# Multiway Point Cloud Mosaicking with Diffusion and Global Optimization

Shengze Jin[1] Iro Armeni[2] Marc Pollefeys[1,3] Dániel Baráth[1]
[1] Department of Computer Science, ETH Zurich, Switzerland
[2] Department of Civil and Environmental Engineering, Stanford University
[3] Microsoft Mixed Reality & AI Lab, Zurich, Switzerland

## Abstract

*We introduce a novel framework for multiway point cloud mosaicking (named Wednesday), designed to co-align sets of partially overlapping point clouds – typically obtained from 3D scanners or moving RGB-D cameras – into a unified coordinate system. At the core of our approach is ODIN, a learned pairwise registration algorithm that iteratively identifies overlaps and refines attention scores, employing a diffusion-based process for denoising pairwise correlation matrices to enhance matching accuracy. Further steps include constructing a pose graph from all point clouds, performing rotation averaging, a novel robust algorithm for re-estimating translations optimally in terms of consensus maximization and translation optimization. Finally, the point cloud rotations and positions are optimized jointly by a diffusion-based approach. Tested on four diverse, large-scale datasets, our method achieves state-of-the-art pairwise and multiway registration results by a large margin on all benchmarks. Our code and models are available at https://github.com/jinsz/Multiway-Point-Cloud-Mosaicking-with-Diffusion-and-Global-Optimization.*

## 1. Introduction

Registering multiple partially overlapping 3D point cloud fragments into a unified coordinate system is crucial to comprehensively representing an environment. This procedure has a wide range of applications in computer vision and robotics, such as in 3D scene understanding [41, 63], augmented reality [59, 62], and autonomous driving [48, 52, 74]. In particular, LiDAR or RGB-D-based mapping is often employed to build large-scale maps in self-driving and mobile robotics due to their direct and accurate 3D point cloud sensing capability. There are typically two steps in building such maps: pairwise and multiway registration.

The pairwise registration of partially overlapping point clouds is a thoroughly investigated problem, with several methods proposed over time. Conventional approaches to the pairwise problem are based on imposing geometric constraints [56, 66, 83] on hand-engineered feature descriptors [30, 43, 58, 68] employing robust estimators [7, 16, 29]. In recent years, research on local descriptors for pairwise registration of 3D point clouds has shifted towards deep learning methodologies [23, 24, 32, 37, 44, 47, 55, 78, 81, 82, 84]. Such approaches have demonstrated impressive results by implicitly learning local and global scene characteristics, which are then distilled into highly distinctive local descriptors. Although these methods have proven to be effective, directly applying them to the multiway problem has conceptual drawbacks: (i) low overlap between adja-

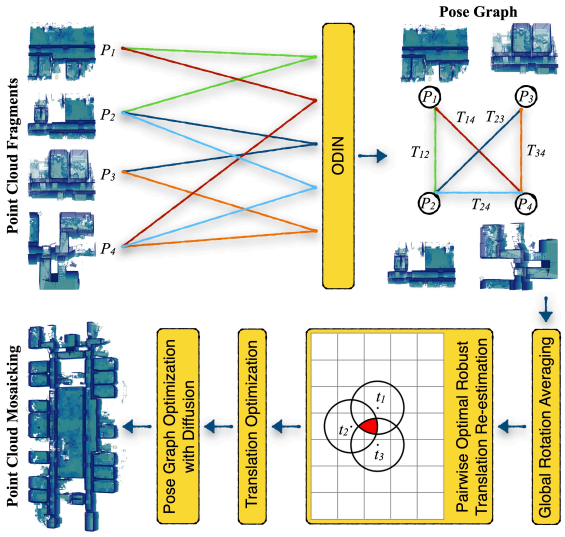


Figure 1. The proposed multiway registration method, Wednesday, starts with pairwise registration of an unordered set of partially overlapping point clouds using the proposed matcher (ODIN). The process then optimizes the constructed pose graph, which includes global point cloud poses (vertices) and relative transforms (edges), through a sequence of steps: (a) global rotation averaging, (b) a novel optimal robust translation re-estimation method conceptualized as finding maximal sphere overlaps, (c) position averaging, and (d) diffusion-based pose graph optimization. The output is the point clouds in a unified coordinate system.

cent point clouds may result in incorrect matches, and (ii) the reliance only on local evidence, which can be problematic in scenes with scarce or repetitive 3D structures.

