Global-Aware Registration of Less-Overlap RGB-D Scans

Che Sun, Yunde Jia, Yi Guo, and Yuwei Wu*
Beijing Laboratory of Intelligent Information Technology, School of Computer Science,
Beijing Institute of Technology, Beijing, 100081, China.
{sunche,jiayunde,guoyi,wuyuwei}@bit.edu.cn

Abstract

We propose a novel method of registering less-overlap RGB-D scans. Our method learns global information of a scene to construct a panorama, and aligns RGB-D scans to the panorama to perform registration. Different from existing methods that use local feature points to register less-overlap RGB-D scans and mismatch too much, we use global information to guide the registration, thereby alleviating the mismatching problem by preserving global consistency of alignments. To this end, we build a scene inference network to construct the panorama representing global information. We introduce a reinforcement learning strategy to iteratively align RGB-D scans with the panorama and refine the panorama representation, which reduces the noise of global information and preserves global consistency of both geometric and photometric alignments. Experimental results on benchmark datasets including SUNCG, Matterport, and ScanNet show the superiority of our method.

1. Introduction

Registering RGB-D scans is the basis of 3D reconstruction and 3D modeling, and has been increasingly studied [5, 9, 19, 26]. Most existing methods [27, 29] usually require a large overlap ($\geq 70\%$) to achieve good registration results. However, in practice, there will inevitably appear to be less-overlap RGB-D scans when cameras are moved suddenly and rapidly, or multiple cameras are deployed in less or no co-visible regions. Rescanning can compensate for the lack of overlap of less-overlap scans [1, 15], but it is somewhat costly and inefficient. Therefore, many researchers have been initiated to investigate directly registering less-overlap scans.

Existing methods [30, 31] use scene completion strategies and the conventional three-step paradigm (i.e., feature extraction, feature matching, and pose estimation) to register less-overlap scans. However, these methods do not work well in registration of blurred and texture-less regions that commonly appear in the completing scene images, because the local feature points they used for matching only contain local neighborhood information around the points. The local neighborhood information is often similar with less discriminative [14, 25], especially in blurred and texture-less regions. Therefore, local feature points are prone to be mismatched, and further incur incorrect pose estimation and registration. In this paper, we propose to use global information (e.g., scene layout and objects' surroundings) of a scene to guide the registration. We align the less-overlap scans with the scene globally in a jigsaw-like manner and preserve global consistency of both geometric and photometric alignments, thereby alleviating the problem caused by less discriminative local feature points.

Using global information to register less-overlap scans is non-trivial. Since the global information is acquired only based on less-overlap RGB-D scans and their completion, much noise will be produced from the unaligned scans and unreliable completion. In particular, we have to face the chicken-and-egg problem: global information relies on good alignments of scans, and aligning scans relies on good global information. Many methods [2, 29] adopt a simple iterative strategy to solve the problem by aligning scans merely based on current global information and refining global information, iteratively. However, the simple iterative strategy ignores the impact of future refined global information on scan alignments. This greedy strategy may lead to suboptimal solutions of alignments, thus obtaining global information with much noise. The noise degrades the fidelity of global information, remaining a significant challenge in the registration of less-overlap scans.

To tackle the challenge, we present a global-aware registration method of less-overlap RGB-D scans by jointly reducing noise and improving alignments in a reinforcement learning process. We use reinforcement learning to align RGB-D scans with the scene on the basis of both current and future global information, and refine the information based on the alignment. Our method makes full use of global information and improves its fidelity by trial-and-error learn-

*Corresponding author

ing in a non-greedy manner. To do this, we build a scene inference network to generate the panorama. The panorama is a weighted initialization representation of the global information that represents reliable regions with less noise. We use global constraints of both photometry and geometry according to the alignment between less-overlap scans and the panorama. We introduce a reinforcement learning strategy to achieve the global constraints for refining the panorama representation and aligning scans with the refined panorama, iteratively.

We evaluate our method by both establishing correspondences and estimating relative poses between RGB-D scans with less than 10% overlap on SUNCG [23], Matterport [4], and ScanNet [6] datasets. Experimental results show that our method outperforms existing state-of-the-art methods.

2. Related Work

Registration of less-overlap scans. Registration methods [3, 12, 13, 22] of low-overlap scans can be broadly categorized into two types: geometry-based and learning-based.

The geometry-based methods assume the scene structures are known, and use traditional multiple view geometry to register scans. Hess *et al.* [11] pre-scanned the indoor scene to obtain its 3D models, and established 3D-2D correspondences to register less-overlap scans. Miyata *et al.* [18] rescanned the scene with omnidirectional cameras to obtain its panorama, and applied the 8-point algorithm to match less-overlap scans to the panorama. These methods acquire high-fidelity scenes via rescanning for registration, but it is somewhat costly and inefficient. Differently, our registration method focuses on acquiring scene structures via learning from data instead of rescanning.

The learning-based methods use deep networks to learn scene structures from data, and complete the scenes in a bottom-up way for registering. Recent works [30, 31] built generative networks to infer invisible regions of scans, and then matched local feature points[1] to both establish correspondences and estimate relative poses for registration. Different from these methods using local feature points for matching, our method makes full use of global information to guide the registration. Our method preserves global consistency of both geometric and photometric alignments, and alleviates the mismatching problem in the registration of blurred and texture-less regions that commonly appear in the learned completing scene images.

Global registration. Existing global registration methods usually use global information to construct global constraints for guiding the registration. For example, iterative closest points (ICP) [2], fast global registration (FGR) [32], and deep global registration (DCP) [5] minimize a global alignment objective of 3D geometry for relative pose estimation. Direct visual odometry [14, 33, 33] and semi-direct visual odometry [8, 10] use the constraint of global or semi-global photometric differences to track consecutive frames. These global registration methods require large overlap ($\geq$ 70%) for reliable information, and don not work well in less-overlap scans. In this paper, we present a global-aware registration method that uses global information to guide the registration of less-overlap ($\leq$ 10%) scans. We also introduce a reinforcement learning strategy to jointly reduce noise of global information and improve global alignments.

[1]The "global module" proposed in [31] still matches local feature points (i.e., SIFT feature points and center points of planar patches). The "global module" aims to use multiple sets of matching results for refinement, which is different from ensuring global consistency in our method.

