Diff-Reg: Diffusion Model in Doubly Stochastic Matrix Space for Registration Problem

Qianliang Wu[1], Haobo Jiang[5], Lei Luo[1], Jun Li[1], Yaqing Ding[4], Jin Xie[2,3]✉, and Jian Yang[1]✉

[1] PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[2] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[3] School of Intelligence Science and Technology, Nanjing University, Suzhou, China
[4] Visual Recognition Group, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
[5] National University of Singapore, Singapore
wuqianliang@njust.edu.cn

Abstract. Establishing reliable correspondences is essential for 3D and 2D-3D registration tasks. Existing methods commonly leverage geometric or semantic point features to generate potential correspondences. However, these features may face challenges such as large deformation, scale inconsistency, and ambiguous matching problems (e.g., symmetry). Additionally, many previous methods, which rely on single-pass prediction, may struggle with local minima in complex scenarios. To mitigate these challenges, we introduce a diffusion matching model for robust correspondence construction. Our approach treats correspondence estimation as a denoising diffusion process within the doubly stochastic matrix space, which gradually denoises (refines) a doubly stochastic matching matrix to the ground-truth one for high-quality correspondence estimation. It involves a forward diffusion process that gradually introduces Gaussian noise into the ground truth matching matrix and a reverse denoising process that iteratively refines the noisy one. In particular, we deploy a lightweight denoising strategy during the inference phase. Specifically, once points/image features are extracted and fixed, we utilize them to conduct multiple-pass denoising predictions in the reverse sampling process. Evaluation of our method on both 3D and 2D-3D registration tasks confirms its effectiveness. The code is available at https://github.com/wuqianliang/Diff-Reg.

1 Introduction

The 3D registration problem, encompassing point cloud registration and image-to-point cloud registration, is critical in various computer vision and com-

✉ Corresponding Authors

puter graphics applications, including 3D reconstructions, localization, AR, and robotics. These applications usually necessitate precise correspondences (matchings) between point cloud pairs or image-to-point cloud pairs for reliable rigid transformation or non-rigid deformation estimations.

The goal of achieving accurate matchings is to identify the most significant correspondences [4, 66, 70] with local or global semantic or geometric consistency [25, 40, 49, 53]. However, this objective would be challenging, especially in situations with globally ambiguous matching patches, large deformation, scale inconsistency, and low overlapping problems.

Recently, deep learning-based feature matching methods [19, 25, 28, 34, 40, 62, 63, 65, 66] have achieved significant progress in point cloud registration by employing UNet-like [17, 45] backbones to extract superpoints (subsampled patches) and their associated features. These methods typically compute an initial matching matrix between superpoints in the feature space. Additionally, outlier rejection techniques [3, 10, 20, 60, 69] propose specialized methodologies to identify improved inlier correspondences based on certain semantic or geometric priors [14, 39, 53, 56–59, 70]. However, these methods usually rely on a single-pass prediction of correspondences, which may not always yield optimal results.

In this paper, drawing inspiration from the diffusion model [2, 18, 43, 48], we introduce a diffusion matching model in the doubly stochastic matrix space [7]. By training a diffusion model with the doubly stochastic matrix space as a feasible solution domain, we effectively learn a generalized optimization algorithm specifically designed for the doubly stochastic matrix space, adapted to the characteristics of the dataset or scene. Our diffusion matching model consists of two main components: a forward diffusion process and a reverse denoising process, which operates within the matrix space. The forward diffusion process gradually introduces Gaussian noise into the ground truth matching matrix, while the reverse denoising process iteratively refines the noisy matrix to the optimal one. For efficiency, we propose a novel and generalized lightweight denoising module that can be adapted to 2D-3D and 3D registration tasks. Finally, we establish a specific variational lower bound associated with our diffusion matching model in the doubly stochastic matrix space and a simplified version of the objective function to train our framework effectively.

Why diffused in the doubly stochastic matrix space? Tasks like image or point cloud registration face challenges like scale inconsistency, large deformation, ambiguous matching, and low overlapping. Several state-of-the-art studies [3, 33, 40, 53, 54, 62, 69] have attempted to encode high-order combinational geometric consistency. However, these manually crafted designs may not encompass all potential effective strategies for various challenging scenarios (e.g., large deformation). The diffusion process in the matrix space is a practical data augmentation technique that can generate additional training samples incorporating any-order combinational geometric consistency, offering a promising approach to address these challenges. A doubly stochastic matching matrix is, in fact, a dual-directional mapping. It offers a one-to-one mapping relation constraint for any kind of two-view (and any two-modality) matching/registration problem. The

diffusion process within the matching matrix space naturally provides a broader range of training samples. Furthermore, the many-to-one relationship between the matching matrix space and the warping operation space can facilitate our method in learning a more optimal sampling path used in the reverse sampling process, in contrast to methods that conduct diffusion in SE(3) space.

Our approach presents several advantages compared to previous registration methods. The forward diffusion process generates diverse training samples, acting as data augmentation for the feature backbone and single-pass prediction head. The drawback of single-pass prediction methods is that if correspondences are predicted in a local minimum, subsequent outlier rejection or post-processing steps may face significant challenges. In contrast, our reverse denoising process, guided by the posterior distribution, allows for escaping from local minima, enabling the process to initiate from either white noise or any initial solution. Another enhancement is eliminating the feature backbone during the reverse denoising process at inference time. This streamlined design enables the denoising sampling process to explore a broader solution space (e.g., the matrix space), increasing diversity and facilitating more iterative steps. The findings from our empirical experiments support these claims.

Our contributions are summarized as follows:

- To our knowledge, we are the first to deploy the diffusion model in the doubly stochastic matrix space for iteratively exploring the optimal matching matrix through the reverse denoising sampling process.
- The lightweight design of our reverse denoising module results in faster convergence in the reverse sampling process. Moreover, our framework can effectively utilize reverse denoising sampling in a noise-to-target fashion or start from a highly reliable initial solution.
- We conducted comprehensive experiments on the real-world 4DMatch [29], 3DMatch [68], and RGB-D Scenes V2 [25, 68] datasets to validate the effectiveness of our diffusion matching model on 3D registration and 2D-3D registration task.

2 Related work

2.1 3D and 2D-3D Registration

The registration problem estimates the transformation between the point cloud or image-to-point-cloud pair. Recently, there have been significant advancements in feature learning-based methods for point cloud registration. Many of these state-of-the-art approaches, such as [4, 19, 40, 64, 65, 67], leverage a backbone architecture similar to KPConv [45] to downsample points and generate features with larger receptive fields. To further enhance the performance of these methods, they integrate prior knowledge and incorporate learnable outlier rejection modules. For instance, GeoTR [40] introduces angle-wise and edge-wise embeddings into the transformer encoder, while RoITr [66] integrates local Point Pair Features (PPF) [13] to improve rotation invariance.

In addition to feature learning-based methods, another category of registration methods focuses on outlier rejection of candidate correspondences. For instance, PointDSC [3] utilizes a maximum clique algorithm in the local patch to cluster inlier correspondences. SC2-PCR [10] constructs a second-order consistency graph for candidate correspondences and theoretically demonstrates its robustness. Building on the second-order consistency graph proposed by SC2-PCR [10], MAC [69] introduces a variant of maximum clique algorithms to generate more reliable candidate inlier correspondences. Moreover, methods such as PEAL [67] and DiffusionPCR [9] employ an iterative refinement strategy to enhance the overlap prior information obtained from a pre-trained GeoTr [40].

