Registration Recall (RR). RR is the proportion of successful registrations within the test set, with a registration deemed successful when the RRE is less than τ_r and the RTE is less than τ_t .

B. Comparison with the State-of-the-Arts

Baselines: We compare our CMR-Agent with 6 recent algorithms, including the three variants of DeepI2P [6], EFGHNet [27], CorrI2P [8] and CFI2P [9]. Note that EFGHNet is a method based on 360° panoramic depth maps, which require good initialization, yet our setup (i.e., 10-meter translation) surpasses the range within which it can work normally. Thus, its performance significantly diverges from what was reported in the original paper [27].

Quantitative Comparison: We report the registration performance of all methods in Table I, where the results of CMR-Agent listed here are from its 10-th iteration, and the thresholds τ_r and τ_t of RR are set to 10° and 5 m, respectively. These statistical results clearly indicate that our CMR-Agent outperforms the state-of-the-art methods by a large margin. On the KITTI-Odometry dataset, CMR-Agent achieves the highest RR, with an average reduction of about 63% in RRE and about 53% in RTE compared to the suboptimal method CFI2P. On the NuScenes dataset, although RR is slightly lower than that of CFI2P, RRE is reduced by about 8% and RTE is reduced by about 33% on average. For a more comprehensive comparison, we plot the error distributions in Fig. 5, with τ_r ranging from 0 to 15° and τ_t from 0 to 5 m. The results indicate that our method

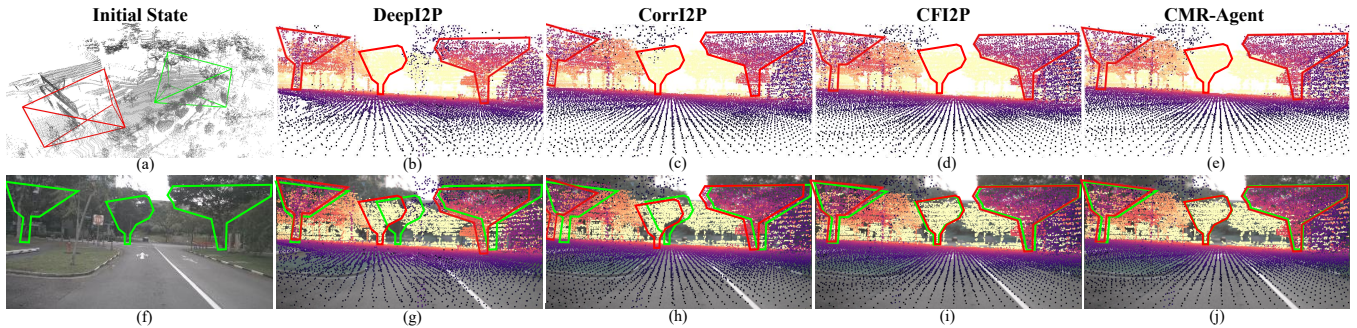


Fig. 6. Qualitative results. (a) The initial (red) and ground-truth (green) camera poses are displayed in the point cloud. (b-e) The point projection maps based on the camera poses estimated by different algorithms. (f) The input RGB image. (g-j) Overlay results of the point projection maps and the RGB image. To clearly compare the alignment, we highlight some objects under the ground-truth (green) and the estimated (red) camera poses.

TABLE III

CONTRIBUTION OF INDIVIDUAL COMPONENTS. THE RESULTS ARE AVERAGED OVER TWO SEQUENCES OF KITTI-ODOMETRY.

Method	RTE(m)	RRE($^{\circ}$)	RR(%)
w/o 3D State	0.292 ± 0.294	0.882 ± 1.176	97.64
w/o 2D State	0.526 ± 0.352	2.107 ± 1.733	99.27
w/o RL Supervision	0.260 ± 0.216	0.891 ± 1.140	99.64
Full Pipeline	0.217 ± 0.181	0.653 ± 0.959	99.80

TABLE IV

PERFORMANCE WITH DIFFERENT NUMBER OF ITERATIONS. THE RESULTS ARE AVERAGED OVER TWO SEQUENCES OF KITTI-ODOMETRY.

Iteration	RTE(m)	RRE($^{\circ}$)	RR(%)	Time(ms)
1	2.272 ± 1.141	4.346 ± 2.800	23.19	0.035
2	1.213 ± 1.008	3.604 ± 2.756	52.83	0.038
3	0.663 ± 0.637	2.649 ± 2.479	78.90	0.042
4	0.426 ± 0.427	1.844 ± 2.082	95.47	0.046
5	0.291 ± 0.261	1.138 ± 1.404	99.46	0.049
6	0.236 ± 0.193	0.814 ± 1.087	99.70	0.053
7	0.222 ± 0.186	0.691 ± 0.976	99.75	0.057
8	0.218 ± 0.187	0.679 ± 0.980	99.77	0.061
9	0.217 ± 0.178	0.658 ± 0.956	99.80	0.064
10	0.217 ± 0.181	0.653 ± 0.959	99.80	0.068