3. Preliminaries

Global registration of large-overlap RGB-D scans has been well investigated. Given two RGB-D scans $\mathbf{I}_1, \mathbf{I}_2 \in \mathbb{R}^{W\times H\times 4}$, registering $\mathbf{I}_1$ and $\mathbf{I}_2$ is to solve their rigid transformation matrix $\mathcal{T} \in SE(3)$. We assume that there exists a point with the world coordinate $\mathrm{M} = [X, Y, Z]^\top$ in the scene. Its camera coordinates in $\mathbf{I}_1$ and $\mathbf{I}_2$ are $\mathrm{M}_1 = [X_1, Y_1, Z_1]^\top$ and $\mathrm{M}_2 = [X_2, Y_2, Z_2]^\top$, respectively. Its pixel image coordinates in $\mathbf{I}_1$ and $\mathbf{I}_2$ are $\mathbf{m}_1 = [u_1, v_1]^\top$ and $\mathbf{m}_2 = [u_2, v_2]^\top$, respectively. Their homogeneous coordinates are represented as $[\mathrm{M}_1; 1], [\mathrm{M}_2; 1], [\mathbf{m}_1; 1]$ and $[\mathbf{m}_2; 1]$. We assume that $\mathbf{I}_1$ and $\mathbf{I}_2$ have the same camera intrinsic matrix $\mathbf{A}$.

The popular global registration methods solve $\mathcal{T}$ by minimizing alignment errors,

$$\begin{aligned} \min_{\mathcal{T}} \sum_{\mathbf{m}_1\in\mathcal{C}_1} \|\mathbf{I}_1(\mathbf{m}_1) - \mathbf{I}_2(\mathbf{m}_2)\|_2^2, \\ s.t.,\ [\mathrm{M}_1; 1] = \mathcal{T}[\mathrm{M}_2; 1], \end{aligned} \tag{1}$$

where $\mathcal{C}_1$ is the coordinate set in the co-visible regions of $\mathbf{I}_1$ and $\mathbf{I}_2$. These methods perform well in registration of large-overlap RGB-D scans. For example, conventional methods [7, 8] solve $\mathcal{T}$ by using the gradient or Gauss-Newton algorithms, and deep methods [14, 16] regress $\mathcal{T}$ directly in deep networks attached with an additional loss function in Eq. (1). These methods, however, do not work well in registering less-overlap RGB-D scans, due to lack of sufficient correspondences for solving $\mathcal{T}$ in such RGB-D scans.

4. Method

We present a global-aware registration method that uses global information to guide the registration of less-overlap RGB-D scans. As illustrated in Fig. 1, our method learns global information to construct an initialized panorama $\mathbf{I}_\mathrm{p}^{(0)}$ in scene inference based on RGB-D scans and their initialized transformation matrices. $\mathbf{I}_\mathrm{p}^{(0)}$ provides global information for global alignments. Since both the panorama

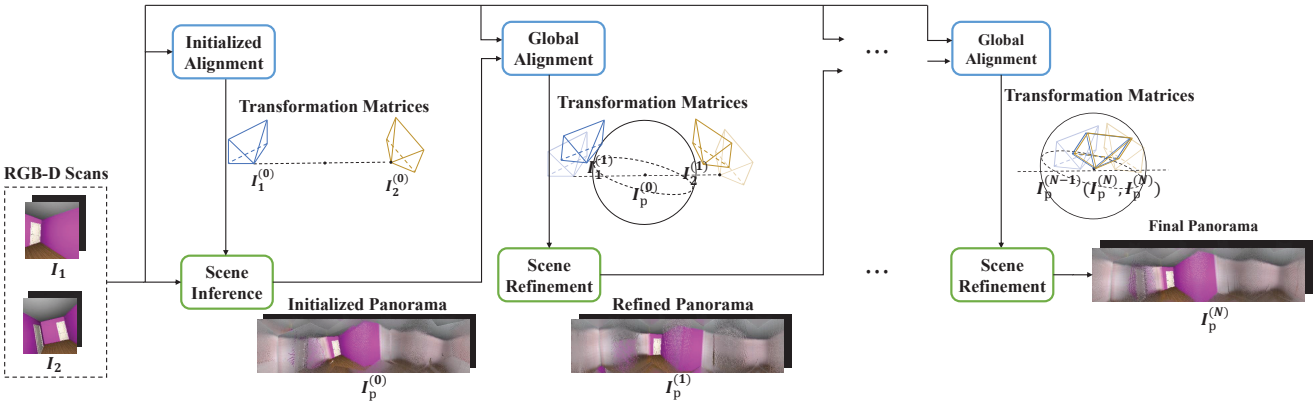

Figure 1. Overview of our global-aware registration. RGB-D scans $\mathbf{I}_1$ and $\mathbf{I}_2$ are initially aligned to obtain their initialized transformation matrices for transforming RGB-D scans as $\mathbf{I}_1^{(0)}$ and $\mathbf{I}_2^{(0)}$. The scene inference is performed to construct an initialized panorama $\mathbf{I}_\mathrm{p}^{(0)}$ that provides global information for alignments. The global alignments are used to refine the transformation matrices for transforming RGB-D scans as $\mathbf{I}_1^{(1)}$ and $\mathbf{I}_2^{(1)}$. The scene is refined to construct the refined panorama $\mathbf{I}_\mathrm{p}^{(1)}$. We iteratively perform global alignments and scene refinement for N times.

construction and global alignments form a chicken-and-egg problem, we use a reinforcement learning strategy to iteratively perform global alignments and refine the panorama. In the n-th iteration, we use the global alignment results to solve the transformation matrices for transforming RGB-D scans as $\mathbf{I}_1^{(n)}$ and $\mathbf{I}_2^{(n)}$, and we refine the panorama as $\mathbf{I}_\mathrm{p}^{(n)}$ based on the solved transformation matrices.

4.1. Problem Formulation

We use a scene inference network (described in Sec. 4.2) to construct an initialized panorama $\mathbf{I}_\mathrm{p}^{(0)} \in \mathbb{R}^{W_\mathrm{p} \times H_\mathrm{p} \times 4}$, and then solve the transformation matrices as well as refine the panorama in a reinforcement learning process (described in Sec. 4.3). We assume that the point with the world coordinate $\mathrm{M} = [X, Y, Z]^\top$ has a camera coordinate M_p in $\mathbf{I}_\mathrm{p}^{(0)}$, and its pixel image coordinate is $\mathbf{m}_\mathrm{p}$. We use the notation $\mathcal{T}_1$ to denote the transformation matrix between $\mathbf{I}_\mathrm{p}^{(0)}$ and $\mathbf{I}_1$, and use $\mathcal{T}_2$ to denote the transformation matrix between $\mathbf{I}_\mathrm{p}^{(0)}$ and $\mathbf{I}_2$.

