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

Anonymous CVPR submission

Paper ID 3168

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 propose 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 [6, 11, 20, 27, 28]. Most existing methods [28, 29, 31] usually require large overlap ($\geq$ 70%) to achieve good registration results. However, in practice, there will inevitably appear to be less-overlap RGB-D scans, such as sudden and rapid camera movement, and less or no co-visible regions of multiple cameras. Rescanning can compensate for the lack of overlap of less-overlap scans [1, 16], but it is somewhat costly and inefficient. Therefore, many researchers have been initiated to investigate directly registering less-overlap scans.

Existing methods [32, 33] 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 somewhat less discriminative [26, 34], 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 mitigating 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. 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 the reinforcement learning to align RGB-D scans with the scene on the basis of the global information and refine the information based on the alignment. Our method makes full use of global information and improves its fidelity in trial-and-error learning. 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 introduce global constraints of both photometry and geometry to align less-overlap scans with the panorama for obtaining global consistency. We propose a reinforcement learning strategy to align scans with the panorama and refine the panorama representation iteratively.

We evaluate our registration method by both establishing correspondences and estimating relative poses between

RGB-D scans that have less than 10% overlap rate on SUNCG [24], Matterport [4], and ScanNet [7] datasets. Experimental results show that our method outperforms existing state-of-the-art methods, demonstrating the superiority of our method.

In summary, our contributions are two folds:

- We propose a global-aware registration method that makes full use of the global information to guide the registration of less-overlap RGB-D scans, eliminating the mismatching problem caused by local feature points by preserving global consistency.
- We introduce a reinforcement learning strategy to iteratively align less-overlap scans with the panorama and refine the panorama representation, therefore improving alignment results and reducing the noise of global information.

2. Related Work

Registration of less-overlap scans. Registration methods [3, 14, 15, 23] of low-overlap scans can be broadly categorized into two types of 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.* [13] pre-scanned the indoor scene to obtain its 3D models, and established 3D-2D correspondences to register less-overlap scans. Miyata *et al.* [19] 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 [32, 33] 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) [35], and deep global registration (DCP) [5] minimize a global alignment objective of 3D geometry for relative pose estimation. Direct visual odometry [34, 37, 37] and semi-direct visual odometry [9, 12] use the constraint of global or semi-global photometric difference 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 propose 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 [33] 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 ($\geq$ 70%) RGB-D scans has been well investigated. Given two RGB-D scans $\mathbf{I}_1, \mathbf{I}_2 \in \mathbb{R}^{W\times H\times 4}$, where W and H denote the width size and height size, respectively, 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 $[\mathbf{m}_2; 1]$, respectively. 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,

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

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 [8, 10] solve $\mathcal{T}$ in Eq. (1) by using the gradient or Gauss-Newton algorithms, and deep methods [17, 34] 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