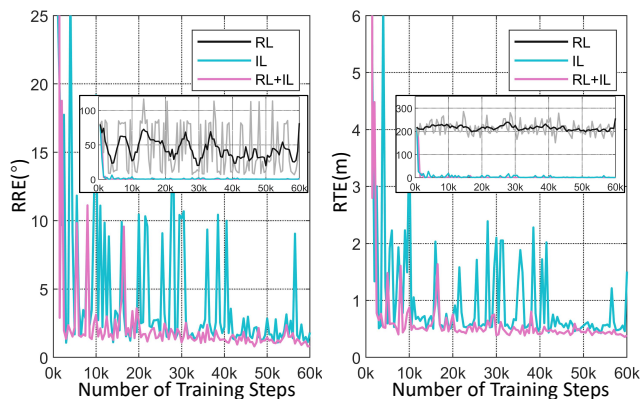


Fig. 7. Convergence of CMR-Agent with different training configurations.

maintains higher accuracy even with smaller thresholds.

Qualitative Comparison: Image-to-point cloud registration is equivalent to the process of aligning 2D pixels with their corresponding 3D points from the same object. For a more intuitive comparison, we provide a visual example in Fig. 6. Owing to our CMR-Agent’s ability to estimate relative camera poses with minimal error, it also outperforms other methods in achieving superior visual alignment.

Efficiency: The computational resources needed for all methods when executed on the KITTI-Odometry dataset are detailed in Table II. Our CMR-Agent has the minimal number of parameters and the lowest memory footprint during execution. Remarkably, as an iterative method, CMR-Agent completes 10 iterations in just 68 milliseconds, making it the second fastest even when compared to the one-shot methods.

C. Ablation studies

Individual Components: We investigate the contribution of individual components by removing them from the full

pipeline of CMR-Agent. The results are listed in Table III. By simply removing the 3D state representation, we can see that RR has dropped significantly compared to the other three rows. This serves as a significant evidence that neutral states (see Section. III-C) exist and fail to provide useful information for action decision. By simply removing the 2D state representation, we can see that RR still maintains a relatively high value, while RTE and RRE have significantly increased. This demonstrates that the 3D state supplements useful information to 2D neutral states, yet its contribution to fine registration is less than that provided by the fine-grained features of RGB images. Thus, the results in the first two rows of Table III demonstrate the necessity of the 2D-3D hybrid state representation. In addition, we also investigate the role of reinforcement learning (RL) and imitation learning (IL) by training our CMR-Agent both separately and jointly. The curves of validation error are displayed in Fig. 7. When trained solely with RL, in line with the prior experience [29], [30], the convergence process is very slow. IL promotes quick convergence but suffers from training instability and limited accuracy. The main bottleneck lies in the greedy expert policy, and theoretically, it is impossible for the agent to learn a policy better than the expert. As shown in Fig. 7 and the last two rows of Table. III, the combination of IL and RL results in training that is both rapid and stable, breaking through the accuracy limitation of IL.

Number of Iterations: We iterate CMR-Agent for different numbers of times and report the results in Table IV. Obviously, the registration performance positively correlates

with the number of iterations, and CMR-Agent only takes 3~4 ms per iteration once the one-shot embeddings are completed. A dynamic demonstration of the entire iterative process is included at the end of the supplementary video.

V. CONCLUSION

In this paper, we propose a novel cross-modal agent (CMR-Agent) for image-to-point cloud registration. Our key insight is to mimic the natural human process of aligning two objects through observation and adjustment. By reformulating the problem as a Markov decision process, we implement CMR-Agent with reinforcement learning. The proposed 2D-3D hybrid state representation fully exploits the fine-grained features of RGB images while reducing the useless neutral states. The proposed point-to-point alignment reward bridges the modality gap between images and point clouds to guide the training. The proposed framework for reusing one-shot embeddings also makes CMR-Agent very efficient. Extensive experiments demonstrate notable improvements in registration performance, showcasing the great potential of our method to achieve accurate camera localization in pre-built LiDAR maps.