We perform global-aware registration by converting registering $\mathbf{I}_1$ and $\mathbf{I}_2$ into jointly registering $\mathbf{I}_1$ and $\mathbf{I}_\mathrm{p}^{(0)}$ as well as registering $\mathbf{I}_2$ and $\mathbf{I}_\mathrm{p}^{(0)}$. Therefore, $\mathcal{T}$ is solved by $\mathcal{T} = \mathcal{T}_1^{-1}\mathcal{T}_2$, and Eq. (1) is converted into an equivalent form

$$\begin{aligned} \min_{\mathcal{T}_1, \mathcal{T}_2} & \sum_{\mathbf{m}_1 \in \mathcal{C}_1} \|\mathbf{I}_1(\mathbf{m}_1) - \mathbf{I}_\mathrm{p}^{(0)}(\mathbf{m}_\mathrm{p})\|_2^2 \\ & + \sum_{\mathbf{m}_2 \in \mathcal{C}_2} \|\mathbf{I}_2(\mathbf{m}_2) - \mathbf{I}_\mathrm{p}^{(0)}(\mathbf{m}_\mathrm{p})\|_2^2, \\ s.t., \ & [\mathrm{M}_\mathrm{p}; 1] = \mathcal{T}_1[\mathrm{M}_1; 1], [\mathrm{M}_\mathrm{p}; 1] = \mathcal{T}_2[\mathrm{M}_2; 1], \end{aligned} \tag{2}$$

where $\mathcal{C}_1$ and $\mathcal{C}_2$ are the coordinate sets in the respective co-visible regions.

4.2. Scene Inference

As mentioned above, we design a scene inference network to construct an initialized panorama $\mathbf{I}_\mathrm{p}^{(0)}$ and refine it as $\mathbf{I}_\mathrm{p}^{(n)}$ in the n-th iteration. The inputs include two RGB-D scans and their transformation matrices that are initialized by using the method proposed by Yang *et al.* [30] and refined in our reinforcement learning. As illustrated in Fig. 2, we use two scan completion sub-networks g_θ to extrapolate RGB-D scans, and then use the panorama inference sub-network h_ϕ to construct the panorama.

The scan completion sub-networks g_θ with shared parameters have an encoder-decoder structure with some convolutional layers. g_θ is used to obtain the extrapolated RGB-D scans that are formulated as a reduced cube-map form excluding floors and ceilings [24]. In the panorama inference sub-network h_ϕ, we first encode the extrapolated RGB-D scans in a siamese encoder, and then perform feature transforming [20] to transform the two extrapolated RGB-D scans at feature levels according to the initialized/refined transformation matrices. The two transformed features are concatenated for constructing the panorama $\mathbf{I}_\mathrm{p}^{(0)}/\mathbf{I}_\mathrm{p}^{(n)}$ in a decoder. The panorama is also formulated as the reduced cube-map form. More details about the network structures can be found in the *supplementary materials* (Supplementary Sec. S1).

The panorama construction relies on the transformation matrices $\mathcal{T}_1$ and $\mathcal{T}_2$ that are exactly what we need to solve for registration. Therefore, we initialize the transformation matrices, and refine them with a help of the reinforcement learning strategy to improve both the panorama construc-

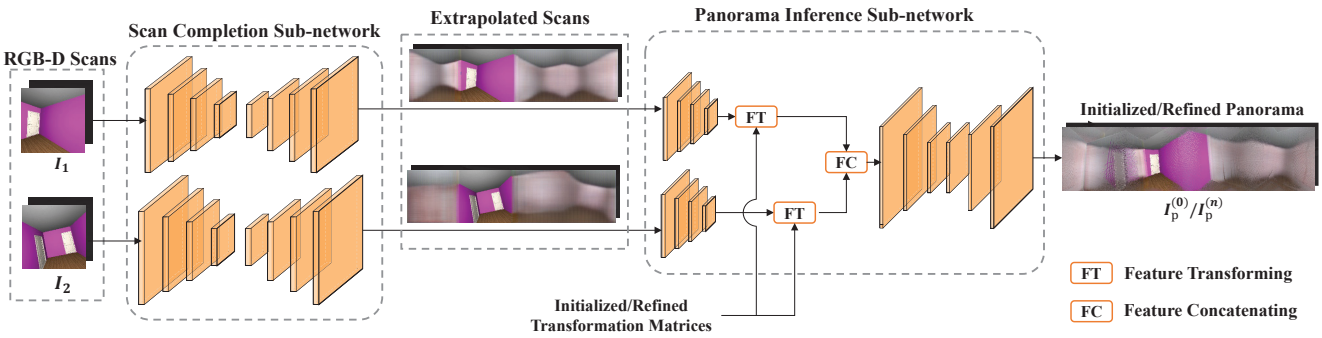

Figure 2. Illustration of the scene inference network. The scene inference network takes the inputs of RGB-D scans $\mathbf{I}_1$ and $\mathbf{I}_2$ to generate the extrapolated RGB-D scans in the scan completion sub-network. The extrapolated RGB-D scans are used to construct the initialized/refined panorama $\mathbf{I}_\text{p}^{(0)}/\mathbf{I}_\text{p}^{(n)}$ in the panorama inference sub-network, where we perform feature transforming at feature levels based on the initialized/refined transformation matrices.

tion and global alignments.

4.3. Reinforcement Learning Strategy

The goal of the reinforcement learning is to maximize the expected sum of the future discounted reward $R = \mathbb{E}[\Sigma_n \gamma_n r_n]$, where $\gamma_n \in [0, 1)$ is the discount factor, and r_n is the immediate reward at the n-th step that depends on the state s_n and the actions a_n. In the n-the iteration, we solve the transformation matrices $\mathcal{T}_1$ and $\mathcal{T}_2$ to transform the RGB-D scans as $\mathbf{I}_1^{(n)}$ and $\mathbf{I}_2^{(n)}$. The state indicates the transformed RGB-D scans $\mathbf{I}_1^{(n)}$ and $\mathbf{I}_2^{(n)}$ at the n-th iteration (refer to Sec. 4.3.1), the actions are defined as self-transformation matrices $\mathcal{T}_1^{(n)}$ and $\mathcal{T}_2^{(n)}$ for estimating transformation matrices $\mathcal{T}_1$ and $\mathcal{T}_2$ (refer to Sec. 4.3.2), and the reward is computed based on the alignment errors among $\mathbf{I}_1^{(n)}, \mathbf{I}_2^{(n)}$ and $\mathbf{I}_\text{p}^{(n)}$ (refer to Sec. 4.3.3). The transformation matrices $\mathcal{T}_1, \mathcal{T}_2$ can be solved by calculating the sequential actions $\mathcal{T}_1 = \prod_{i=1}^{n-1} \mathcal{T}_1^{(n-i)}$ and $\mathcal{T}_2 = \prod_{i=1}^{n-1} \mathcal{T}_2^{(n-i)}$ after n iterations, and the proof is given in the *supplementary materials* (Supplementary Sec. S2).