In contrast to pairwise registration, point cloud mosaicking (*i.e.*, the globally consistent multiway alignment of unorganized point clouds) has received much less attention. Traditional approaches [18, 20, 27, 60, 73, 75, 85] primarily tackle this problem from the robotics perspective in scenarios where the differences between adjacent point clouds are minimal (*e.g.*, they come from subsequent RGB-D frames), and the recorded data distinctly reveals the trajectory of the robot as it captures the scene. In such cases, pairwise registration provides accurate initialization for the multiway problem, with the trajectory providing additional constraints [86] that can be incorporated into the optimization process. Other methods [2–4, 9, 15, 34, 49, 69], focusing on having unordered point cloud sets, frame the challenge as an optimization procedure. However, in practice, it is still sensitive to the failures in the pairwise registration.

Recent progress in the field has seen the introduction of end-to-end pipelines [18, 33, 70, 79] that aim at learning specific local and global characteristics of the scene. However, such methods prioritize ease of training over efficacy during inference. To facilitate differentiability, the representations (*e.g.*, rotation manifold) and algorithms (*e.g.*, iteratively reweighted least-squares) utilized are selected for their compatibility with the end-to-end pipeline rather than for their potential to yield optimal performance at inference.

This paper focuses on designing a pipeline for accurate point cloud registration even in challenging environments with spatial and temporal changes. We enhance state-of-the-art pairwise registration algorithms based on two observations: (i) the predicted matching matrix often contains noise, and its denoising leads to improved 3D-3D matches; (ii) while finding individual 3D point matches is key to estimating the rigid transformation, the underlying objective is to find the best point cloud overlap. This can be directly measured and integrated into the matching process. To achieve accurate multiway registration, we rely on classical geometric optimization-based approaches known for their accuracy and generalizability. The proposed pipeline is the result of carefully selected and *new* optimization techniques for the best accuracy, combining learning-based and classical algorithms to benefit from data-driven approaches while maintaining the efficiency and applicability of geometric methods. The contributions are as follows:

- A novel pipeline (see Fig. 1) for multiway point cloud registration consisting of modules for pairwise estimation, global rotation averaging, translation re-estimation and averaging, and diffusion-based final optimization.
- A novel pairwise point cloud registration method, ODIN, incorporating point cloud overlap scores into attention learning and diffusion-based correlation matrix denoising for highly accurate pairwise matching.
- An efficient and globally optimal robust consensus (*i.e.*, inlier number) maximization approach for re-estimating relative translations given known global orientations.
- As a technical contribution, we adapt a recent diffusion-based pose graph optimization [71] to point clouds.

The proposed advancements and other methods fused into a single pipeline achieve state-of-the-art accuracy by a *great* margin. It achieves 82% rotation error reduction on the most challenging dataset [61]. It also reduces the average position error by 27% across the tested datasets.

## 2. Related Work

**Pairwise registration** is traditionally a two-step process. The first step is the coarse alignment stage, where initial estimates of the relative transformations are obtained. The second step is the refinement stage, where the global poses are iteratively refined to minimize the 3D registration error, assuming a rigid transformation. Coarse alignment often uses handcrafted [30, 43, 58, 68] or learned [37, 55, 81, 82] 3D local feature descriptors to establish tentative pointwise correspondences. They are used with a RANSAC-like robust estimator [29] or geometric hashing [12, 26, 36] to find the pose parameters and the consistent matches. Another approach uses 4-point congruent sets to establish correspondences [1, 50, 65]. In the refinement stage, coarse transformation parameters are fine-tuned using a variant of the iterative closest point (ICP) algorithm [10]. However, ICP-like algorithms [19, 45, 76] are not robust against outliers. ICP algorithms can be extended to use additional radiometric, temporal, or odometry constraints [86].

Recent work [37, 42, 47, 72, 82] either directly regress transformations or refine correspondences by considering the information from both point clouds, *e.g.*, with attention layers. Our proposed pairwise method falls into this category, building upon transformer-based alternatives by incorporating predicted point cloud overlap scores and diffusion-based denoising of the estimated correlation matrices.

**Multiview registration** methods [2–4, 9, 15, 33, 34, 49, 69, 70, 79] reconstruct a complete scene from a collection of partially overlapping point clouds. The first family of methods employs a multiview ICP to optimize camera poses and 3D correspondences [13, 28, 40, 51]. However, these methods often struggle with the increased correspondence estimation complexity. To address this, some approaches focus solely on optimizing motion, using the point clouds to evaluate errors [11, 66, 86]. However, such ICP-based methods are prone to inaccuracies in the pairwise poses that provide the starting point for the multi-view procedure.