Recently, significant advancements have been made in 2D-3D registration methods [24, 25, 50, 51]. These methods face similar challenges to 3D registration tasks, with the additional complexity of scale inconsistency caused by the perspective projection of images. To address the issue of scale inconsistency, we propose the incorporation of a pre-trained feature backbone, DINO v2 [36], which offers superior multiscale features. Additionally, implementing diffused data augmentation in our diffusion matching model can enhance the ability to identify prominent combinational and consistent correspondences.

2.2 Diffusion Models for 3D Registration

Recently, the diffusion model [18,43,44] has made great development in many fields, including human pose estimation [15,42], camera pose estimation [52], object detection [8], segmentation [5,16]. These developments have been achieved through a generative Markov Chain process based on the Langevin MCMC [37] or a reversed diffusion process [43]. Recognizing the power of the diffusion model to iteratively approximate target data distributions from white noise using hierarchical variational decoders, researchers have started applying it to point cloud registration and 6D pose estimation problems.

The pioneer work [46] that applied the diffusion model in the SE(3) space was accomplished by utilizing NCSN [44] to learn a denoising score matching function. This function was then used for reverse sampling with Langevin MCMC in SE(3) space to evaluate 6DoF grasp pose generation. Additionally, [21] implemented DDPM [18] in the SE(3) space for 6D pose estimation by employing a surrogate point cloud registration baseline model. Similarly, GeoTR [40] served as a denoising module in [9], gradually denoising the overlap prior given by the pre-trained model, following a similar approach to PEAL [67].

3 The proposed Approach

3.1 Problem Formulation

Given source point clouds $\mathbf{P} \in \mathbb{R}^{N\times 3}$ and target point clouds $\mathbf{Q} \in \mathbb{R}^{M\times 3}$, the 3D registration task is to find top-k correspondences $\mathcal{C}$ from matching matrix $\mathbf{E}$ and to conduct warping transformation ($\mathbf{\Gamma} \in$ SE(3) for rigid transformations,

and 3D flow fields for non-rigid transformations) to align the overlap region of $\mathbf{P}$ and $\mathbf{Q}$. In the context of 2D-3D registration, with a source image $\mathbf{X} \in \mathbb{R}^{H\times W\times 2}$ and target point cloud $\mathbf{Y} \in \mathbb{R}^{M\times 3}$, the standard pipeline involves determining the top-k correspondences $\mathcal{C} = \{(\mathbf{x}_i, \mathbf{y}_j)|\mathbf{x}_i \in \mathbb{R}^2, \mathbf{y}_j \in \mathbb{R}^3\}$, and then estimating the rigid transformation $\mathbf{\Gamma} \in \mathrm{SE}(3)$ by minimizing the 2D projection error:

$$\min_{\mathbf{\Gamma}\in SE(3)} \sum_{\mathbf{x}_i,\mathbf{y}_j\in\mathcal{C}} ||Proj(\mathbf{\Gamma}(\mathbf{y}_j), \mathbf{K}) - \mathbf{x}_i||_2$$

where $\mathbf{K}$ represents the camera intrinsic matrix, and $Proj(\cdot,\cdot)$ denotes the projection function from 3D space to the image plane.

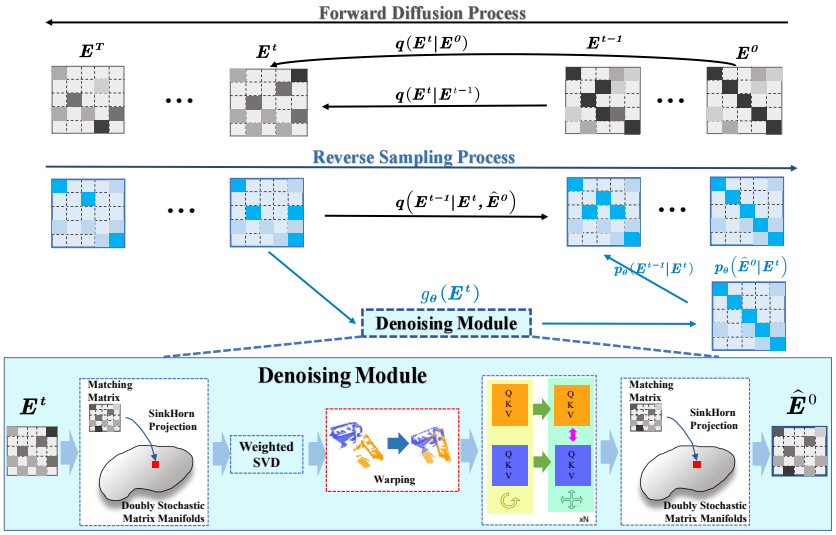


Fig. 1: Overview of our diffusion matching model. The forward diffusion process is driven by the Gaussian transition kernel $q(\mathbf{E}^t|\mathbf{E}^{t-1})$, which has a closed form $q(\mathbf{E}^t|\mathbf{E}^0)$. The denoising model $g_\theta(\mathbf{E}^t)$ learns a reverse denoising gradient that points to the target solution $\mathbf{E}^0$. During inference, in the reverse sampling process, we utilize the predicted $\hat{\mathbf{E}}_0$ and DDIM [43] to sampling $\mathbf{E}^{t-1}$.

3.2 Overview

Our framework comprises a feature backbone (e.g., KPConv [45]/ResNet [17]) and a diffusion matching model [18]. In the 3D registration task, the KPConv backbone takes source point clouds $\mathbf{P}$ and target $\mathbf{Q}$ as input and performs downsamplings to obtain the superpoints $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$, along with their associated features $\mathbf{F}^{\hat{\mathbf{P}}} \in \mathbb{R}^{N\times d}$ and $\mathbf{F}^{\hat{\mathbf{Q}}} \in \mathbb{R}^{M\times d}$. In the 2D-3D registration task, a ResNet [17] with FPN [32] downsamples the image $\mathbf{X} \in \mathbb{R}^{H\times W\times 2}$ (in image-point-coud pair

$(\mathbf{X}, \mathbf{Q})$) to the superpixels $\hat{\mathbf{X}} \in \mathbb{R}^{\hat{H} \times \hat{W} \times 2}$ and give the associated image feature $\mathbf{F}^{\hat{\mathbf{X}}} \in \mathbb{R}^{\hat{H} \times \hat{W} \times d}$, while the target point cloud $\mathbf{Q}$ is processed by a KPConv backbone, similar to the 3D registration task. Additionally, a depth map $\mathbf{D}^{\hat{\mathbf{X}}}$ and new superpixel feature $\mathbf{F}^{\hat{\mathbf{X}}}_{dino}$ for superpixels $\hat{\mathbf{X}}$ are given by the pre-trained depth estimation model [61] and visual feature backbone DINO v2 [36], respectively.

Our diffusion module g_θ (refer to Section 3.4) takes these superpoints (or superpixels) and associated features as inputs. We employ one loss for our denoising module and another for the single-pass prediction head during the training stage. During inference, we utilize g_θ in the reverse sampling process to predict the target matching matrix $\mathbf{E}^0$ from any noisy one and exploit DDIM [43] to sample a more accurate matching matrix. The denoising module g_θ primarily consists of four components: (1) Sinkhorn Projection [12], (2) Weighted SVD [6], (3) Warping Function (4) Denoising Transformer Network [47] and (5) Matching function. More details of our framework can be found in section 3.4 and the appendix.