4.3.1 State

The state s_n denotes RGB-D scans' interactions with environments, which should be instrumental for the RGB-D scans to decide how to transform themselves for alignments. At the n-th iteration, the RGB-D scans $\mathbf{I}_1^{(n)}$ and $\mathbf{I}_2^{(n)}$ in the state s_n are transformed into new RGB-D scans $\mathbf{I}_1^{(n+1)}$ and $\mathbf{I}_2^{(n+1)}$ in the state s_{n+1} through the current actions a_n (i.e., the self-transformation matrices $\mathcal{T}_1^{(n)}, \mathcal{T}_2^{(n)}$). The scan transformation indicates moving the point at $\mathbf{m}_1^{(n)}$ and $\mathbf{m}_2^{(n)}$ to new coordinates $\mathbf{m}_1^{(n+1)}$ and $\mathbf{m}_2^{(n+1)}$, respectively, where $\mathbf{m}_1^{(n+1)} = \mathbf{A}\mathrm{M}_1^{(n+1)}$, $[\mathrm{M}_1^{(n+1)}; 1] = \mathcal{T}_1^{(n)}[\mathrm{M}_1^{(n)}; 1]$, and $\mathrm{M}_1^{(n)} = \mathbf{A}^{-1}\mathbf{m}_1$. $\mathbf{A}$ denotes the camera intrinsic matrix and $\mathbf{m}_2^{n+1}$ is calculated in a similar way.

4.3.2 Action

The action a_n is regarded as the rigid transformation matrices $\mathcal{T}_1^{(n)}$ and $\mathcal{T}_2^{(n)}$ at the n-th iteration. The goal of the action is to maximize the expected future reward based on the alignment errors.

We disentangle the 6D self-transformation matrices $\mathcal{T}_1^{(n)}$ and $\mathcal{T}_2^{(n)}$ as rotation matrices $\mathbf{R}_1^{(n)}, \mathbf{R}_2^{(n)} \in SO(3)$ and translation vectors $\mathbf{t}_1^{(n)}, \mathbf{t}_2^{(n)} \in \mathbb{R}^3$. The disentangled rotation and translation are not mutually affected during the prediction. We use a policy network f_π with a pre-trained embedding network e_ψ as the backbone to predict the action. The inputs to the policy network include transformed RGB-D scans $\mathbf{I}_1^{(n)}$ and $\mathbf{I}_2^{(n)}$ and the previously refined panorama $\mathbf{I}_\text{p}^{(n-1)}$. We first convert the RGB-D values to the colored point clouds, and then use the embedding network e_ψ built by a Siamese DGCNN [28] to generate point embeddings. The embeddings are fed into a cascaded two-branch network to predict distributions of the disentangled rotation $p(\mathbf{R}_1^{(n)}|s_n)$ and $p(\mathbf{R}_2^{(n)}|s_n)$ as well as the translation $p(\mathbf{t}_1^{(n)}|s_n)$ and $p(\mathbf{t}_2^{(n)}|s_n)$. The rotations $\mathbf{R}_1^{(n)}$ and $\mathbf{R}_2^{(n)}$ as well as the translations $\mathbf{t}_1^{(n)}$ and $\mathbf{t}_2^{(n)}$ are sampled from the distributions parameterized by

$$\begin{aligned}
\mathbf{R}_1^{(n)} &\sim p(\mathbf{R}_1^{(n)}|s_n) = \mathcal{N}(\mu(\mathbf{R}_1^{(n)}), \Sigma(\mathbf{R}_1^{(n)})), \\
\mathbf{R}_2^{(n)} &\sim p(\mathbf{R}_2^{(n)}|s_n) = \mathcal{N}(\mu(\mathbf{R}_2^{(n)}), \Sigma(\mathbf{R}_2^{(n)})), \\
\mathbf{t}_1^{(n)} &\sim p(\mathbf{t}_1^{(n)}|s_n) = \mathcal{N}(\mu(\mathbf{t}_1^{(n)}), \Sigma(\mathbf{t}_1^{(n)})), \\
\mathbf{t}_2^{(n)} &\sim p(\mathbf{t}_2^{(n)}|s_n) = \mathcal{N}(\mu(\mathbf{t}_2^{(n)}), \Sigma(\mathbf{t}_2^{(n)})),
\end{aligned} \tag{3}$$

where $\mathcal{N}$ refers to multivariate Gaussian distributions with mean values μ and variance matrices Σ. The mean values μ and variance matrices Σ are the outputs of the policy

network. More details about the network structure can be found in the *supplementary materials* (Supplementary Sec. S2).

4.3.3 Reward

At each iteration, a reward signal r_n is constructed for the policy update, which is regarded as the global constraint of both geometric and photometric alignments. We design a weighted reward r_n based on Eq. (2):

$$\begin{aligned} r_n =& \frac{1}{1+d_n}, \\ d_n =& \sum_{\mathbf{m}_1 \in \mathcal{C}_1} \frac{\|\mathbf{F}_1^{(n)}(\mathbf{m}_1) - \mathbf{F}_\mathrm{p}^{(n)}(\mathbf{m}_1)\|_2^2}{1+\mathbf{U}(\mathbf{m}_1)} \\ &+ \sum_{\mathbf{m}_2 \in \mathcal{C}_2} \frac{\|\mathbf{F}_2^{(n)}(\mathbf{m}_2) - \mathbf{F}_\mathrm{p}^{(n)}(\mathbf{m}_2)\|_2^2}{1+\mathbf{U}(\mathbf{m}_2)}, \end{aligned} \tag{4}$$

where $\mathbf{F}_1^{(n)}, \mathbf{F}_2^{(n)}$ and $\mathbf{F}_\mathrm{p}^{(n)}$ indicate the geometric and photometric featrure representations of $\mathbf{I}_1^{(n)}, \mathbf{I}_2^{(n)}$ and $\mathbf{I}_\mathrm{p}^{(n)}$, replacing the RGB-D values in Eq. (2) for obtaining robust results. Following the work of [30], the feature representations include specifying color, depth, normal, semantic class, and a learned descriptor. $\mathbf{U} \in \mathbb{R}^{W_\mathrm{p} \times H_\mathrm{p}}$ denotes the uncertainty maps generated by the scene inference network, increasing the importance of points in higher fidelity regions for computing the reward.

5. Network Training

There are two networks to train: the scene inference network for constructing panorama and the policy network for reinforcement learning. The scene inference network is pre-trained and its parameters are fixed when performing reinforcement learning. The policy network is optimized through both pre-training and fine-tuning for better convergence.

5.1. Scene Inference Network

The scene inference network consists of the scan completion sub-network g_θ and panorama inference sub-network h_ϕ, and they are end-to-end trained via minimizing a reconstruction loss function

$$\begin{aligned} \mathcal{L}^\mathrm{g} =& \|\mathbf{F}_1 - (\mathbf{F}_1)^*\|_F^2 + \|\mathbf{F}_2 - (\mathbf{F}_2)^*\|_F^2 + \Big\| \frac{1}{2\mathbf{U}^2} \\ &\times \mathrm{Avg}\Big(\big(\mathbf{F}_\mathrm{p} - (\mathbf{F}_\mathrm{p})^*\big)^2 \Big) + \frac{1}{2}\log(\mathbf{U}^2) \Big\|_F^2, \end{aligned} \tag{5}$$

where $(\cdot)^*$ indicates the ground-truth labels and $\mathbf{U}^2$ represents the Hadamard product between $\mathbf{U}$ and $\mathbf{U}$. $\mathrm{Avg}(\cdot)$ denotes the average pooling performed at the channel dimension (i.e., $\mathbb{R}^{W_\mathrm{p} \times H_\mathrm{p} \times D} \rightarrow \mathbb{R}^{W_\mathrm{p} \times H_\mathrm{p}}$). $\mathbf{F}_1$ and $\mathbf{F}_2$ are feature representations of the extrapolated RGB-D scans. $\mathbf{F}_\mathrm{p}$ is the feature representation of $\mathbf{I}_\mathrm{p}^{(0)}$. We use the first two F-norm terms in Eq. (5) to minimize differences between extrapolated scans and their ground-truth labels. We design the last term in Eq. (5) to simultaneously infer the panorama and measure its uncertainty by estimating parameters of a Gaussian distribution, where the mean and variance denote the panorama and its uncertainty, respectively.