Other modern methods take a different approach, using global cycle consistency to optimize poses starting from an initial set of pairwise maps. This so-called synchronization method is known for its efficiency [2, 3, 9, 11, 14, 39, 49, 64,

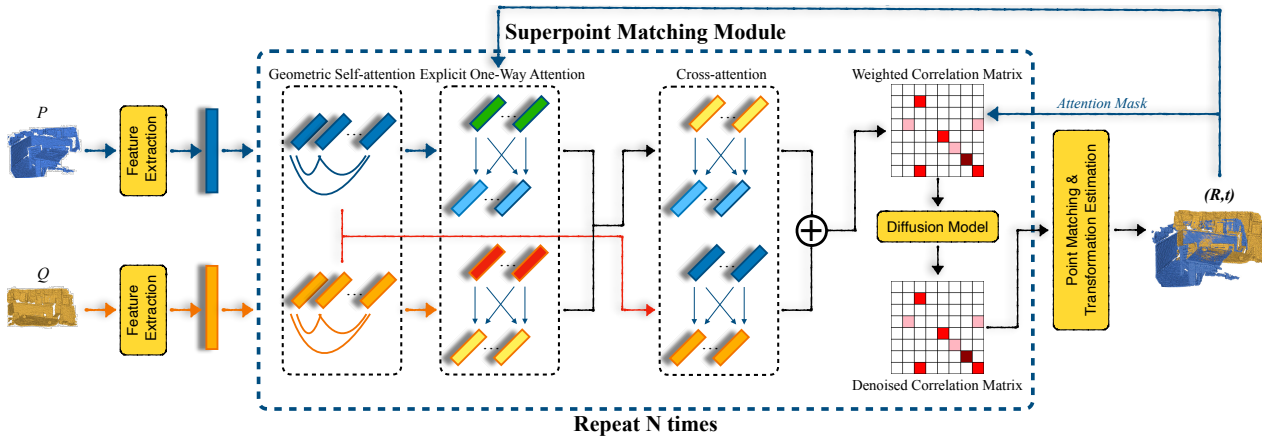


Figure 2. **Two-view registration**. Given two points clouds as input, ODIN (Section 3.1) first extracts features that are then processed by geometric self-attention to learn point-specific attention features. Next, the process is separated into two parallel streams: In (a), the features are processed by explicit one-way self and cross-attentions. This process incorporates overlap scores determined in the final stage. In (b), the features directly go through cross-attention. The determined correlation matrix is the weighted average of the correlations from the two streams. A diffusion-based denoising cleans the correlations. Finally, point matching and transformation estimation are performed. The overlap scores implied by the estimated transform are sent back to the attention learning module as a mask and the process starts over.

66, 69, 86]. Global Structure-from-Motion [22] synchronizes observed relative motions and decomposes them into rotation, translation, and scale components. Our pipeline follows a similar global approach, with geometric optimization at its core, combined with recent advancements in deep learning to leverage the best of both worlds.

Recent methods [4, 20, 38] employ an iteratively reweighted least-squares (IRLS) scheme to adaptively downweight noisy pairwise estimates. However, the iterative refinement of IRLS can become trapped in local minima and may fail to remove outlier edges. To tackle this challenge, recent learning-based advances [33, 39, 70, 79] adopt a data-driven strategy to learn robust reweighting functions. While these approaches allow for end-to-end training, they often make design choices prioritizing ease of training over performance during inference. In contrast, our paper takes a different approach. Instead of prioritizing end-to-end trainability, which may not be directly relevant in real-world scenarios, we aim to design a framework estimating highly accurate multiway registration from a set of partially overlapping point clouds. Our focus is on precision and reliability, guiding our design choices throughout the development of the framework.

## 3. Pairwise and Multiway Registration

**Problem Definition.** Point cloud mosaicking from pairwise registrations can be formalized as a rigid transform averaging, recovering 3D orientations $\boldsymbol{R}_i \in \mathrm{SO}(3)$ and positions $\boldsymbol{t}_i \in \mathbb{R}^3$ from a set of estimated relative pairwise motions $(\boldsymbol{R}_{ij}, \boldsymbol{t}_{ij})$, where $i, j \in [1, n]$ and $n \in \mathbb{N}$ is the number of partially overlapping point clouds ($i \neq j$). We can express the information as pose graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each vertex $v \in \mathcal{V}$ represents a global pose, and each edge $(i, j) \in \mathcal{E}$ is the relative motion of point clouds $\mathcal{P}_i$ and $\mathcal{P}_j$. The relative and global transforms are related by constraints:

$$\boldsymbol{R}_{ij} = \boldsymbol{R}_j \boldsymbol{R}_i^{\mathrm{T}}, \quad \boldsymbol{t}_{ij} = \boldsymbol{R}_i(\boldsymbol{t}_j - \boldsymbol{t}_i), \quad \forall (i, j) \in \mathcal{E}. \quad (1)$$

Relative transforms are obtained from pairwise registration methods and are corrupted by noise and outliers. Thus, a solution that satisfies all constraints in Eq. 1 cannot be found. To circumvent this, transformation averaging seeks to recover global transforms with minimum consistency error.