3.3 Diffusion Model in Doubly Stochastic Matrix Space

In this section, we introduce the construction of our diffusion matching model for generating the matching matrix between two scans. We denote the matching matrix as $\mathbf{E} \in \{0, 1\}^{N \times M}$, and we assume $\mathbf{E}$ is defined in a nonsquare "doubly stochastic" matrix space $\mathcal{M}$ (refer to Appendix).

Forward Diffusion Process. As mentioned in DDPM [18], the forward diffusion process is fixed to a Markovian chain, denoted as $q(\mathbf{E}^{1:T}|\mathbf{E}^0)$, which generates a sequence of latent variables $\mathbf{E}^t$ by gradually injecting Gaussian noise into the ground truth matching matrix $\mathbf{E}^0$. The diffused matching matrix $\mathbf{E}^t$ at arbitrary timestamp t has a closed form:

$$\mathbf{E}^t \sim q(\mathbf{E}^t|\mathbf{E}^0) = \mathcal{N}(\mathbf{E}^t; \sqrt{\bar{\alpha}}\mathbf{E}^0, (1 - \bar{\alpha})\mathbf{I})). \tag{1}$$

where the added noise over each element of the matrix is sampled independently and identically distributed (i.i.d.). However, this diffused $\mathbf{E}^t \sim q(\mathbf{E}^t|\mathbf{E}^0)$ is a continuous matrix in $\mathbb{R}^{N \times M}$, which is outside the feasible solution space of matching matrices (i.e., doubly stochastic matrix manifolds). To address this issue, we apply the following projection to confine the matrix $\mathbf{E}^t$ to the feasible solution space $\mathcal{M}$:

$$\begin{aligned} \text{(Rigid)} \quad & \mathbf{E}^t = \sqrt{\bar{\alpha}_t}\mathbf{E}^0 + \sqrt{1 - \bar{\alpha}_t} f_\epsilon(\epsilon_0), \quad \tilde{\mathbf{E}}^t = \mathbf{E}^t - \mathrm{Min}(\mathbf{E}^{\mathrm{t}}), \\ \text{(Deformable)} \quad & \mathbf{E}^t = \sqrt{\bar{\alpha}_t}\mathbf{E}^0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_0, \quad \tilde{\mathbf{E}}^t = \mathrm{Sigmoid}(\mathbf{E}^{\mathrm{t}}), \\ & \tilde{\mathbf{E}}^t = \mathbf{f}_{\mathrm{sinkhorn}}(\tilde{\mathbf{E}}^t) \end{aligned} \tag{2}$$

where the $\mathbf{f}_{\mathrm{sinkhorn}}$ operation is from the Sinkhorn algorithm [12] and $f_\epsilon = (\epsilon \% 1)(abs(\epsilon)/\epsilon)\eta$. We empirically set $\eta = 1.5$ and $\epsilon_0 \sim \mathcal{N}(\epsilon; 0, \mathbf{I})$.

Reverse Denoising Sampling Process. Given a diffusion Markovian chain $\mathbf{E}^0 \rightarrow \mathbf{E}^1 \rightarrow ... \rightarrow \mathbf{E}^T$, we need to learn a reverse transition kernel with the posterior distribution $q(\mathbf{E}^{t-1}|\mathbf{E}^t, \mathbf{E}^0)$ to sample the reverse Markovian chain

$\mathbf{E}^T \rightarrow \mathbf{E}^{T-1} \rightarrow ... \rightarrow \mathbf{E}^0$ from a white noise $\mathbf{E}^T$ to achieve the target matching matrix. The posterior distribution $q(\mathbf{E}^{t-1}|\mathbf{E}^t, \mathbf{E}^0)$ conditioned on $\mathbf{E}^0$ and $\mathbf{E}^t$ is defined as:

$$\begin{aligned} q(\mathbf{E}^{t-1}|\mathbf{E}^t, \mathbf{E}^0) &= \frac{q(\mathbf{E}^t|\mathbf{E}^{t-1}, \mathbf{E}^0) q(\mathbf{E}^{t-1}|\mathbf{E}^0)}{q(\mathbf{E}^t|\mathbf{E}^0)} \\ &\propto \mathcal{N}(\mathbf{E}^{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\mathbf{E}^t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\mathbf{E}^0}{1-\bar{\alpha}_t}}_{\mu_q(\mathbf{E}^t, \mathbf{E}^0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\Sigma_q(t)}). \end{aligned} \tag{3}$$

To effectively train our denoising network, we derive the variational lower bound of the log-likelihood of the training samples $\mathbf{E}^0$:

$$log p(\mathbf{E}^0) \geq \mathbb{E}_{\mathbf{E}^{1:T} \sim q(\mathbf{E}^{1:T}|\mathbf{E}^0)} \left[log \left(\frac{p_\theta(\mathbf{E}^{0:T})}{q(\mathbf{E}^{1:T}|\mathbf{E}^0)} \right) \right] \propto \mathbb{E}_q \left[\sum_{t=2}^{T} log(p_\theta(\mathbf{E}^0|\mathbf{E}^t)) \right] \tag{4}$$

Based on the derivation of Eqn. (4), we can further simplify the variational lower bound above to train $p_\theta(\mathbf{E}^0|\mathbf{E}^t)$:

$$L_{simple} = -\mathbb{E}_{q(\mathbf{E}^0)} \left[\sum_{t-1}^{T} \mathbb{E}_{q(\mathbf{E}^t|\mathbf{E}^0)} log p_\theta(\mathbf{E}^0|\mathbf{E}^t) \right]. \tag{5}$$

3.4 The Lightweight Denoising Module g_θ

This section outlines the architecture of the lightweight denoising module g_θ. In 3D registration tasks, g_θ take the superpoints $\hat{\mathbf{P}}$, $\hat{\mathbf{Q}}$ with associated points features $\mathbf{F}^{\hat{\mathbf{P}}}$, $\mathbf{F}^{\hat{\mathbf{Q}}}$ as the inputs in the reverse sampling process. Similarly, in 2D-3D registration tasks, the inputs of g_θ are superpixels $\hat{\mathbf{X}}$ and superpoints $\hat{\mathbf{Q}}$ with associated features $\{\mathbf{F}^{\hat{\mathbf{X}}}, \mathbf{D}^{\hat{\mathbf{X}}}, \mathbf{F}^{\hat{\mathbf{X}}}_{dino}, \mathbf{F}^{\hat{\mathbf{Q}}}\}$. At inference time, g_θ inputs a noised matching matrix $\mathbf{E}^t$ and outputs a predicted target matching matrix $\hat{\mathbf{E}}_0$.

We define the denoising module g_θ by sequentially stacking five components as a differentiable layer:

Sinkhorn Projection: $\mathbf{f}_{\text{sinkhorn}}(\cdot)$. To constrain the matching matrix $\mathbf{E}^t$ within the doubly stochastic matrices manifolds, we utilize the SinkHorn [12] iterations to project $\mathbf{E}^t$. We treat this operation as a key role in our framework rather than a post-processing in other methods.