REFERENCES

- [1] Bernhard Hofmann-Wellenhof, Herbert Lichtenegger, and James Collins, *Global positioning system: theory and practice*, Springer Science & Business Media, 2012.
- [2] Lucas de Paula Veronese, Fernando Auat-Cheein, Filipe Mutz, and et al., “Evaluating the limits of a lidar for an autonomous driving localization,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1449–1458, 2020.
- [3] Changhao Chen, Bing Wang, Chris Xiaoxuan Lu, and et al., “Deep learning for visual localization and mapping: A survey,” *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [4] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler, “Predator: Registration of 3d point clouds with low overlap,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4267–4276.
- [5] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, “LoftR: Detector-free local feature matching with transformers,” in *Computer Vision and Pattern Recognition*, 2021.
- [6] Jiaxin Li and Gim Hee Lee, “Deepi2p: Image-to-point cloud registration via deep classification,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15960–15969.
- [7] 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 *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4790–4796.
- [8] Siyu Ren, Yiming Zeng, Junhui Hou, and Xiaodong Chen, “Corri2p: Deep image-to-point cloud registration via dense correspondence,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1198–1208, 2023.
- [9] Gongxin Yao, Yixin Xuan, Yiwei Chen, and Yu Pan, “CFI2P: Coarse-to-fine cross-modal correspondence learning for image-to-point cloud registration,” *arXiv preprint arXiv:2307.07142*, 2023.
- [10] Shuhang Zheng, Yixuan Li, Zhu Yu, and et al., “I2p-rec: Recognizing images on large-scale point cloud maps through bird’s eye view projections,” in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 1395–1400.
- [11] Csaba Szepesvári, *Algorithms for reinforcement learning*, Springer Nature, 2022.
- [12] John Schulman, Filip Wolski, Prafulla Dhariwal, and et al., “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [13] Abdoulaye O Ly and Moulay Akhloufi, “Learning to drive by imitation: An overview of deep behavior cloning methods,” *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 2, pp. 195–209, 2020.
- [14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, “Vision meets robotics: The kitti dataset,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [15] Holger Caesar, Varun Bankiti, Alex H Lang, and et al., “nuscenes: A multimodal dataset for autonomous driving,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11621–11631.
- [16] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe, “Patch2pix: Epipolar-guided pixel-level correspondences,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4669–4678.
- [17] Mingliang Zhai, Xuezhi Xiang, Ning Lv, and Xiangdong Kong, “Optical flow and scene flow estimation: A survey,” *Pattern Recognition*, vol. 114, pp. 107861, 2021.
- [18] Yawen Lu, Qifan Wang, Siqi Ma, and et al., “Transflow: Transformer as flow learner,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18063–18073.
- [19] Tao Xie, Kun Dai, Ke Wang, and et al., “Deepmatcher: a deep transformer-based network for robust and accurate local feature matching,” *Expert Systems with Applications*, vol. 237, pp. 121361, 2024.
- [20] P.J. Besl and Neil D. McKay, “A method for registration of 3-d shapes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.
- [21] Jiaolong Yang, Hongdong Li, and Yunde Jia, “Go-icp: Solving 3d registration efficiently and globally optimally,” in *IEEE International Conference on Computer Vision*, 2013, pp. 1457–1464.
- [22] Zheng Qin, Hao Yu, Changjian Wang, and et al., “Geotransformer: Fast and robust point cloud registration with geometric transformer,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [23] Théodore Papadopoulos and Manolis IA Lourakis, “Estimating the jacobian of the singular value decomposition: Theory and applications,” in *Computer Vision-ECCV 2000: 6th European Conference on Computer Vision Dublin, Ireland, June 26–July 1, 2000 Proceedings, Part 1 6*. Springer, 2000, pp. 554–570.
- [24] Chongjian Yuan, Xiyuan Liu, Xiaoping Hong, and Fu Zhang, “Pixel-level extrinsic self calibration of high resolution lidar and camera in targetless environments,” *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7517–7524, 2021.
- [25] Xiyuan Liu, Chongjian Yuan, and Fu Zhang, “Targetless extrinsic calibration of multiple small fov lidars and cameras using adaptive voxelization,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.
- [26] Ming-Fang Chang, Joshua Mangelson, Michael Kaess, and Simon Lucey, “Hypermap: Compressed 3d map for monocular camera registration,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 11739–11745.
- [27] Yurim Jeon and Seung-Woo Seo, “EFGHNet: A versatile image-to-point cloud registration network for extreme outdoor environment,” *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7511–7517, 2022.
- [28] Alexander Krull, Eric Brachmann, Sebastian Nowozin, and et al., “Poseagent: Budget-constrained 6d object pose estimation via reinforcement learning,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [29] Jianzhun Shao, Yuhang Jiang, Gu Wang, Zhigang Li, and Xiangyang Ji, “Pfrl: Pose-free reinforcement learning for 6d pose estimation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [30] Dominik Bauer, Timothy Patten, and Markus Vincze, “Reagent: Point cloud registration using imitation and reinforcement learning,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14586–14594.
- [31] Yi Li, Gu Wang, Xiangyang Ji, and et al., “Deepim: Deep iterative matching for 6d pose estimation,” in *European Conference on Computer Vision (ECCV)*, 2018, pp. 683–698.
- [32] Daniele Cattaneo, Matteo Vaghi, Augusto Ballardini, and et al., “Cmnet: Camera to lidar-map registration,” in *2019 IEEE intelligent transportation systems conference (ITSC)*. IEEE, 2019, pp. 1283–1289.
- [33] Yifan Sun, Changmao Cheng, Yuhang Zhang, and et al., “Circle loss: A unified perspective of pair similarity optimization,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6398–6407.
- [34] John Schulman, Philipp Moritz, Sergey Levine, and et al., “High-dimensional continuous control using generalized advantage estimation,” *arXiv preprint arXiv:1506.02438*, 2015.