**Pipeline Summary.** The proposed pipeline (called Wednesday) performing pairwise and then multiway point cloud registration is depicted in Fig. 1. It begins by iterating through pairs of point clouds, where tentative 3D point correspondences are established and relative poses are estimated, as proposed in Section 3.1. These pairs are utilized to construct a pose graph. To further refine the estimated poses, the pipeline adopts a decoupled approach. This involves first optimizing the global orientations as described in Section 3.2, then re-estimating the relative translations based on the global orientations (Section 3.3), and finally, optimizing the global positions as outlined in Section 3.4. At last, diffusion-based optimization, detailed in Section 3.5, is applied to further optimize the pose graph.

### 3.1. Overlap and Diffusion-based Registration

This section introduces the Overlap-aware, Diffusion-aided paIrwise registratioN (ODIN) methoD, enhancing SOTA frameworks like GeoTransformer [55] and PEAL [82] by incorporating diffusion-based denoising and iteratively optimizing point cloud overlap within the attention learning.

**Feature extraction.** ODIN begins by extracting features from individual points and superpoints (clusters of points),

as in GeoTransformer [55]. Utilizing the KPConv-FPN backbone [67], we downsample input point clouds to multi-level features $\mathbf{F}^{\hat{\mathcal{P}}} \in \mathbb{R}^{|\hat{\mathcal{P}}| \times d}$ and $\mathbf{F}^{\hat{\mathcal{Q}}} \in \mathbb{R}^{|\hat{\mathcal{Q}}| \times d}$. The sets of coarsest resolution points, treated as superpoints, are denoted as $\hat{\mathcal{P}}$ and $\hat{\mathcal{Q}}$, along with their associated features. Additionally, dense sets of correspondences, $\tilde{\mathcal{P}}$ and $\tilde{\mathcal{Q}}$, and their features $\mathbf{F}^{\tilde{\mathcal{P}}} \in \mathbb{R}^{|\tilde{\mathcal{P}}| \times d}$ and $\mathbf{F}^{\tilde{\mathcal{Q}}} \in \mathbb{R}^{|\tilde{\mathcal{Q}}| \times d}$ are computed at half the original resolution. Each point in $\tilde{\mathcal{P}}$ is assigned to its nearest superpoint. The feature matrix associated with the points in $\mathcal{G}_i^{\tilde{\mathcal{P}}}$ is denoted as $\mathbf{F}_i^{\tilde{\mathcal{P}}} \subseteq \mathbf{F}^{\tilde{\mathcal{P}}}$. Superpoints without assignments are removed. Patches $\{\mathcal{G}_i^{\hat{\mathcal{Q}}}\}$ and features $\{\mathbf{F}_i^{\hat{\mathcal{Q}}}\}$ for point cloud $\hat{\mathcal{Q}}$ are computed similarly.

**Geometric Self-Attention.** Following [55], we employ self-attention mechanisms within the superpoints $\hat{\mathcal{P}}$ and $\hat{\mathcal{Q}}$ to learn point-specific attention features. The self-attention is formalized as follows: Given input feature matrix $\boldsymbol{X} \in \mathbb{R}^{|\hat{\mathcal{P}}| \times d_t}$, we compute output feature matrix $\boldsymbol{Z} \in \mathbb{R}^{|\hat{\mathcal{P}}| \times d_t}$. Each element in $\boldsymbol{Z}$ is a cumulative sum of the weighted, projected input features as: $\boldsymbol{Z}_i = \sum_{j=1}^{|\hat{\mathcal{P}}|} a_{i,j}(x_j \boldsymbol{W}^V)$, where $a_{i,j}$ is the weight coefficient for the $i$th and $j$th superpoint, obtained via a row-wise softmax applied to the attention scores $e_{i,j}$. These scores are calculated as $e_{i,j} = ((\boldsymbol{x_i} \boldsymbol{W}^Q)(\boldsymbol{x_j} \boldsymbol{W}^K + \boldsymbol{r_{i,j}} \boldsymbol{W}^R)^T)/\sqrt{d_t}$, where $\boldsymbol{r_{i,j}} \in \mathbb{R}^{d_t}$ represents a vector embedding geometric structural information, capturing pairwise distances and angular relationships among points. Projection matrices $\boldsymbol{W}^Q$, $\boldsymbol{W}^K$, $\boldsymbol{W}^V$, and $\boldsymbol{W}^R \in \mathbb{R}^{d_t \times d_t}$ correspond to queries, keys, values, and geometric structure embeddings, respectively. The outcome of this process is matrices $\boldsymbol{X}^{\hat{\mathcal{P}}}$ and $\boldsymbol{X}^{\hat{\mathcal{Q}}}$, representing the learned attention features for superpoints in $\hat{\mathcal{P}}$ and $\hat{\mathcal{Q}}$.