Weighted SVD: **soft_procrustes**$(\cdot, \cdot, \cdot)$. Given top-k confident correspondences κ, we utilize the weighted SVD algorithm [1] (differentiable) to compute the transformation $\mathbf{R}, \mathbf{t}$ in a closed form:

$$\begin{aligned} \mathbf{H} &= \sum_{(i,j) \in C} \tilde{\mathbf{E}}(i,j) \hat{\mathbf{p}}_i \hat{\mathbf{q}}_j^\top, \ \mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top, \\ \mathbf{R} &= \mathbf{U}\text{diag}(1, 1, \det(\mathbf{U}\mathbf{V}^\top))\mathbf{V}, \\ \mathbf{t} &= \frac{1}{|\kappa|} \left(\sum_{(i,\cdot) \in \kappa} \hat{\mathbf{p}}_i - \mathbf{R} \sum_{(\cdot,j) \in \kappa} \hat{\mathbf{q}}_j \right) \end{aligned} \tag{6}$$

where $(\hat{\mathbf{p}}_i, \hat{\mathbf{q}}_j)$ is a superpoint correspondences. In the 2D-3D registration task, we first project a superpixel depth map $\mathbf{D}^{\hat{\mathbf{X}}}$ to a point cloud $\mathbf{P}_D^{\hat{\mathbf{X}}}$ using camera intrinsic. Then we utilize this SVD decomposition to compute $\mathbf{R}, \mathbf{t}$ with $\mathbf{P}_D^{\hat{\mathbf{X}}}$ and $\hat{\mathbf{Q}}$ as inputs.

The Weighted SVD operation can be viewed as a differentiable projection that arises from the matching matrix $\tilde{\mathbf{E}}$ obtained from the first Sinkhorn Projection. Alternatively, in the 2D-3D registration task, this Weighted SVD operation can be substituted with a differentiable PnP (Perspective-n-Point) [26].

Warping Function: **warping**$(\cdot, \cdot, \cdot)$. After obtaining transformation $\mathbf{R}, \mathbf{t}$, the rigid warping of source point clouds is computed by $\mathbf{W}(\hat{\mathbf{p}}_i) = \mathbf{R}\hat{\mathbf{p}}_i + \mathbf{t}$. In this paper, we use rigid warping for rigid and deformable registration cases to demonstrate our design. As for deformable registration, with the denoised correspondences, we can compute the flow fields for all points in $\hat{\mathbf{P}}$ by performing nearest neighbor interpolation with the predicted inlier correspondences as anchors.

Denoising Transformer: $f_\theta(\cdot, \cdot, \cdot, \cdot, \cdot, \cdot)$. We observed empirically that a simple noise model does not hurt performance. Thus, we exploit a lightweight Transformer [47] as our denoising network. Specifically, we utilize a 6-layer inter-leaved attention layers transformer f_θ for denoising feature embedding. Worth noting that, in each denoising step, only coarse level source point cloud $\hat{\mathbf{P}}_t$ and its position encoding $\Theta(\hat{\mathbf{P}}_t)$ (or target point cloud $\hat{\mathbf{Q}}$ in 2D-3D registration task) have their values changed according to the warping operation, while other input parameters remain fixed. This is the key to our fast sampling speed.

Attention Layer in f_θ: In the 3D registration task, following [28], the vectors $\mathbf{q}, \mathbf{k}, \mathbf{v}$ in the self-attention, are computed as:

$$\begin{aligned} &\mathbf{q}_i = \Theta(\mathbf{p}_i)\mathbf{W}_q f^{\hat{\mathbf{P}}_i},\ \mathbf{k}_j = \Theta(\mathbf{p}_j)\mathbf{W}_k f^{\hat{\mathbf{P}}_j},\ \mathbf{v}_j = \mathbf{W}_v f^{\hat{\mathbf{P}}_j}, \\ &f^{\hat{\mathbf{P}}_i} = f^{\hat{\mathbf{P}}_i} + \mathrm{MLP}(\mathrm{cat}[\mathbf{q}_\mathrm{i}, \Sigma_\mathrm{j}\alpha_{\mathrm{ij}}\mathbf{v}_\mathrm{j}]), \end{aligned} \tag{7}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d\times d}$ are the attention weights, $\alpha_{ij} = softmax(\mathbf{q}_i\mathbf{k}_j^\top/\sqrt{d})$, and $\Theta(\cdot)$ is the relative rotationary position encoding [28]. MLP$(\cdot)$ is a 3-layer fully connected network, and $cat[\cdot, \cdot]$ is the concatenating operator. The cross attention layer is the standard form that $\mathbf{q}$ and $\mathbf{k}, \mathbf{v}$ are computed by source and target point clouds, respectively.

In the 2D-3D registration task, we take image inputs $\{\hat{\mathbf{X}}, \mathbf{F}^{\hat{\mathbf{X}}}, \mathbf{F}_{dino}^{\hat{\mathbf{X}}}\}$ and point cloud input $\{\hat{\mathbf{Q}}, \mathbf{F}^{\hat{\mathbf{Q}}}\}$ to compute $\mathbf{q}, \mathbf{k}, \mathbf{v}$ by utilizing standard attention layers [47]. We also take Fourier embedding function [35] to embed superpixels $\hat{\mathbf{X}}$ and superpoints $\hat{\mathbf{Q}}$ for positional encoding.

Matching Function: **matching_logits**$(\cdot, \cdot, \cdot, \cdot)$. We compute matching "logits" between $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$ by features $\mathbf{F}^{\hat{\mathbf{P}}}$ (or $\mathbf{F}^{\hat{\mathbf{X}}}$) and $\mathbf{F}^{\hat{\mathbf{Q}}}$: $\tilde{\mathbf{E}}(i,j) = \frac{1}{\sqrt{d}}\left\langle f^{\hat{\mathbf{P}}_i}, f^{\hat{\mathbf{q}}_j}\right\rangle$.

For the sake of clarity, we provide pseudo-code in Algorithm.1 to describe the logic of our entire denoising module g_θ for the 3D Registration task. A similar definition of g_θ for the 2D-3D registration task can be found in the appendix.

Algorithm 1 Denoising Module g_θ for 3D Registration Task.

Require: Sampled matching matrix $\mathbf{E}^t \in \mathbb{R}^{N\times M}$; Point clouds $\hat{\mathbf{P}}, \hat{\mathbf{Q}} \in \mathbb{R}^3$ and associated point features $\mathbf{F}^{\hat{\mathbf{P}}}, \mathbf{F}^{\hat{\mathbf{Q}}}$.
Ensure: Target matching matrix $\hat{\mathbf{E}}_0$.
function $g_\theta(\mathbf{E}^t, \hat{\mathbf{P}}, \hat{\mathbf{Q}}, \mathbf{F}^{\hat{\mathbf{P}}}, \mathbf{F}^{\hat{\mathbf{Q}}})$
  $\tilde{\mathbf{E}}_t \leftarrow \mathbf{f_{sinkhorn}}(\mathbf{E}^t)$
  $\hat{\mathbf{R}}_t, \hat{\mathbf{t}}_t \leftarrow \mathbf{soft\_procrustes}(\tilde{\mathbf{E}}_t, \hat{\mathbf{P}}, \hat{\mathbf{Q}}),\ \hat{\mathbf{P}}_t \leftarrow \mathbf{warping}(\hat{\mathbf{P}}, \hat{\mathbf{R}}_t, \hat{\mathbf{t}}_t)$
  $\tilde{\mathbf{F}}^{\hat{\mathbf{P}}_t}, \tilde{\mathbf{F}}^{\hat{\mathbf{Q}}_t} \leftarrow f_\theta(\hat{\mathbf{P}}_t, \hat{\mathbf{Q}}, \mathbf{F}^{\hat{\mathbf{P}}}, \mathbf{F}^{\hat{\mathbf{Q}}}, \Theta(\hat{\mathbf{P}}_t), \Theta(\hat{\mathbf{Q}}))$
  $\tilde{\mathbf{E}}_0 \leftarrow \mathbf{matching\_logits}(\tilde{\mathbf{F}}^{\hat{\mathbf{P}}_t}, \tilde{\mathbf{F}}^{\hat{\mathbf{Q}}_t}, \Theta(\hat{\mathbf{P}}_t), \Theta(\hat{\mathbf{Q}}))$
  $\hat{\mathbf{E}}_0 \leftarrow \mathbf{f_{sinkhorn}}(\tilde{\mathbf{E}}_0)$
  return $\hat{\mathbf{E}}_0$
end function