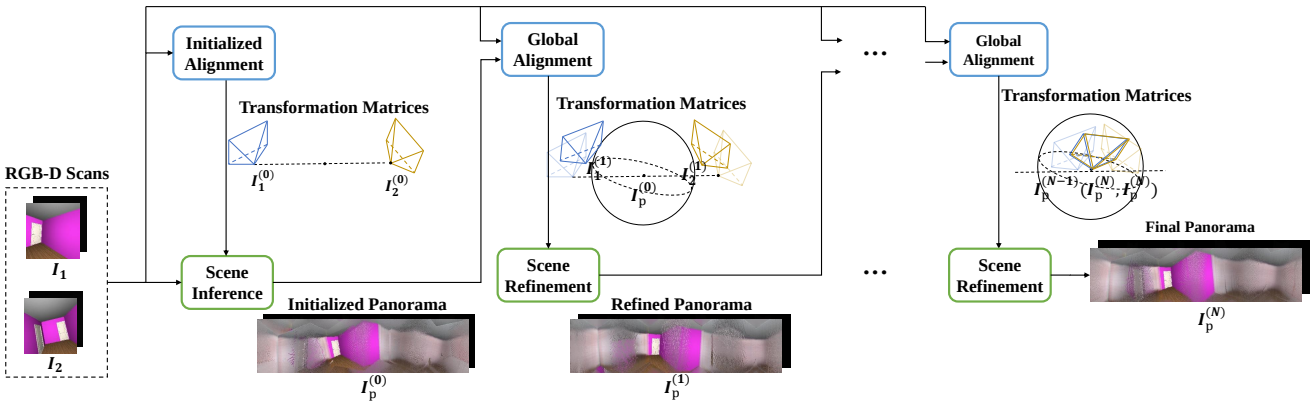

Figure 1. Overview of our global-aware registration. RGB-D scans $\mathbf{I}_1$ and $\mathbf{I}_2$ are initially aligned to obtain their initialized transformation matrices for transforming RGB-D scans as $\mathbf{I}_1^{(0)}$ and $\mathbf{I}_2^{(0)}$. The scene inference is performed to construct an 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.

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. $\mathbf{I}_\mathrm{p}^{(0)}$ provide global information for global alignments. Since both the panorama construction and global alignments form a chicken-and-egg problem, we perform 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 first 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 $\mathbf{M}_\mathrm{p}$ in $\mathbf{I}_\mathrm{p}^{(0)}$, and its pixel image coordinate is $\mathbf{m}_\mathrm{p}$. We use notation $\mathcal{T}_1$ to denote the transformation matrices between $\mathbf{I}_\mathrm{p}^{(0)}$ and $\mathbf{I}_1$, and use $\mathcal{T}_2$ to denote the transformation matrices 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 an existing method [32] and refined in our reinforcement learning. As illustrated in Fig. 2, we first 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 [25]. 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 [21] 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

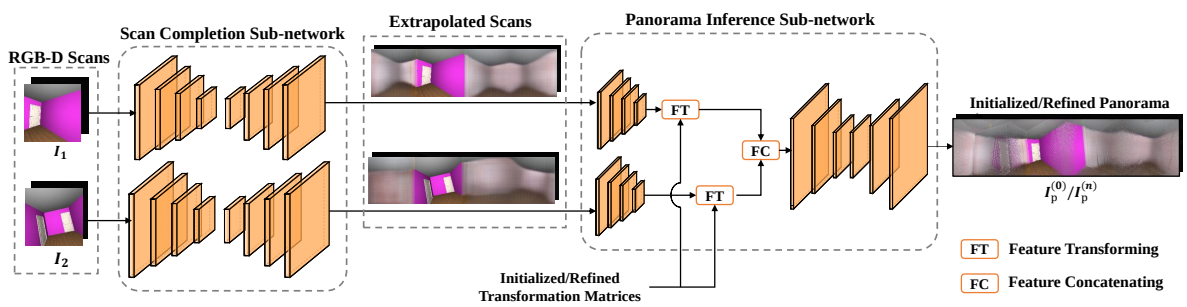

Figure 2. Illustration of the scene inference network. The scene inference network takes the inputs of RGB-D scans $\mathbf{I}_1$ and $\mathbf{I}_2$ to generate the extrapolated RGB-D scans in the scan completion sub-network. The extrapolated RGB-D scans are used to construct the initialized/refined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

$\mathbf{I}_\mathrm{p}^{(0)}/\mathbf{I}_\mathrm{p}^{(n)}$ 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 in a reinforcement learning strategy to improve both the panorama construction and global alignments.

4.3. Reinforcement Learning Process

We introduce an iterative reinforcement learning strategy to align the RGB-D scans $\mathbf{I}_1$ and $\mathbf{I}_2$ with the panorama and refine the panorama as $\mathbf{I}_\mathrm{p}^{(n)}$. 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 $\mathbf{I}_2^{(n)}$ by minimizing the alignment errors, and refine the panorama representation according to the transformation matrices.

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. We assume that 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 matrics $\mathcal{T}_1^{(n)}$ 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 $\mathbf{I}_\mathrm{p}^{(n)}$ (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 state s_n are transformed into new RGB-D scans $\mathbf{I}_1^{(n+1)}$ and $\mathbf{I}_2^{(n+1)}$ in state s_{n+1} through the current actions a_n (i.e., the self-transformation matrics $\mathcal{T}_1^{(n)}, \mathcal{T}_2^{(n)}$). 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 $\mathrm{M}_1^{(n)} = \mathbf{A}^{-1}\mathbf{m}_1$. $\mathbf{m}_2^{n+1}$ is calculated in a similar way.