5.2. Policy Network

Pre-training. The backbone (i.e., the embedding network e_ψ) is pre-trained before the reinforcement learning process. We follow the work of [27] to use the embedding network e_ψ to generate point embeddings of $\mathbf{I}_1, \mathbf{I}_2$ and $\mathbf{I}_\mathrm{p}^{(0)}$. The network is used to establish a mapping between $\mathbf{I}_1$ and $\mathbf{I}_\mathrm{p}^{(0)}$ and another mapping between $\mathbf{I}_2$ and $\mathbf{I}_\mathrm{p}^{(0)}$ based on the similarity of the embeddings. The mappings are used to estimate transformation matrices $\mathcal{T}_1$ and $\mathcal{T}_2$ in a differentiable SVD. A regression loss function is introduced to pre-train e_ψ:

$$\begin{aligned} \mathcal{L}^\mathrm{e} =& \|\mathrm{inv}(\mathbf{R}_1)(\mathbf{R}_1)^* - \mathbf{1}\|_F^2 + \|\mathbf{t}_1 - (\mathbf{t}_1)^*\|_2^2 \\ &+ \|\mathrm{inv}(\mathbf{R}_2)(\mathbf{R}_2)^* - \mathbf{1}\|_F^2 + \|\mathbf{t}_2 - (\mathbf{t}_2)^*\|_2^2, \end{aligned} \tag{6}$$

where $\mathbf{R}_1$ and $\mathbf{R}_2$ denote the predicted rotation matrices, $\mathbf{t}_1$ and $\mathbf{t}_2$ denote the predicted translation vectors, and $\mathbf{1} \in \mathbb{R}^{3\times3}$ is an identity matrix. $\mathrm{inv}(\cdot)$ is the inverse function of the matrix.

Fine-tuning. The policy network f_π with the pre-trained backbone is fine-tuned during the reinforcement learning process. The goals of the policy network include maximizing the expected discounted reward $R_n = \mathbb{E}[\sum_{j=1}^{j=n} \gamma_j r_j]$ and regressing the transformation matrices in a supervised manner. To this end, we use the proximal policy optimization (PPO) algorithm [21] to acquire the maximum reward, and use an extra supervised transformation loss function $\mathcal{L}^\mathrm{s}$ at each iteration. The supervised transformation loss function $\mathcal{L}^\mathrm{s}$ is

$$\begin{aligned} \mathcal{L}^\mathrm{s} =& \|\mathrm{inv}(\mathbf{R}_1^{(n)})(\mathbf{R}_1^{(n)})^* - \mathbf{1}\|_F^2 + \|\mathbf{t}_1^{(n)} - (\mathbf{t}_1^{(n)})^*\|_2^2 \\ &+ \|\mathrm{inv}(\mathbf{R}_2^{(n)})(\mathbf{R}_2^{(n)})^* - \mathbf{1}\|_F^2 + \|\mathbf{t}_2^{(n)} - (\mathbf{t}_2^{(n)})^*\|_2^2, \end{aligned} \tag{7}$$

where,

$$\begin{aligned} \begin{bmatrix} (\mathbf{R}_1^{(n)})^* & (\mathbf{t}_1^{(n)})^* \\ \mathbf{0}^\top & 1 \end{bmatrix} &= (\mathcal{T}_1)^* \mathrm{inv} \left(\prod_{i=1}^{n-1} \mathcal{T}_1^{(n-i)} \right), \\ \begin{bmatrix} (\mathbf{R}_2^{(n)})^* & (\mathbf{t}_2^{(n)})^* \\ \mathbf{0}^\top & 1 \end{bmatrix} &= (\mathcal{T}_2)^* \mathrm{inv} \left(\prod_{i=1}^{n-1} \mathcal{T}_2^{(n-i)} \right). \end{aligned} \tag{8}$$

For the PPO optimization algorithm, please refer to the *supplementary materials* (Supplementary Sec. S2).

6. Experiments

6.1. Datasets

We evaluate our method on three benchmark datasets: SUNCG [23], Matterport [4], and ScanNet [6]. The three datasets contain $45k$ synthetic 3D scenes, 925 real 3D scenes, and 1513 real 3D scenes. We use the same training/testing split as the work of [30]. For training, 9892 training scenes in the SUNCG dataset and all training scenes in the other two datasets are selected, where 25, 50, and 25 RGB-D scans are sampled at each scene. For testing, 1000 pairs of RGB-D scans are sampled from the scenes never seen during training.

6.2. Evaluation Metric

The evaluation strategies include the relative angular error $acos\frac{\|(\mathbf{R})^*\mathbf{R}^\top\|_\mathcal{F}}{\sqrt{2}}$ and the relative translation error $\|\mathbf{t}-(\mathbf{t})^*\|_2$, where the predicted rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{t}$ are derived from the transformation matrix $\mathcal{T}=\mathcal{T}_1^{-1}\mathcal{T}_2$, and $(\cdot)^*$ denotes the ground-truth labels. We also evaluate point correspondences $\{\mathbf{m}_1,\mathbf{m}_2|[\mathrm{M}_1;1]=\mathcal{T}[\mathrm{M}_2;1]\}$ in co-visible regions by computing the true-positive rate and recall at top-K correspondences. We sort all correspondences according to the feature representation error $\|\mathbf{F}_1(\mathbf{m_1})-\mathbf{F}_2(\mathbf{m}_2)\|_2^2$ in ascending order for obtaining top-K correspondences. If their actual Euclidean distance $\|[\mathrm{M}_1;1]-(\mathcal{T})^*[\mathrm{M}_2;1]\|_2$ in 3D space is less than $1m$, the correspondence is treated as positive, and larger than $1m$ is treated as negative.

During the evaluation, the testing RGB-D scans are divided into two categories of large-overlap and less-overlap. The large-overlap category contains scan pairs $\mathbf{I}_1$ and $\mathbf{I}_2$ that are overlapped more than 10% in terms of a ratio $o(\mathbf{I}_1,\mathbf{I}_2)=|\mathbf{I}_1\cap\mathbf{I}_2|/\min(|\mathbf{I}_1|,|\mathbf{I}_2|)$, and the less-overlap one contains the remaining scan pairs.