4 Experiments

4.1 3D Non-Rigid Registration Task

Datasets. 4DMatch/4DLoMatch [28] is an 3D non-rigid benchmark generated by the animation sequences from DeformingThings4D [30]. We follow the dataset split provided in [28], which has a wide range of overlap ratio, that 45%-92% in 4DMatch and 15%-45% in 4DLoMatch.

Implementation Details. Our framework utilize a KPConv [45] backbone to produce the superpoints $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$ and the asccociated features $\mathbf{F}^{\hat{\mathbf{P}}}$ and $\mathbf{F}^{\hat{\mathbf{Q}}}$. The dimension d of superpoint features $\mathbf{F}^{\hat{\mathbf{P}}}$ and $\mathbf{F}^{\hat{\mathbf{Q}}}$ is set as $d = 432$. Subsequently, we employ a repositioning transformer [28] to provide a single-pass prediction of the matching matrix and the resulting transformation $[\mathbf{R}, \mathbf{t}]$, both of which are supervised by the matching loss L_M and warping loss L_W introduced in [28]. We utilize a focal loss L_{simple} (modified from Eqn.5) to guide the training of the denoising module g_θ. The total loss function is defined as $L = L_M + L_W + L_{simple}$.

We train the model for 30 epochs on the 4DMatch dataset with a batch size of 2. We adopt the training/validation/test split strategy from Predator [19] and Lepard [28]. At inference time, we conduct 20 iterations in the reverse sampling process while the total diffusion steps during training are set to 1000.

Metrics. Following Lepard [28], we utilize two evaluation metrics to assess the quality of predicted matches. (1) Inlier Ratio (IR): The correct fraction in the correspondences prediction $\mathcal{K}_{pred}$. (2) Non-rigid Feature Matching Recall (NFMR): The fraction of ground truth correspondences $(u, v) \in \mathcal{K}_{gt}$ that can be successfully recovered by using the predicted correspondences $\mathcal{K}_{pred}$ as anchors. The NFMR metric provides a better characterization of the global rationality of overall body deformation, directly indicating whether the anchor $\mathcal{K}_{pred}$ effectively captures the body movements.

Quantitative Results. We compare our method with two categories of state-of-the-art methods. The first category includes Scene Flow Methods such as

Table 1: Quantitative results on the 4DMatch and 4DLoMatch benchmarks. The best results are highlighted in bold, and the second-best results are underlined.

Category	Method	4DMatch		4DLoMatch	
		NFMR(%)	IR(%)	NFMR(%)	IR (%)
Scene Flow	PointPWC [55]	21.60	20.0	10.0	7.20
	FLOT [38]	27.10	24.90	15.20	10.70
Feature Matching	D3Feat [4]	55.50	54.70	27.40	21.50
	Predator [19]	56.40	60.40	32.10	27.50
	Lepard [28]	83.60	82.64	66.63	55.55
	GeoTR [40]	83.20	82.20	65.40	63.60
	RoITr [66]	83.00	<u>84.40</u>	69.40	<u>67.60</u>
	Diff-Reg(Backbone)	<u>85.47</u>	81.15	72.37	59.50
	Diff-Reg(steps=1)	85.23	83.85	<u>73.19</u>	65.26
	Diff-Reg(steps=20)	**88.40**	**86.41**	**76.23**	**67.80**

PWC [55], FLOT [38], and NSFP [27]. The second category encapsulates Feature Matching-Based Methods, namely D3Feat [4], Predator [19], Lepard [28], GeoTR [40], and RoITr [66].

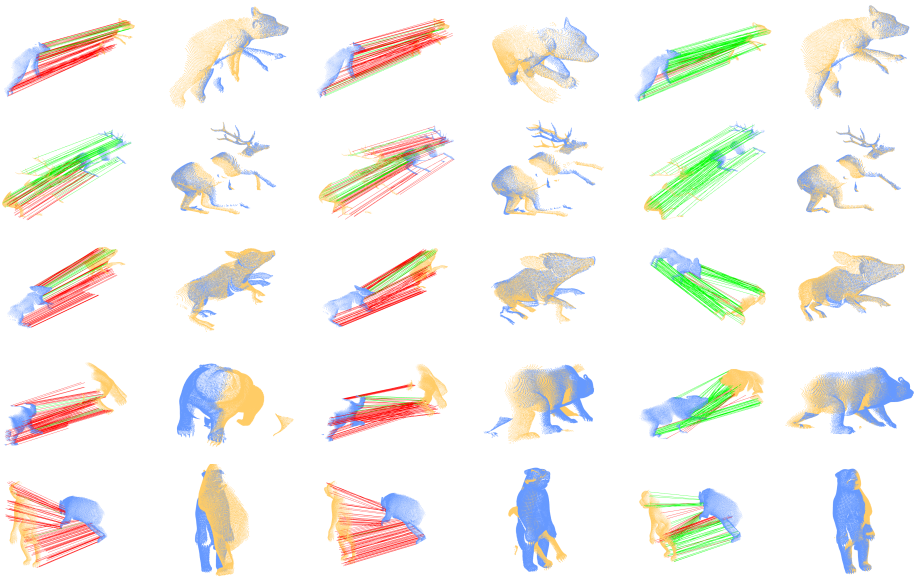
As illustrated in Table 1, our method demonstrates significant improvements compared to the single-pass baselines. "Diff-Reg(Backbone)" refers to the single-pass prediction head (i.e., reposition transformer in Lepard [28]), while "Diff-Reg(steps=1)" and "Diff-Reg(steps=20)" denotes our denoising module g_θ with one single step and 20 steps of reverse sampling. For both NFMR and IR metrics, "Diff-Reg(steps=20)" achieves the best performance. The improvement in NFMR of "Diff-Reg(Backbone)" compared to the baselines indicates that our diffused training samples in the matching matrix space enhance the feature backbone's representation, enabling the capture of crucial salient correspondences that are helpful for consistent global deformation. The significant enhancement of "Diff-Reg(steps=20)" over "Diff-Reg(steps=1)" demonstrates that the reverse denoising sampling process indeed searches for a better solution guided by the learned posterior distribution.