4.3.2 Action

The action a_n is regarded as the phased self-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 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 $\mathbf{t}_1^{(n)}, \mathbf{t}_2^{(n)} \in \mathbb{R}^3$:

$$\mathcal{T}_1^{(n)} = \begin{bmatrix} \mathbf{R}_1^{(n)} & \mathbf{t}_1^{(n)} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \mathcal{T}_2^{(n)} = \begin{bmatrix} \mathbf{R}_2^{(n)} & \mathbf{t}_2^{(n)} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \quad (3)$$

where $\mathbf{0}^\top \in \mathbb{R}^3$ is a zero row vector. 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 of the policy network include transformed RGB-D scans $\mathbf{I}_1^{(n)}$ and $\mathbf{I}_2^{(n)}$ and the previously refined panorama $\mathbf{I}_\mathrm{p}^{(n-1)}$. 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 [30] 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

$$\begin{aligned}
\mathbf{R}_1^{(n)} &\sim p(\mathbf{R}_1^{(n)}|s_n) = \mathcal{N}(\mu(\mathbf{R}_1^{(n)}), \Sigma(\mathbf{R}_1^{(n)})),\\
\mathbf{R}_2^{(n)} &\sim p(\mathbf{R}_2^{(n)}|s_n) = \mathcal{N}(\mu(\mathbf{R}_2^{(n)}), \Sigma(\mathbf{R}_2^{(n)})),\\
\mathbf{t}_1^{(n)} &\sim p(\mathbf{t}_1^{(n)}|s_n) = \mathcal{N}(\mu(\mathbf{t}_1^{(n)}), \Sigma(\mathbf{t}_1^{(n)})),\\
\mathbf{t}_{2,n} &\sim p(\mathbf{t}_2^{(n)}|s_n) = \mathcal{N}(\mu(\mathbf{t}_2^{(n)}), \Sigma(\mathbf{t}_2^{(n)})),
\end{aligned} \tag{4}$$

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, we construct a reward signal r_n 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):

$$\begin{aligned}
r_n =& \frac{\lambda_1}{\lambda_1 + d_n},\\
d_n =& \sum_{\mathbf{m}_1 \in \mathcal{C}_1} \frac{\|\mathbf{F}_1^{(n)}(\mathbf{m}_1) - \mathbf{F}_\mathrm{p}^{(n)}(\mathbf{m}_1)\|_2^2}{1 + \lambda_2 \mathbf{U}(\mathbf{m}_1)}\\
&+ \sum_{\mathbf{m}_2 \in \mathcal{C}_2} \frac{\|\mathbf{F}_2^{(n)}(\mathbf{m}_2) - \mathbf{F}_\mathrm{p}^{(n)}(\mathbf{m}_2)\|_2^2}{1 + \lambda_2 \mathbf{U}(\mathbf{m}_2)},
\end{aligned} \tag{5}$$

where λ_1 and λ_2 are the scaling parameters for scaling the reward into a suitable interval and avoiding excessive gradients. $\mathbf{F}_1^{(n)}, \mathbf{F}_2^{(n)}$ 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). 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 that need to be trained: the scene inference network for constructing panorama and the policy network in the reinforcement learning. The scene inference network is trained offline as we need to render ground-truth panoramas offline. Its network parameters are fixed when performing reinforcement learning. The policy network is optimized through offline pre-training and online 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, where they are end-to-end trained via minimizing a reconstruction loss function

$$\begin{aligned}
\mathcal{L}^\mathrm{g} =& \|\mathbf{F}_1 - (\mathbf{F}_1)^\mathrm{gt}\|_F^2 + \|\mathbf{F}_2 - (\mathbf{F}_2)^\mathrm{gt}\|_F^2\\
&+ \Big\| \frac{1}{2} * \Big(\frac{1}{\mathbf{U}} * \frac{1}{\mathbf{U}} \Big) * \mathrm{Avg}\Big((\mathbf{F}_\mathrm{p} - (\mathbf{F}_\mathrm{p})^\mathrm{gt})\\
&* (\mathbf{F}_\mathrm{p} - (\mathbf{F}_\mathrm{p})^\mathrm{gt}) \Big) + \frac{1}{2} \log(\mathbf{U} * \mathbf{U}) \Big\|_F^2,
\end{aligned} \tag{6}$$

where $(\cdot)^\mathrm{gt}$ indicates the ground truth and $*$ denotes the Hadamard product. $\mathrm{Avg}(\cdot)$ 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 $\mathbf{I}_\mathrm{p}^{(0)}$.