6.3. Results

We compare our method with several state-of-the-art methods: Super4PCS (Mellado *et al.* [17]), RobustGR (Zhou *et al.* [32]), ScanComplete (Yang *et al.* [30]), and HybridRepresentation (Yang *et al.* [31]), where the work of Yang *et al.* [30] is the baseline of our method for estimating transformation matrices between less-overlap RGB-D scans.

Comparisons on Transformation Matrices. Tab. 1 shows the quantitative comparison results between our method and some existing methods. It can be seen that the performance of our method is superior in registering less-overlap ($\leq 10\%$) RGB-D scans. Our method reduces the mean rotation/translation errors by $6.13°/0.13m$, $3.24°/0.48m$ and $11.18°/0.23m$, compared with the method [31] on the three datasets, showing the superiority of our method. When the overlapped regions are more than 10%, our method also achieves competitive results compared with these state-of-the-arts. On average, our method can achieve the best results in both real and synthetic datasets.

	SUNCG		Matterport		ScanNet	
	Rotation	Trans.	Rotation	Trans.	Rotation	Trans.
Mellado *et al.* [17] ($\geq 10\%$)	$75.18°$	$1.30m$	$46.83°$	$1.40m$	$55.01°$	$1.04m$
Zhou *et al.* [32] ($\geq 10\%$)	$41.98°$	$0.83m$	$53.85°$	$0.78m$	$49.08°$	$0.71m$
Yang *et al.* [30] ($\geq 10\%$)	$12.32°$	$0.33m$	$10.20°$	$0.27m$	$27.27°$	$0.53m$
Yang *et al.* [31] ($\geq 10\%$)	$19.40°$	$0.24m$	$\mathbf{8.15°}$	$0.29m$	$17.12°$	$0.67m$
Ours ($\geq 10\%$)	$\mathbf{10.67°}$	$\mathbf{0.24}m$	$8.29°$	$\mathbf{0.24}m$	$\mathbf{15.16°}$	$\mathbf{0.54}m$
Yang *et al.* [30] ($\leq 10\%$)	$78.80°$	$0.52m$	$87.30°$	$2.19m$	$78.95°$	$1.60m$
Yang *et al.* [31] ($\leq 10\%$)	$35.34°$	$0.50m$	$52.00°$	$1.15m$	$44.91°$	$1.00m$
Ours ($\leq 10\%$)	$\mathbf{29.21°}$	$\mathbf{0.37}m$	$\mathbf{48.76°}$	$\mathbf{0.67}m$	$\mathbf{33.73°}$	$\mathbf{0.77}m$
Yang *et al.* [30] (all)	$44.50°$	$0.65m$	$50.02°$	$1.24m$	$40.97°$	$1.09m$
Yang *et al.* [31] (all)	$31.12°$	$0.39m$	$36.07°$	$0.75m$	$24.29°$	$0.75m$
Ours (all)	$\mathbf{22.56°}$	$\mathbf{0.29}m$	$\mathbf{34.23°}$	$\mathbf{0.56}m$	$\mathbf{20.67°}$	$\mathbf{0.61}m$

Table 1. Evaluations of the relative angular error and the relative translation error of our method and baseline approaches.

	True-Positve Rate (%)			Recall (%)		
	top-30	top-50	top-100	top-30	top-50	top-100
Yang *et al.* [30]	39.1	39.7	39.0	17.6	29.8	58.5
Yang *et al.* [31]	41.0	41.1	40.4	18.6	30.3	60.8
Ours	**63.4**	**63.3**	**64.0**	**27.8**	**44.5**	**70.8**

Table 2. Comparisons of the true-positive rate and recall of correspondences on the Matterport dataset.

We convert several RGB-D scans to point clouds, and visualize the results of registering less-overlap ($\leq 10\%$) RGB-D scans in Fig. 3. We fix the green point clouds and transform the red point clouds through transformation matrices. When RGB-D scans overlap slightly, our method performs better than these state-of-the-art methods [30, 31]. For more visualization results, please refer to the *supplementary materials* (Supplementary Sec. S3).

Comparisons on Point Correspondences. We compare quantitative results of point correspondences, where RGB-D scans have less than 10% overlap regions. For fair comparisons with [30, 31], we use the extrapolated RGB-D scans, instead of original input RGB-D scans, to collect correspondences by traversing all the pixels. These pixels are converted to 3D points with the ground-truth depth for calculating the Euclidean distance.

The true-positive rate and recall on the Matterport dataset are shown in Tab. 2. Our method generates accurate correspondences in registration of the noisy RGB-D scans, with improvements of $22.2\%-23.6\%$ and $10.0\%-14.2\%$ in terms of the true-positive rate and recall compared with the method of Yang *et al.* [31]. This verifies the effectiveness of using global information for registering less-overlap scans. Preserving global consistency will improve the true-positive rate and recall of point correspondences.

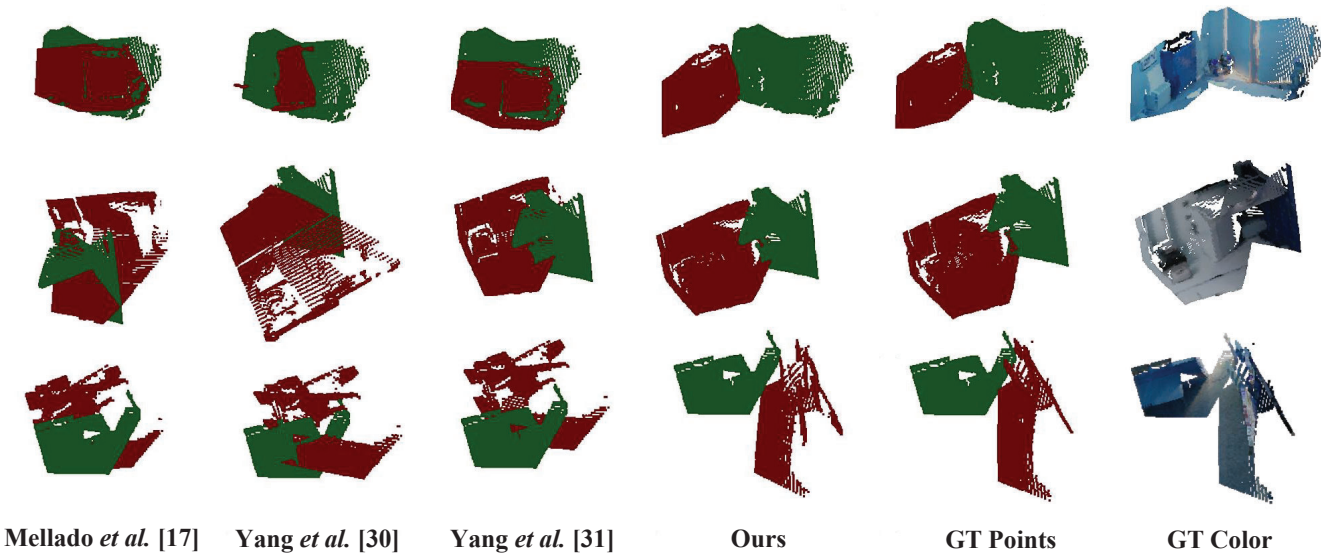

Figure 3. Qualitative results of Mellado *et al.* [17], Yang *et al.* [30], Yang *et al.* [31] and ours on the ScanNet dataset. The green point clouds are fixed and the red point clouds are transformed through predicted transformation matrices.