To validate that the predicted correspondences indeed improve deformable registration, we conducted experiments using the state-of-the-art registration method GraphSCNet [41]. As indicated in Table 2, our predicted correspondences are beneficial for deformable registration, particularly in the more challenging 4DLoMatch benchmark.

Qualitative Results. We provide a visualization to demonstrate our method's effectiveness in Fig.2. For a fair comparison, we exploit the source point cloud's "metric index" (i.e., the test point set in the 4DMatch/4DLoMatch dataset) for all methods. Taking the predicted correspondences from RoITr [66], GeoTr [66], and "Our (steps=20)" as anchor correspondences, we calculate the deformation flow for the source test points by applying neighborhood k-nearest neighbors (KNN) interpolation based on the anchors. The deformable registration of the bear's two front paws in the first row and second/fourth column reveals that dealing with ambiguous matching patches of asymmetric objects can be highly challenging. However, our denoising process can handle this scenario perfectly.

Table 2: Non-rigid registration results of 4DMatch/4DLoMatch. Given predicted correspondences, we utilize the non-rigid registration method GraphSCNet [41] to conduct the deformable registration. We retrain RoITr* using the authors' code. The modified 4DMatch-F and 4DLoMatch-F datasets [29] exclude data involving near-rigid movements. The metrics (following [29, 41]) are 3D End Point Error (EPE), 3D Accuracy Strict (AccS) (<2.5cm or 5%), 3D Accuracy Relaxed (AccR)(<5cm or 5%), and Outlier Ratio (OR) (>30%).

Method	4DMatch-F				4DLoMatch-F			
	EPE↓	AccS↑	AccR↑	OR↓	EPE↓	AccS↑	AccR↑	OR ↓
PointPWC [55]	0.182	6.25	21.49	52.07	0.279	1.69	8.15	55.70
FLOT [38]	0.133	7.66	27.15	40.49	0.210	2.73	13.08	42.51
GeomFmaps [9]	0.152	12.34	32.56	37.90	0.148	1.85	6.51	64.63
Synorim-pw [19]	0.099	22.91	49.86	26.01	0.170	10.55	30.17	31.12
Lepard [28]+GraphSCNet [41]	0.042	70.10	83.80	9.20	0.102	40.00	59.10	17.50
GeoTR [40]+GraphSCNet [41]	0.043	72.10	84.30	9.50	0.119	41.00	58.40	20.60
RoITr* [66] +GraphSCNet [41]	0.056	59.60	80.50	12.50	0.118	32.30	56.70	20.50
Diff-Reg+GraphSCNet [41]	**0.041**	**73.20**	**85.80**	**8.30**	**0.095**	**43.80**	**62.90**	**15.50**



(a) GeoTR [40] **(b)** RoITr [66] **(c)** Diff-Reg(steps=20)

Fig. 2: The qualitative results of non-rigid registration in the 4DMatch/4DLoMatch benchmark. The top two lines are from 4DMatch, while the bottom three are from 4DLoMatch. The blue and yellow colors denote the source and target point cloud, respectively. The green and red lines indicate whether the predicted deformable flow from the source points is accepted by the threshold. The deformable registration is built by GraphSCNet [25]. Zoom in for details.

The deformable registration results from the top three rows (from the 4DMatch benchmark) indicate that the baseline methods struggle to compute reliable and consistent correspondences between scans with large deformations. The bottom two rows (from the 4DLoMatch benchmark) also demonstrate that low overlapping combined with deformation results in a disaster. These visualizations demonstrate that our denoising module in the matching matrix space provides a more effective approach for tackling deformable registration tasks.

4.2 2D-3D Registration Task: RGB-D SCENES V2

The 2D-3D registration task is non-trivial because the 2D image data is a perspective projection of the 3D scene, which creates the scale ambiguity problem [25]. Since there is no good robust differentiable PnP [23] solver that can be integrated into our denoising module, we deploy the SOTA depth estimation model DepthAnything [61] to generate "affine-invariant depth" for the image. Subsequently, we utilize the camera's intrinsic matrix to project the estimated depth map onto a new non-metric point cloud paired with the corresponding real point cloud in the dataset, creating a challenge scale ambiguous registration problem. To alleviate the image data's scale ambiguity, we use the SOTA self-supposed pre-trained visual feature backbone DINOv2 [36] to enhance the image features.

Datasets. RGB-D Scenes V2 [22] are generated from 14 indoor scenes comprising 11,427 RGB-D frames. Following [25], we split the 14 sequences into image-to-point-cloud pairs data, where scenes 0-8/11-14/9-10 are used for training/validation/testing. The resulting dataset contains 1,748 training pairs, 236 validation pairs, and 497 testing pairs of image-to-point-clouds.

Implementation Details. For the single-pass backbone design, we follow 2D3D-MATR [25]. Specifically, for images in data pair, we utilize a ResNet [17] with FPN [31] to generate down-sampled superpixel and associated features. For the real point cloud in the data pair, we exploit KPConv [45] to extract the down-sampled superpoints with associated features. A transformer [47] is deployed with inputs of superpixel and associated image features (including ResNet features and DINO features) from the image and superpoint and associated features from real point cloud to predict the cross-modality features. The denoising transformer in g_θ design has a similar definition. The new non-metric point cloud generated from the image depth map output by the DepthAnything model [61] is used as inputs of the weighted SVD function in g_θ to compute $\mathbf{R}, \mathbf{t}$.

We utilize the coarse level circle loss [25] and fine level matching loss [25] for the singe-pass backbone [25], while a focal loss for our denoising module g_θ. More details about network design are in the appendix. We train our model about 30 epochs with batch size 1.

Metrics. We evaluate our method using the Registration Recall (RR) metric: the ratio of image-to-point-cloud pairs' RMSE is under 10cm.

Table 3: Evaluation results on RGB-D Scenes V2 [25]. The best results are highlighted in bold, and the second-best results are underlined.

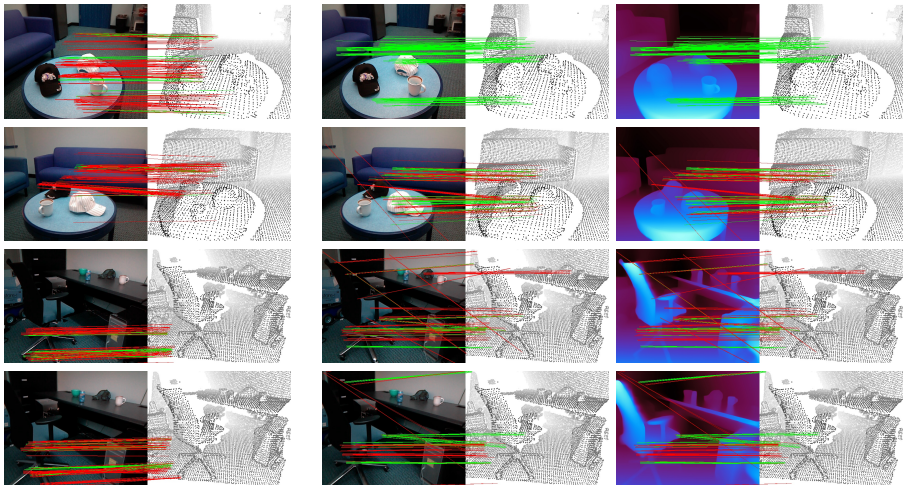
Method	Scene-11	Scene-12	Scene-13	Scene-14	Mean
Mean depth (m)	1.74	1.66	1.18	1.39	1.49
Registration Recall(%)↑					
FCGF-2D3D [11]	26.4	41.2	37.1	16.8	30.4
P2-Net [50]	40.3	40.2	41.2	31.9	38.4
Predator-2D3D [19]	44.4	41.2	21.6	13.7	30.2
2D3D-MATR [25]	63.9	53.9	58.8	49.1	56.4
FreeReg+Kabsch [51]	38.7	51.6	30.7	15.5	34.1
FreeReg+PnP [51]	74.2	72.5	54.5	27.9	57.3
Diff-Reg(dino)	87.5	86.3	63.9	60.6	74.6
Diff-Reg(dino/backbone)	79.2	86.3	75.3	**71.2**	78.0
Diff-Reg(dino/steps=1)	<u>94.4</u>	<u>98.0</u>	<u>85.6</u>	<u>63.7</u>	<u>85.4</u>
Diff-Reg(dino/steps=10)	**98.6**	**99.0**	**86.6**	<u>63.7</u>	**87.0**
Diff-Reg(dino/backbone$_{epnp}$)	95.8	96.1	88.7	69.0	87.4