We distinguish anchor and non-anchor superpoints, determined by an attention mask based on the overlap scores predicted later. As the overlap is unknown in the first iteration, we use an identity mask (updated later), making all superpoints anchors. Matrices $\boldsymbol{X}^{\hat{\mathcal{P}}_A}$, $\boldsymbol{X}^{\hat{\mathcal{Q}}_A}$ are features for anchors, and $\boldsymbol{X}^{\hat{\mathcal{P}}_N}$, $\boldsymbol{X}^{\hat{\mathcal{Q}}_N}$ are for non-anchors. At this point, the attention learning splits into two separate streams. The steps discussed next are contributions of this paper.

**A) Explicit One-Way and Cross-Attention.** This stream is responsible for incorporating overlap information into attention learning. It starts with an explicit one-way attention module [82] to learn the intra-frame correlations with anchor superpoints, which is critical to encode inter-frame geometric consistency. The module begins by differentiating anchor and non-anchor superpoints, working with their respective feature matrices $\boldsymbol{X}^{\hat{\mathcal{P}}_A}$ and $\boldsymbol{X}^{\hat{\mathcal{P}}_N}$. The attention features for non-anchor superpoints, denoted as $\boldsymbol{Z}^{\hat{\mathcal{P}}_N}$ are computed by leveraging the attention features of the anchor superpoints as $\boldsymbol{Z}_m^{\hat{\mathcal{P}}_N} = \sum_{n=1}^{|\hat{\mathcal{P}}_A|} \alpha_{m,n} \left(\boldsymbol{X}_n^{\hat{\mathcal{P}}_A} \boldsymbol{W}^V\right)$. Here, $\alpha_{m,n}$ indicates the attention score, obtained through a row-wise softmax function, representing the feature correlation between non-anchor $\boldsymbol{X}^{\hat{\mathcal{P}}_N}$ and anchor $\boldsymbol{X}^{\hat{\mathcal{P}}_A}$ superpoints. The specific attention score $e_{m,n}$ is given by:

$$e_{m,n} = \frac{\left(\boldsymbol{X}_m^{\hat{\mathcal{P}}_N} \boldsymbol{W}_A^P\right)\left(\boldsymbol{X}_n^{\hat{\mathcal{P}}_A} \boldsymbol{W}^K\right)^T}{\sqrt{d_t}}. \tag{2}$$

This is similarly applied to update attention features for $\boldsymbol{X}^{\hat{\mathcal{Q}}_N}$, while features $\boldsymbol{X}^{\hat{\mathcal{Q}}_A}$ and $\boldsymbol{X}^{\hat{\mathcal{P}}_A}$ remain unchanged.

**B) Cross-Attention Stream.** Complementing the previous calculations, the framework employs another stream directly utilizing the self-attention embeddings, determined in feature extraction, to facilitate cross-attention learning [55].

**Correlation Maps.** Each stream outputs a correlation map where the value in the $i$th row and $j$th column signifies the correlation of the $i$th superpoint in the first point cloud and the $j$th in the second. The final correlation map is obtained as a weighted sum of correlations from both streams. It is designed to evolve by being updated at the end of the pipeline, capturing overlap information by leveraging the predicted transformation from prior iterations. The initial weight of one-way attention is set to zero. The upper stream will gradually gain more attention during training. We also apply an attention mask to the weighted correlation matrix, in which high-confidence matches gain more attention. This significantly accelerates the training process. Different from [82], our method can be trained from scratch and does not rely on the initial overlapping prediction.

**Correlation Denoising by Diffusion.** This step focuses on improving the predicted correlations by reducing the effect of noise. We employ diffusion models, a type of probabilistic generative model that learns to transform a noisy sample $h_K \sim \mathcal{N}(0, \mathbf{I})$ into a clean one $h_0$. Each noisy $h_k$ is expressed as a linear mix of the source sample $h_0$ and the noise variable $\epsilon$ as $h_k := \sqrt{\alpha_k} h_0 + \sqrt{1-\alpha_k}\epsilon, \quad \epsilon \sim \mathcal{N}(0,1)$. Using sample $h_0$ and the forward diffusion-generated noisy samples $\{h_k\}_{k=1}^K$, diffusion model $g$ is optimized to approximate the reverse process. Finally, the reverse step is recurrently performed to generate a high-quality sample $h_0$ from the noisy one $h_K$ using the trained model $g$.