We also visualize point correspondences on several scenes in Fig. 4. Considering that the compared methods [30, 31] extrapolate less-overlap RGB-D scans for matching feature points, we obtain the point correspondences by transforming the extrapolated RGB-D scans with ground-truth depth in 3D spaces, and visualize the correspondences on 2D images. Fig. 4, from left to right, shows the input RGB-D scans, extrapolated RGB-D scans, correspondence results of Yang *et al.* [30], Yang *et al.* [31] and ours. Green lines indicate correct correspondences and red lines denote incorrect ones. It can be seen that our method tends to establish globally consistent correspondences based on relatively high fidelity regions, thus achieving better registration results.

6.4. Ablation Study

Analysis of Panorama Representation. As illustrated in Fig. 5, the global representation of panorama provides sufficient information of a scene for registration. To verify its effectiveness, we conduct an experiment of using extrapolated RGB-D scans $\mathbf{I}_1$ and $\mathbf{I}_2$ instead of the panorama to represent global features, where we use the same panorama inference network to obtain the extrapolated scans and panorama for a fair comparison. In the experiment, the RGB-D scan $\mathbf{I}_1$ is fixed (i.e., $\forall \mathcal{T}_1^{(n)} = \mathbf{1}$) and the RGB-D scan $\mathbf{I}_2$ is aligned towards the fixed RGB-D scan through the transformation matrix $\mathcal{T}_2 = \prod_{i=1}^{N-1} \mathcal{T}_2^{N-i}$. Experiment results of "w/o panorama" in Tab. 3 demonstrate the effectiveness of the panorama representation. The average errors on the Matterport and ScanNet datasets are reduced from $37.11°/0.60m$ and $24.33°/0.65m$ to $34.23°/0.56m$ and $20.67°/0.61m$.

	Matterport		ScanNet	
	Rotation	Trans.	Rotation	Trans.
w/o panorama	$37.11°$	$0.60m$	$24.33°$	$0.65m$
w/o weights	$40.95°$	$0.72m$	$27.42°$	$0.75m$
w/o reward	$44.25°$	$0.78m$	$28.10°$	$0.76m$
Ours	$\mathbf{34.23°}$	$\mathbf{0.56m}$	$\mathbf{20.67°}$	$\mathbf{0.61m}$

Table 3. The relative pose errors of different components of our method on the Matterport and ScanNet dataset.

Analysis of Reward. To verify the contributions of the weighted reward, we design two experiments about the reward to estimate transformation matrices. As shown in Tab. 3, "w/o weights" means that all pixels in co-visible regions contribute equally, where the uncertainty matrix $\mathbf{U}$ is a zero matrix. The average relative pose errors increase from $34.23°/0.56m$ to $40.95°/0.72m$ on the Matterport dataset and $20.67°/0.61m$ to $27.42°/0.75m$ on the ScanNet dataset. This verifies the importance of the weighted reward for guiding the alignments. "w/o reward" represents that the policy network is only optimized by the supervised regression loss function in Eq. (7) and the reward loss function is removed, which forms a direct supervised regression method via deep networks. From Tab. 3, it can be seen that the reward significantly improves the performance, reducing the average relative errors by $10.02°/0.22m$ and $7.43°/0.15m$ on the two datasets, respectively.

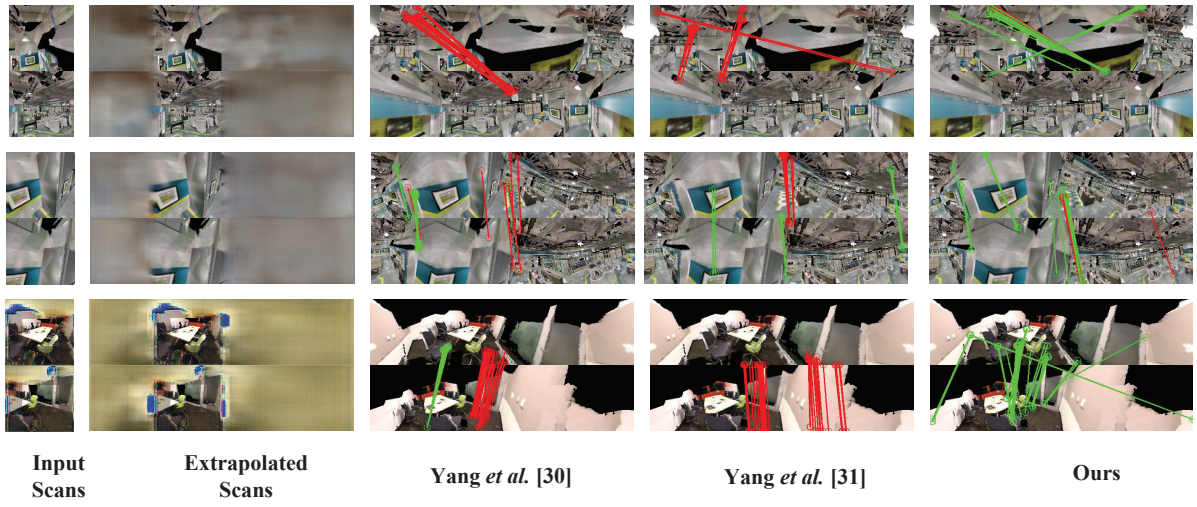

Figure 4. Visualizations of our method and baseline methods on the Matterport and ScanNet datasets. Green lines indicate correct correspondences and red lines denote incorrect ones.

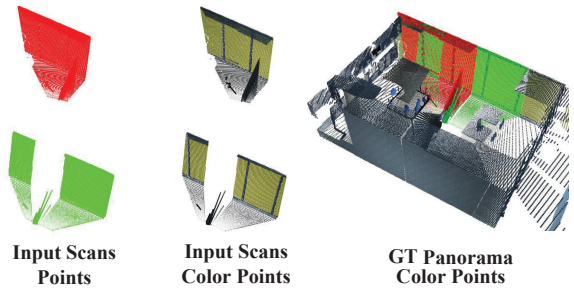

Figure 5. Visualizations of global information in an example panorama. We show the point clouds of two input scans, color points of two input scans and the ground-truth (GT) panorama.