Quantitative Results. We introduce a single-pass baseline "Diff-Reg(dino)", in which we integrate the visual foundation model DINOv2 [36] into the single-pass model 2D3D-MATR's [25] ResNet. "Diff-Reg(dino/backbone)" represents the single-pass prediction head derived from "Diff-Reg(dino)" after joint training with our denoising module g_θ. "Diff-Reg(steps=1)" and "Diff-Reg(steps=10)" refer to our diffusion matching model with one step and ten steps of reverse denoising sampling.

As demonstrated in Table 3, the results for "Diff-Reg(dino/backbone)" indicate that the diffused training samples in the matrix space serve as data augmentation to enhance the representation of the ResNet feature backbone and single-pass prediction head. Additionally, the outcome for "Diff-Reg(dino/steps=10)" reveals that our denoising module g_θ effectively tackles the scale ambiguous issue in the 2D-3D registration. Furthermore, the result for "Diff-Reg(dino/steps=1)" reveals that the diffused training samples within the matrix space indeed enhance the single-pass prediction head in "Diff-Reg(dino)."

We also carried out an additional experiment to show that the performance improvement of "Diff-Reg(dino/backbone)" compared to "Diff-Reg(dino)" is attributed to our diffusion matching model. In this experiment, we did not use the depth map (from DepthAnything model [61]), opting instead to employ a differentiable weighted EPnP [23] solver in the denoising module g_θ to replace the weighted SVD layer. We take only superpixels and superpoints as inputs of the EPnP solver, with correspondence weights computed by associated superpixel and superpoint features. This setting denoted by "Diff-Reg(dino/backbone$_{epnp}$)" (in Table 3) achieved a high recall rate of 87.4%, indicating that our diffused samples in matching matrix space indeed enhance the feature backbone.

Qualitative Results. The prediction examples of "Diff-Reg(dino/steps=10)" in Fig. 3 reveal that our diffusion matching model excels at capturing salient correspondences crucial for combinatorial consistency, regardless of their distance



(a) Diff-Reg(dino) **(b)** Diff-Reg(dino/steps=10)

Fig. 3: The qualitative results of top-200 predicted correspondences on the RGB-D Scenes V2 benchmark [22]. The green/red color indicates whether the matching score is accepted based on a threshold value. Zoom in for details.

from the camera. On the other hand, the matches generated by "Diff-Reg(dino)" tend to be more focused at specific distances. For instance, in the third row, the correspondences produced by "Diff-Reg(dino)" are located very close to the camera, while the correspondences from "Diff-Reg(dino/steps=10)" encompass objects such as hats on the black table that are situated at a greater distance. In the first row, "Diff-Reg(dino)" fails to capture the correspondences on the sofa, and in the second row, the correspondences of the white hat on the table are lost. An extreme case in the fourth row demonstrates that "Diff-Reg(dino)" misses the farthest correspondence on the column bookshelf or wall.

5 Conclusion

This paper presents a novel diffusion module that leverages a diffusion matching model in the doubly stochastic matrix space to learn a posterior distribution for guiding the reverse denoising sampling process within the matrix space. Moreover, we have integrated a lightweight design into the denoising module to decrease the time cost associated with iterative reverse sampling. Experimental results on both 3D registration and 2D-3D registration tasks confirm the effectiveness and efficiency of our proposed denoising module.

Acknowledgments. This work was partially supported by the National Science Fund of China (Grant Nos. 62361166670, 62276144, 62072242, 62276135) and the Czech Science Foundation (GACR) JUNIOR STAR Grant No. 22-23183M.

References

1. Arun, K.S., Huang, T.S., Blostein, S.D.: Least-squares fitting of two 3-d point sets. IEEE Transactions on pattern analysis and machine intelligence (1987)
2. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. Advances in Neural Information Processing Systems **34**, 17981–17993 (2021)
3. Bai, X., Luo, Z., Zhou, L., Chen, H., Li, L., Hu, Z., Fu, H., Tai, C.L.: Pointdsc: Robust point cloud registration using deep spatial consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
4. Bai, X., Luo, Z., Zhou, L., Fu, H., Quan, L., Tai, C.L.: D3feat: Joint learning of dense detection and description of 3d local features. In: CVPR (2020)
5. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. ArXiv (2021)
6. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Sensor fusion IV: control paradigms and data structures (1992)
7. Caron, R.M., Li, X., Mikusiński, P., Sherwood, H., Taylor, M.D.: Nonsquare "doubly stochastic" matrices. Lecture Notes-Monograph Series **28**, 65–75 (1996), http://www.jstor.org/stable/4355884
8. Chen, S., Sun, P., Song, Y., Luo, P.: Diffusiondet: Diffusion model for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
9. Chen, Z., Ren, Y., Zhang, T., Dang, Z., Tao, W., Süsstrunk, S., Salzmann, M.: Diffusionpcr: Diffusion models for robust multi-step point cloud registration. arXiv preprint arXiv:2312.03053 (2023)
10. Chen, Z., Sun, K., Yang, F., Tao, W.: Sc2-pcr: A second order spatial compatibility for efficient and robust point cloud registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
11. Choy, C., Park, J., Koltun, V.: Fully convolutional geometric features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
12. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems (2013)
13. Deng, H., Birdal, T., Ilic, S.: Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
14. Fu, K., Liu, S., Luo, X., Wang, M.: Robust point cloud registration framework based on deep graph matching. In: CVPR (2021)
15. Gong, J., Foo, L.G., Fan, Z., Ke, Q., Rahmani, H., Liu, J.: Diffpose: Toward more reliable 3d pose estimation. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
16. Gu, Z., Chen, H., Xu, Z., Lan, J., Meng, C., Wang, W.: Diffusioninst: Diffusion model for instance segmentation. ArXiv (2022)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems (2020)
19. Huang, S., Gojcic, Z., Usvyatsov, M., Wieser, A., Schindler, K.: Predator: Registration of 3d point clouds with low overlap. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition (2021)