We assume that the correlation matrix generated previously is a noisy observation of the actual, unknown correlation. Thus, we use a noise reduction diffusion process, enhancing the subsequent matching procedure. Drawing inspiration from [5] and [25], we adopt a UNet architecture and the overlap-aware circle loss [55]. The loss on superpoint patch $G_i^P$ (*i.e.*, a continuous representation of the underlying local surface) is defined as follows:

$$L_{oc}^P = \frac{1}{|A|} \sum_{G_i^P \in A}$$

$$\log\left(1 + \sum_{G_j^Q \in \epsilon_p} e^{\beta_{i,j}^p \lambda_i^j \left(d_i^j - \Delta_p\right)} - \sum_{G_k^Q \in \epsilon_n} e^{\beta_{i,k}^n \left(\Delta_n - d_i^k\right)}\right),$$

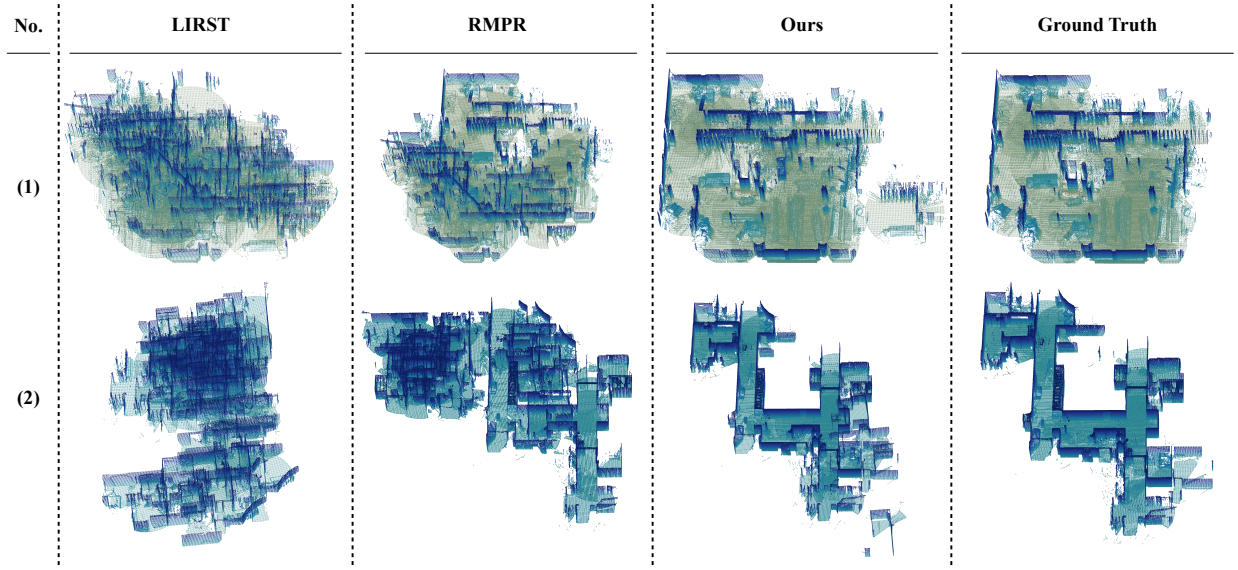


Figure 3. **Multiway point cloud registration** results on two scenes from the challenging NSS dataset [61] with the recent LIRST [79] and RMPR [70] and the proposed methods (ceilings are not shown). Also, we visualize the provided ground truth. We show results for LIRST and RMPR as they are the best-performing alternatives in Tab. 2. Such results are common output of these methods on this dataset.

where $d_i^j = \|\hat{h}_i^P - \hat{h}_j^Q\|_2$ is the distance in the feature space, $\lambda_i^j = (o_i^j)^{1/2}$ and $o_i^j$ represents the overlap ratio between $G_i^P$ and $G_j^Q$. The positive and negative weights are computed for each sample individually as $\beta_p^{i,j} = \gamma(d_i^j - \Delta_p)$ and $\beta_n^{i,k} = \gamma(\Delta_n - d_i^k)$. The loss $L_{oc}^Q$ on $Q$ is calculated similarly. The overall loss is $L_{oc} = (L_{oc}^P + L_{oc}^Q)/2$. The operational model is conditioned on the features of the superpoints. These features undergo iterative updates during each step of the denoising process. The overlap-aware circle loss is employed as a strategic optimization mechanism to refine and optimize these features. The foundational ground truth correlation matrix is the normalized matrix of overlapping ratios, where each constituent element represents the overlapping proportion of respective superpoint patches.

**Point Matching Module.** With the correlation matrix denoised, the next step establishes superpoint correspondences and extracts 3D point correspondences through the Point Matching Module. This step is similar to what is done in [55].