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 [29] to use the embedding network e_ψ to generate point embeddings of $\mathbf{I}_1, \mathbf{I}_2$ and $\mathbf{I}_\mathrm{p}^{(0)}$, and establish two mappings between $\mathbf{I}_1$ and $\mathbf{I}_\mathrm{p}^{(0)}$ as well as $\mathbf{I}_2$ and $\mathbf{I}_\mathrm{p}^{(0)}$, respectively, based on the similarity of their 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:

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

where $\mathbf{R}_1, \mathbf{R}_2, \mathbf{t}_1$ and $\mathbf{t}_2$ denote predicted rotation matrices and translation vectors, respectively, 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 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

$$\begin{aligned}
\mathcal{L}^\mathrm{s} =& \|\mathrm{inv}(\mathbf{R}_1^{(n)})(\mathbf{R}_1^{(n)})^\mathrm{gt} - \mathbf{1}\|_F^2 + \|\mathbf{t}_1^{(n)} - (\mathbf{t}_1^{(n)})^\mathrm{gt}\|_2^2\\
&+ \|\mathrm{inv}(\mathbf{R}_2^{(n)})(\mathbf{R}_2^{(n)})^\mathrm{gt} - \mathbf{1}\|_F^2 + \|\mathbf{t}_2^{(n)} - (\mathbf{t}_2^{(n)})^\mathrm{gt}\|_2^2,
\end{aligned} \tag{8}$$

where,

$$\begin{bmatrix} (\mathbf{R}_1^{(n)})^{\mathrm{gt}} & (\mathbf{t}_1^{(n)})^{\mathrm{gt}} \\ \mathbf{0}^\top & 1 \end{bmatrix} = (\mathcal{T}_1)^{\mathrm{gt}}\mathrm{inv}\left(\prod_{i=1}^{n-1}\mathcal{T}_1^{(n-i)}\right), \\ \begin{bmatrix} (\mathbf{R}_2^{(n)})^{\mathrm{gt}} & (\mathbf{t}_2^{(n)})^{\mathrm{gt}} \\ \mathbf{0}^\top & 1 \end{bmatrix} = (\mathcal{T}_2)^{\mathrm{gt}}\mathrm{inv}\left(\prod_{i=1}^{n-1}\mathcal{T}_2^{(n-i)}\right). \tag{9}$$

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 [24], Matterport [4], and ScanNet [7]. The three datasets contain $45k$ synthetic 3D scenes, 925 real 3D scenes and 1513 real 3D scenes, respectively. We follow the work of [32] to construct experimental settings. For training, we sample 25, 50 and 25 RGB-D scans at each scene, where 9892 scenes are selected from the SUNCG dataset and all scenes in the other two datasets are used. For testing, 1000 pairs of RGB-D scans are sampled from the scenes never seen during training.

6.2. Evaluation Metric

We evaluate our method by estimating transformation matrices between RGB-D scans. The evaluation strategies include the relative angular error $acos\frac{\|(\mathbf{R})^{\mathrm{gt}}\mathbf{R}^\top\|_{\mathcal{F}}}{\sqrt{2}}$ and the relative translation error $\|\mathbf{t}-(\mathbf{t})^{\mathrm{gt}}\|_2$, 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 $(\cdot)^{\mathrm{gt}}$ denotes the ground truth. 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 $\|[\mathrm{M}_1;1]-(\mathcal{T})^{\mathrm{gt}}[\mathrm{M}_2;1]\|_2$ in 3D space is less than 1m, the correspondence is treated as positive and larger than 1m means 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 baseline methods: Super4PCS [18], RobustGR [36], ScanComp. [32], and HybridRep. [33], where ScanComp. and HybridRep. are state-of-the-art approaches for estimating transformation matrices between less-overlap RGB-D scans.