20. Jiang, H., Dang, Z., Wei, Z., Xie, J., Yang, J., Salzmann, M.: Robust outlier rejection for 3d registration with variational bayes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
21. Jiang, H., Salzmann, M., Dang, Z., Xie, J., Yang, J.: Se (3) diffusion model-based point cloud registration for robust 6d object pose estimation. arXiv preprint arXiv:2310.17359 (2023)
22. Lai, K., Bo, L., Fox, D.: Unsupervised feature learning for 3d scene labeling. In: IEEE International Conference on Robotics and Automation (ICRA) (2014)
23. Lepetit, V., Moreno-Noguer, F., Fua, P.: Ep n p: An accurate o (n) solution to the p n p problem. International journal of computer vision (2009)
24. Li, J., Lee, G.H.: Deepi2p: Image-to-point cloud registration via deep classification. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
25. Li, M., Qin, Z., Gao, Z., Yi, R., Zhu, C., Guo, Y., Xu, K.: 2d3d-matr: 2d-3d matching transformer for detection-free registration between images and point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
26. Li, S., Xu, C., Xie, M.: A robust o (n) solution to the perspective-n-point problem. IEEE transactions on pattern analysis and machine intelligence (2012)
27. Li, X., Kaesemodel Pontes, J., Lucey, S.: Neural scene flow prior. Advances in Neural Information Processing Systems (2021)
28. Li, Y., Harada, T.: Lepard: Learning partial point cloud matching in rigid and deformable scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
29. Li, Y., Harada, T.: Non-rigid point cloud registration with neural deformation pyramid. Advances in Neural Information Processing Systems (2022)
30. Li, Y., Takehara, H., Taketomi, T., Zheng, B., Nießner, M.: 4dcomplete: Non-rigid motion estimation beyond the observable surface. In: ICCV (2021)
31. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017)
32. Lin, T.Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
33. Mei, G., Huang, X., Yu, L., Zhang, J., Bennamoun, M.: Cotreg: Coupled optimal transport based point cloud registration. arXiv preprint arXiv:2112.14381 (2021)
34. Mei, G., Tang, H., Huang, X., Wang, W., Liu, J., Zhang, J., Van Gool, L., Wu, Q.: Unsupervised deep probabilistic approach for partial point cloud registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
35. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM (2021)
36. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
37. Parisi, G.: Correlation functions and computer simulations. Nuclear Physics B (1981)
38. Puy, G., Boulch, A., Marlet, R.: Flot: Scene flow on point clouds guided by optimal transport. In: European conference on computer vision (2020)

39. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
40. Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Xu, K.: Geometric transformer for fast and robust point cloud registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
41. Qin, Z., Yu, H., Wang, C., Peng, Y., Xu, K.: Deep graph-based spatial consistency for robust non-rigid point cloud registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
42. Shan, W., Liu, Z., Zhang, X., Wang, Z., Han, K., Wang, S., Ma, S., Gao, W.: Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
43. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
44. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems (2019)
45. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF international conference on computer vision (2019)
46. Urain, J., Funk, N., Peters, J., Chalvatzaki, G.: Se (3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion. In: 2023 IEEE International Conference on Robotics and Automation (ICRA) (2023)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems (2017)
48. Vignac, C., Krawczuk, I., Siraudin, A., Wang, B., Cevher, V., Frossard, P.: Digress: Discrete denoising diffusion for graph generation. arXiv preprint arXiv:2209.14734 (2022)
49. Deng, H., Birdal, T., Ilic, S.: Ppfnet: Global context aware local features for robust 3d point matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)
50. Wang, B., Chen, C., Cui, Z., Qin, J., Lu, C.X., Yu, Z., Zhao, P., Dong, Z., Zhu, F., Trigoni, N., Markham, A.: P2-net: Joint description and detection of local features for pixel and point matching. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
51. Wang, H., Liu, Y., Wang, B., Sun, Y., Dong, Z., Wang, W., Yang, B.: Freereg: Image-to-point cloud registration leveraging pretrained diffusion models and monocular depth estimators. ArXiv (2023)
52. Wang, J., Rupprecht, C., Novotný, D.: Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
53. Wu, Q., Ding, Y., Luo, L., Zhou, C., Xie, J., Yang, J.: Sgfeat: Salient geometric feature for point cloud registration. arXiv preprint arXiv:2309.06207 (2023)
54. Wu, Q., Shen, Y., Jiang, H., Mei, G., Ding, Y., Luo, L., Xie, J., Yang, J.: Graph matching optimization network for point cloud registration. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2023)
55. Wu, W., Wang, Z., Li, Z., Liu, W., Fuxin, L.: Pointpwc-net: A coarse-to-fine network for supervised and self-supervised scene flow estimation on 3d point clouds. arXiv preprint arXiv:1911.12408 (2019)

56. Yan, Z., Lin, Y., Wang, K., Zheng, Y., Wang, Y., Zhang, Z., Li, J., Yang, J.: Tri-perspective view decomposition for geometry-aware depth completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4874–4884 (2024)
57. Yan, Z., Wang, K., Li, X., Zhang, Z., Li, G., Li, J., Yang, J.: Learning complementary correlations for depth super-resolution with incomplete data in real world. IEEE transactions on neural networks and learning systems (2022)
58. Yan, Z., Wang, K., Li, X., Zhang, Z., Li, J., Yang, J.: Rignet: Repetitive image guided network for depth completion. In: European Conference on Computer Vision. pp. 214–230. Springer (2022)
59. Yan, Z., Wang, K., Li, X., Zhang, Z., Li, J., Yang, J.: Desnet: Decomposed scale-consistent network for unsupervised depth completion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3109–3117 (2023)
60. Yang, H., Shi, J., Carlone, L.: Teaser: Fast and certifiable point cloud registration. IEEE Transactions on Robotics **37**(2), 314–333 (2020)
61. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. arXiv preprint arXiv:2401.10891 (2024)
62. Yao, R., Du, S., Cui, W., Ye, A., Wen, F., Zhang, H., Tian, Z., Gao, Y.: Hunter: Exploring high-order consistency for point cloud registration with severe outliers. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
63. Yew, Z.J., Lee, G.H.: Regtr: End-to-end point cloud correspondences with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6677–6686 (2022)
64. Yu, H., Hou, J., Qin, Z., Saleh, M., Shugurov, I., Wang, K., Busam, B., Ilic, S.: Riga: Rotation-invariant and globally-aware descriptors for point cloud registration. arXiv preprint arXiv:2209.13252 (2022)
65. Yu, H., Li, F., Saleh, M., Busam, B., Ilic, S.: Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. Advances in Neural Information Processing Systems (2021)
66. Yu, H., Qin, Z., Hou, J., Saleh, M., Li, D., Busam, B., Ilic, S.: Rotation-invariant transformer for point cloud matching. In: CVPR (2023)
67. Yu, J., Ren, L., Zhang, Y., Zhou, W., Lin, L., Dai, G.: Peal: Prior-embedded explicit attention learning for low-overlap point cloud registration. In: CVPR (2023)
68. Zeng, A., Song, S., Nießner, M., Fisher, M., Xiao, J., Funkhouser, T.: 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1802–1811 (2017)
69. Zhang, X., Yang, J., Zhang, S., Zhang, Y.: 3d registration with maximal cliques. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
70. Zhong, Y.: Intrinsic shape signatures: A shape descriptor for 3d object recognition. In: 2009 IEEE 12th international conference on computer vision workshops, ICCV workshops (2009)