For each identified superpoint match $\hat{C}_i = (\hat{\mathcal{P}}_{x_i}, \hat{\mathcal{Q}}_{y_i})$, we employ an optimal transport layer to ascertain dense point correspondences between $G_{x_i}^P$ and $G_{y_i}^Q$. The process begins by constructing a cost matrix $\boldsymbol{C}_i$ as follows: $\boldsymbol{C}_i = \boldsymbol{F}_{x_i}^P (\boldsymbol{F}_{y_i}^Q)^T / \sqrt{d}$, where $n_i = |G_{x_i}^P|$ and $m_i = |G_{y_i}^Q|$. Cost matrix $\boldsymbol{C}_i$ is augmented by appending a new row and column filled with a learnable dustbin parameter $\alpha$. We convert $\boldsymbol{C}_i$ into a soft assignment matrix $\boldsymbol{Z}_i$ with the Sinkhorn algorithm, which serves as the confidence matrix for candidate matches. Point correspondences are determined through mutual top-$k$ selection, identifying matches as those ranking among the top $k$ entries in their respective rows and columns: $C_i = \{(G_{x_i}^P(x_j), G_{y_i}^Q(y_j)) | (x_j, y_j) \in \text{mutual\_top}_k(\boldsymbol{Z}_i^{x,y})\}$. The correspondences from each superpoint match are combined to construct dense correspondences as $C = \bigcup_{i=1}^{N_c} C_i$. They are then used to regress the rigid pose parameters.

Subsequently, the computed pose is employed to assess the overlap of point clouds via a nearest neighbor search in the Euclidean space as done in [82]. This overlap score is considered a 3D overlap prior. In contrast to prior work, we reintegrate this score into the attention learning module, which restarts the learning process with updated attention masks. This iterative procedure is repeated $I$ times, progressively learning the overlap-aware attention scores, thereby enabling accurate matching of dense 3D point pairs.

Given the estimated relative transforms, we construct pose graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V}$ representing the global poses of individual point clouds and $\mathcal{E}$ denoting the relative transforms between them, estimated previously. We will now discuss the proposed multiway algorithm: Wednesday.

### 3.2. Global Rotation Averaging

The goal of rotation averaging is to deduce the global orientations decoupled from the positions given pairwise rotations. We employ the method of [17] to obtain global point cloud orientations, finding it exceptionally efficient for our problem. In brief, this method performs a two-step procedure. First, it employs an L1 optimization to yield coarse rotation estimates robust to outliers. Next, it utilizes an iteratively re-weighted least-squares approach to refine these initial estimates and obtain accurate global rotations.

Selecting the somewhat older method of [17] might seem unconventional when aiming for a pipeline that de-

| Method | NSS [61] | | | 3DMatch [84] | | | 3DLoMatch [37] | | | KITTI [37] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RR (%)↑ | RRE (°)↓ | RTE (m)↓ | RR (%)↑ | RRE (°)↓ | RTE (m)↓ | RR (%)↑ | RRE (°)↓ | RTE (m)↓ | RR (%)↑ | RRE (°)↓ | RTE (cm)↓ |
| FPFH [57] | 11.70 | 45.32 | 2.23 | 0.851 | – | – | – | – | – | – | – | – |
| FCGF [21] | 24.43 | 39.89 | 2.04 | – | – | – | 0.401 | – | – | 96.0 | 0.30 | 9.5 |
| D3Feat [6] | 22.73 | 33.09 | 2.26 | 0.816 | – | – | 0.372 | – | – | **99.8** | 0.30 | 7.2 |
| RegTR [80] | – | – | – | 0.919 | 5.31 | 0.170 | 0.646 | 23.05 | 0.644 | – | – | – |
| Predator [37] | 64.97 | 13.52 | 0.65 | 0.893 | 6.80 | 0.202 | 0.604 | 30.07 | 0.762 | **99.8** | 0.27 | 6.8 |
| GeoTr. [55] | 39.07 | 22.93 | 0.99 | 0.925 | 7.04 | 0.194 | 0.741 | 23.15 | 0.583 | **99.8** | 0.24 | 6.8 |
| PEAL [82] | 58.72 | 15.78 | 0.71 | 0.941 | 4.23 | 0.152 | 0.788 | 15.79 | 0.485 | **99.8** | 0.23 | 6.8 |
| **ODIN** | **69.73** | **11.96** | **0.54** | **0.958** | **3.15** | **0.108** | **0.812** | **12.61** | **0.402** | **99.8** | **0.14** | **3.6** |

Table 1. **Pairwise point cloud registration** on the NSS [61], 3DMatch [84], 3DLoMatch [37] and KITTI [31] datasets. The reported metrics are the Registration Recall (RR), which measures the fraction of successfully registered pairs; the Relative Rotation Error (RRE); and the Relative Translation Error (RTE). The best results are in **bold**.

livers state-of-the-art accuracy. However, this approach proved to be the most applicable out-of-the-box solution for our problem. We explored several recent learning-based alternatives, including NeuRoRa [53], PoGO-Net [46], MSP [77], and DMF-Net [64]. Although NeuRoRa and MSP demonstrate commendable accuracy in their experiments, we could not reproduce these results, even after retraining the models. PoGO-Net lacks a public implementation. DMF-Net does not yield better results than [17] according to their own experiments, and its optimization process is particularly time-consuming, especially when compared to [17] that runs only for a few seconds in practice.