	SUNCG		Matterport		ScanNet	
	Rotation	Trans.	Rotation	Trans.	Rotation	Trans.
Super4PCS ($\geq 10\%$)	75.18°	$1.30m$	46.83°	$1.40m$	55.01°	$1.04m$
RobustGR ($\geq 10\%$)	41.98°	$0.83m$	53.85°	$0.78m$	49.08°	$0.71m$
ScanComp. ($\geq 10\%$)	12.32°	$0.33m$	10.20°	$0.27m$	27.27°	$0.53m$
HybridRep. ($\geq 10\%$)	19.40°	$0.24m$	$\mathbf{8.15^\circ}$	$0.29m$	17.12°	$0.67m$
Ours ($\geq 10\%$)	$\mathbf{10.67^\circ}$	$\mathbf{0.24}m$	8.29°	$\mathbf{0.24}m$	$\mathbf{15.16^\circ}$	$\mathbf{0.54}m$
ScanComp. ($\leq 10\%$)	78.80°	$0.52m$	87.30°	$2.19m$	78.95°	$1.60m$
HybridRep. ($\leq 10\%$)	35.34°	$0.50m$	52.00°	$1.15m$	44.91°	$1.00m$
Ours ($\leq 10\%$)	$\mathbf{29.21^\circ}$	$\mathbf{0.37}m$	$\mathbf{48.76^\circ}$	$\mathbf{0.67}m$	$\mathbf{33.73^\circ}$	$\mathbf{0.77}m$
ScanComp.(all)	44.50°	$0.65m$	50.02°	$1.24m$	40.97°	$1.09m$
HybridRep. (all)	31.12°	$0.39m$	36.07°	$0.75m$	24.29°	$0.75m$
Ours (all)	$\mathbf{22.56^\circ}$	$\mathbf{0.29}m$	$\mathbf{34.23^\circ}$	$\mathbf{0.56}m$	$\mathbf{20.67^\circ}$	$\mathbf{0.61}m$

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
ScanComp. [32]	39.1	39.7	39.0	17.6	29.8	58.5
HybridRep. [33]	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.

Comparisons on Transformation Matrices. Tab. 1 presents the quantitative comparison results between our method and some existing methods. We observe that the performance of our method is better than existing methods in registering less-overlap ($\leq 10\%$) RGB-D scans. Our method reduces the mean rotation/translation errors by $6.13^\circ/0.13m$, $3.24^\circ/0.48m$ and $11.18^\circ/0.23m$, compared with the state-of-the-art method HybridRep. [33] on the three datasets, respectively. The results prove the superiority of our method in estimating transformation matrices between less-overlap RGB-D scans. When overlapped regions are more than 10%, our method also achieves competitive results compared with these state-of-the-art methods. Overall, our method can achieve the best or comparable results in both real and synthetic datasets with superior stability and generability.

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 works [32, 33]. For more visualization results, please refer to the *supplementary materials* (Supplementary Sec. S3).

Comparisons on Point Correspondences. We compare quantitative results of point correspondences in Tab. 2, where RGB-D scans have less than 10% overlap regions. For fair comparisons with [32, 33], we use the extrapolated