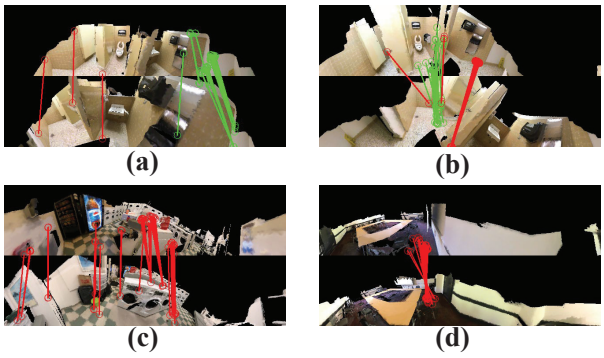

Figure 6. Several failure cases of our method. Green lines indicate correct correspondences and red lines denote incorrect ones.

6.5. Limitations

We discuss the limitations of our method by showing some failure cases in Fig. 6. (1) In an indoor scene, the occlusions are likely to cause the registration errors, as shown in Fig. 6 (a) and Fig. 6 (b). (2) When the views of two RGB-D scans change significantly, our method may be failed for registration, as represented in Fig. 6 (c). (3) The symmetrical scanning scene may mislead the registration, and a typical example is exhibited in Fig. 6 (d). These scenes are difficult to be mapped onto a single panorama for registration, and may be solved by introducing 3D global representations or multiple panoramas in the future study.

7. Conclusion

We have presented a global-aware registration method that can make full use of global information to guide the registration of less-overlap RGB-D scans. Our method can preserve global consistency of both geometric and photometric alignments for eliminating the mismatching problem caused by local feature points. We have built a panorama inference network to construct a panorama representing global information. We have also introduced a reinforcement learning strategy that can jointly reduce the noise of the global information and improve alignments in trial-and-error learning. The experiments show that our method can better register less-overlap RGB-D scans with globally consistent point correspondences.

Acknowledgments. This work was supported by the Natural Science Foundation of China (NSFC) under Grants No. 62176021 and No. 62172041.

References

[1] Esra Ataer-Cansizoglu, Yuichi Taguchi, Srikumar Ramalingam, and Yohei Miki. Calibration of non-overlapping cameras using an external slam system. In *International Conference on 3D Vision*, pages 509–516, 2014. 1

[2] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606, 1992. 1, 2

[3] Yaron Caspi and Michal Irani. Aligning non-overlapping sequences. *International Journal of Computer Vision*, 48(1):39–51, 2002. 2

[4] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *International Conference on 3D Vision*, pages 667–676, 2017. 2, 6

[5] Christopher Bongsoo Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2511–2520, 2020. 1, 2

[6] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2432–2443, 2017. 2, 6

[7] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: large-scale direct monocular SLAM. In *Proceedings of the European Conference on Computer Vision*, pages 834–849, 2014. 2

[8] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *Proceedings of the IEEE international conference on robotics and automation*, pages 15–22, 2014. 2

[9] Maciej Halber and Thomas Funkhouser. Fine-to-coarse global registration of rgb-d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1755–1764, 2017. 1

[10] Lionel Heng and Benjamin Choi. Semi-direct visual odometry for a fisheye-stereo camera. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, pages 4077–4084. IEEE, 2016. 2

[11] Wolfgang Hess, Damon Kohler, Holger Rapp, and Daniel Andor. Real-time loop closure in 2d lidar slam. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1271–1278, 2016. 2

[12] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4267–4276, 2021. 2

[13] Xiaoshui Huang, Guofeng Mei, and Jian Zhang. Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11366–11374, 2020. 2

[14] Qizeng Jia, Yuechuan Pu, Jingyu Chen, Junda Cheng, Chunyuan Liao, and Xin Yang. D2vo: Monocular deep direct visual odometry. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, pages 10158–10165, 2020. 1, 2

[15] Kenji Koide and Emanuele Menegatti. Non-overlapping rgb-d camera network calibration with monocular visual odometry. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, pages 9005–9011, 2020. 1

[16] Shunkai Li, Xin Wu, Yingdian Cao, and Hongbin Zha. Generalizing to the open world: Deep visual odometry with online adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13184–13193, 2021. 2

[17] Nicolas Mellado, Dror Aiger, and N. Mitra. Super 4pcs fast global pointcloud registration via smart indexing. *Computer Graphics Forum*, 33(5):205–215, 2014. 6, 7

[18] Shogo Miyata, Hideo Saito, Kosuke Takahashi, Dan Mikami, Mariko Isogawa, and Akira Kojima. Extrinsic camera calibration without visible corresponding points using omnidirectional cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2210–2219, 2018. 2

[19] Gabriel Moreira, Manuel Marques, and Joao Paulo Costeira. Fast pose graph optimization via krylov-schur and cholesky factorization. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1898–1906, 2021. 1

[20] C. R. Qi, H. Su, Kaichun Mo, and L. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 77–85, 2017. 3

[21] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017. 5

[22] Mohammad Amin Shabani, Weilian Song, Makoto Odamaki, Hirochika Fujiki, and Yasutaka Furukawa. Extreme structure from motion for indoor panoramas without visual overlaps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5703–5711, 2021. 2

[23] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas A. Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 190–198, 2017. 2, 6

[24] Shuran Song, Andy Zeng, Angel X. Chang, Manolis Savva, Silvio Savarese, and Thomas A. Funkhouser. Im2pano3d: Extrapolating 360° structure and semantics beyond the field of view. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3847–3856, 2018. 3

[25] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 1

[26] Zachary Teed and Jia Deng. Tangent space backpropagation for 3d transformation groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10338–10347, 2021. 1

[27] Yue Wang and Justin Solomon. Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3522–3531. IEEE, 2019. 1, 5

[28] Yue Wang, Yongbin Sun, Z. Liu, S. Sarma, M. Bronstein, and J. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics*, 38:1 – 12, 2019. 4

[29] Jiaolong Yang, Hongdong Li, Dylan Campbell, and Yunde Jia. Go-icp: A globally optimal solution to 3d icp point-set registration. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2241–2254, 2015. 1

[30] Zhenpei Yang, Jeffrey Z Pan, Linjie Luo, Xiaowei Zhou, Kristen Grauman, and Qixing Huang. Extreme relative pose estimation for rgb-d scans via scene completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4531–4540, 2019. 1, 2, 3, 5, 6, 7

[31] Zhenpei Yang, Siming Yan, and Qixing Huang. Extreme relative pose network under hybrid representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2455–2464, 2020. 1, 2, 6, 7

[32] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *Proceedings of the European Conference on Computer Vision*, pages 766–782, 2016. 2, 6

[33] Kaiying Zhu, Xiaoyan Jiang, Zhijun Fang, Yongbin Gao, Hamido Fujita, and Jenq-Neng Hwang. Photometric transfer for direct visual odometry. *Knowledge-Based Systems*, 213:106671, 2021. 2