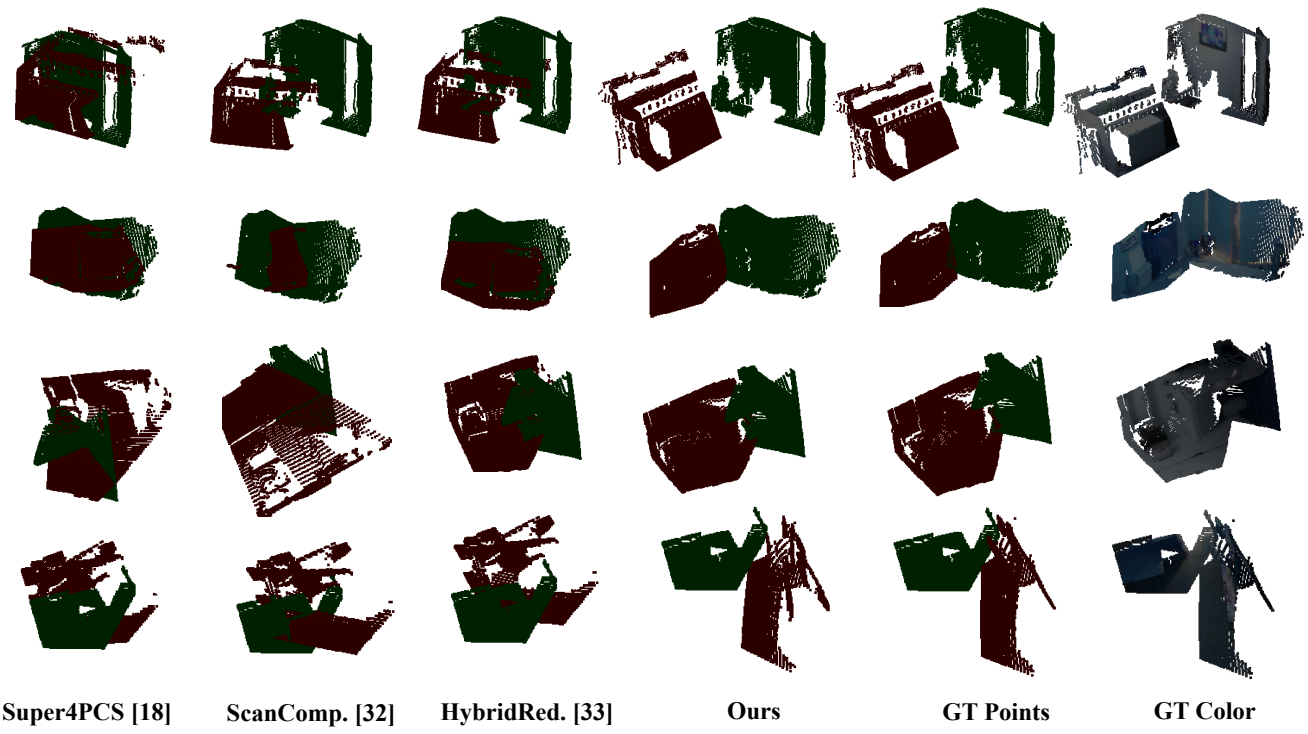

Figure 3. Qualitative results of Super4PCS [18], ScanComp. [32], HybridRep. [33], and our method on the ScanNet dataset. The green point clouds are fixed and the red point clouds are transformed through predicted transformation matrices.

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 exhibited in Tab. 2. We observe that our method generates more accurate correspondences in registration of the noisy RGB-D scans, with significant improvements of $22.2\% - 23.6\%$ and $10.0\% - 14.2\%$ in terms of the true-positive rate and recall compared with the top-performing method HybridRep. [33]. 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.

We also visualize point correspondences on several scenes in Fig. 4. Considering that the compared methods [32, 33] 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. From left to right, Fig. 4 shows the ground-truth extrapolated RGB-D scans, correspondence results of ScanComp., HybridRep. and ours. Green lines indicate correct correspondences and red lines denote incorrect ones. We observe that our method tends to establish globally consistent correspondences based on relatively high fidelity regions, thereby achieving better registration results.

6.4. Ablation Study

Analysis of Global Representation. The global representation of panorama $\mathbf{I}_{\mathrm{p}}^{(N)}$ provides sufficient information of a scene. To verify its effectiveness, we conduct an experiment by removing the panorama inference sub-network h_ϕ. Instead, we use the extrapolated RGB-D scan of $\mathbf{I}_1$ to replace the panorama $\mathbf{I}_{\mathrm{p}}^{(N)}$, where 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 global representation. The average errors is reduced from $37.11°/0.60m$, $24.33°/0.65m$ to $34.23°/0.56m$, $20.67°/0.61m$ on the Matterport and ScanNet datasets, respectively.

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 poses 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 Scan-

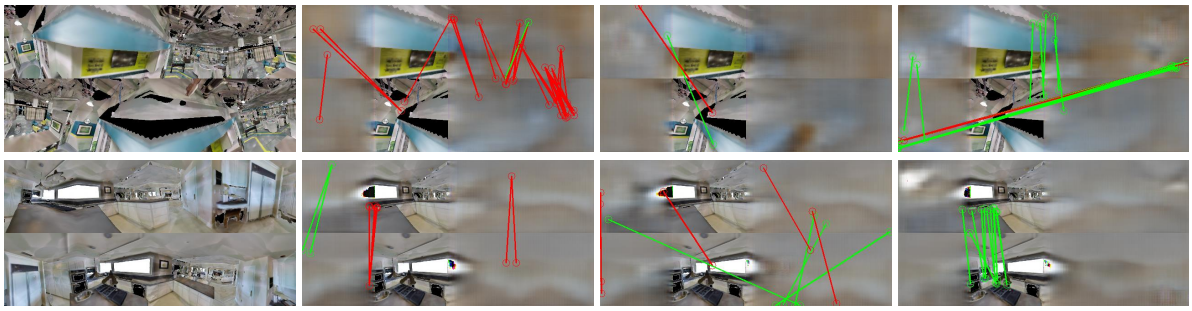

Figure 4. Visualizations of our method and baseline methods on the Matterport dataset. From left to right, we show the results of ground truth, ScanComp [32] ., HybridRep. [33] and ours. Green lines indicate correct correspondences and red lines denote incorrect ones.

	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	**34.23°**	**0.56**m	**20.67°**	**0.61**m

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

Step Length	Time	$\leq 10\%$		$\geq 10\%$	
		Rotation	Trans.	Rotation	Trans.
1	1.51 pps	41.74°	0.86m	18.48°	0.57m
2	1.47 pps	36.85°	0.78m	16.44°	0.55m
3	1.39 pps	36.85°	0.75m	15.68°	0.54m
4	1.32 pps	33.10°	0.77m	15.16°	0.54m
5	1.25 pps	32.74°	0.78m	15.00°	0.54m

Table 4. Relative errors and testing time with respect to step length on the ScanNet dataset. "pps" means pairs per second.

Net dataset without considering weights. 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. (8) and the reward loss function is removed, which forms a direct supervised regression method via deep networks. From Tab. 3, we observe 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.

Analysis of Reinforcement Learning. We use the reinforcement learning to iteratively align RGB-D scans for better results. During the training of the reinforcement learning, we observe that the two scans are usually aligned well within 5 steps, so we set the maximum action step N to 5. The relative errors and test speed are summarized with respect to steps in Tab. 4. When the step length is greater than 4, the performance improvement is not obvious, so we set the step to 4 during testing. It should be noted that when the step size is set to 1, the reinforcement learning degenerates into a simple regression process with an extra loss function of alignment errors. From Tab. 4, a simple regression network can obtain a small translation error, but still maintains a large rotation error. This verifies that the iterative reinforcement learning can get better refinements for boosting alignments. More details of reinforcement learning can be found in the *supplementary materials* (Supplementary Sec. S3).

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.

In future work, we will learn more reliable global information from multiple less-overlap scans for registration, further reducing the noise of global information and advancing our global-aware registration method.

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*, volume 1, 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*, volume 1611, pages 586–606, 1992. 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 Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2514–2523, 2020. 2

[6] Christopher Bongsoo Choy, Wei Dong, and Vladlen Koltun. Deep global registration. pages 2511–2520, 2020. 1

[7] 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

[8] 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

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

[10] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO: fast semi-direct monocular visual odometry. In *International Conference on Robotics and Automation*, pages 15–22, 2014. 2

[11] 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

[12] Lionel Heng and Benjamin Choi. Semi-direct visual odometry for a fisheye-stereo camera. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4077–4084. IEEE, 2016. 2

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

[14] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 4267–4276, 2021. 2

[15] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 11366–11374, 2020. 2

[16] Kenji Koide and Emanuele Menegatti. Non-overlapping rgb-d camera network calibration with monocular visual odometry. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9005–9011, 2020. 1

[17] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 13184–13193, 2021. 2

[18] Nicolas Mellado, Dror Aiger, and N. Mitra. Super 4pcs fast global pointcloud registration via smart indexing. *Computer Graphics Forum*, 33, 2014. 6, 7

[19] 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

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

[21] 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

[22] John Schulman, F. Wolski, Prafulla Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[23] 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/CVF International Conference on Computer Vision*, pages 5703–5711, 2021. 2

[24] 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

[25] 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

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

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

[28] Teng Wan, S. Du, Wenting Cui, Runzhao Yao, Yuyan Ge, Ce Li, Yue Gao, and Nanning Zheng. Rgb-d point cloud registration based on salient object detection. *IEEE transactions on neural networks and learning systems*, PP, 2021. 1

[29] 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

[30] 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

[31] 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

[32] 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, 6, 7, 8

[33] 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, 8

[34] Chaoqiang Zhao, Yang Tang, Qiyu Sun, and Athanasios V Vasilakos. Deep direct visual odometry. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 1, 2

[35] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *European conference on computer vision*, pages 766–782, 2016. 2

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

[37] 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