### 3.3. Optimal Robust Translation Re-estimation

The next phase in our pipeline is the re-estimation of relative translations $\boldsymbol{t}_{ij}$, using the estimated global point cloud orientations $\{\boldsymbol{R}_i\}_{i\in[1,n]}$. This step is crucial as the initial relative translations are computed alongside rotations, posing a more complex problem than the robust estimation of Euclidean translations with known orientations. This phase is key to achieving better translation initialization for the subsequent position averaging stage.

Given rotations $\boldsymbol{R}_i$ and $\boldsymbol{R}_j$, and point correspondence $(\boldsymbol{X}_i, \boldsymbol{X}_j) \in \mathcal{X}_{ij}$ in frames $i$ and $j$, we have constraint $\boldsymbol{R}_j\boldsymbol{X}_j = \boldsymbol{R}_i\boldsymbol{X}_i + \boldsymbol{t}_{ij}$, where $\boldsymbol{t}_{ij} \in \mathbb{R}^3$ is the unknown relative translation. With known $\boldsymbol{R}_i$ and $\boldsymbol{R}_j$, we derive equation

$$\boldsymbol{t}_{ij} = \boldsymbol{R}_j\boldsymbol{X}_j - \boldsymbol{R}_i\boldsymbol{X}_i \tag{3}$$

to calculate the updated translation from a single correspondence. To address the outlier correspondences, methods like RANSAC [29] or L1 optimization [35] could be applied. However, our problem allows for solving the maximum consensus problem – finding the model with the highest number of inliers – in a globally optimal fashion.

Assuming an inlier-outlier threshold $\epsilon \in \mathbb{R}^+$, our objective is to find optimal translation

$$\boldsymbol{t}^*_{ij} = \arg\max_{\boldsymbol{t}_{ij}} \sum_{(\boldsymbol{X}_i,\boldsymbol{X}_j)\in\mathcal{X}_{ij}} \left[\!\left[ |\boldsymbol{t}_{ij} - \boldsymbol{R}_j\boldsymbol{X}_j + \boldsymbol{R}_i\boldsymbol{X}_i|_2 < \epsilon \right]\!\right],$$

where $[\![\cdot]\!]$ is the Iverson bracket that equals 1 if the condition inside holds and 0 otherwise. In our case, this translates to finding the maximum overlap of 3D spheres.

For each translation estimate $\hat{\boldsymbol{t}}_{ij}$ coming from a 3D correspondence via Eq. 3, the points that lead to translations falling inside a sphere with radius $\epsilon$ centered on $\hat{\boldsymbol{t}}_{ij}$ are the inliers. Therefore, finding the 3D subspace with the highest sphere density yields the solution with the maximum inliers. This is a similar process as proposed for 1D problems in [8].

Analytically solving the maximum sphere overlap problem is complex. Instead, we propose a fast "branch-and-bound"-like numerical approach that achieves global optimum. The process starts by iterating through all correspondences $(\boldsymbol{X}^k_i, \boldsymbol{X}^k_j) \in \mathcal{X}_{ij}$, calculating the implied translations $\hat{\boldsymbol{t}}^k_{ij}$. We then create a uniform 3D grid where each cell (*i.e.*, a 3D box) counts the number of intersecting spheres. The cell, with a size arbitrarily set to $\epsilon$ for this stage, allows efficient calculation of box-sphere intersections by intersecting the boundary planes of a cell with the spheres. The intersections can fall into three categories: (i) The sphere does not intersect any boundary planes, and the distance functions indicate it is outside the box, (ii) the sphere intersects one or more planes, or (iii) the sphere is entirely within the box as indicated by the distance functions. We repeatedly zoom into the cells with the highest sphere density, re-initializing a uniform grid on these selected cells. The procedure stops at the zoom level, where the selected cells fall inside the affected spheres without intersecting their boundaries. As we may end up with multiple cell candidates, we choose the one with the highest inlier number. Finally, the translation is estimated by least squares fitting on all inliers.

It is important to note that with a sparse grid and proper hashing functions, this procedure is $\mathcal{O}(|\mathcal{X}|)$, scaling linearly with the number of correspondences. In practice, only the first step checks all correspondences, with subsequent steps focusing on a significantly smaller set of candidates.

### 3.4. Translation Optimization

Given the estimated global orientations and refined relative translations, the next step is estimating the global positions of the point clouds. To do so, we employ a Levenberg-Marquardt numerical optimization implemented in the Ceres library, minimizing pairwise constraints $\boldsymbol{t}_{ij} = \boldsymbol{R}_i(\boldsymbol{t}_j - \boldsymbol{t}_i)$. We use the truncated soft-L1 robust loss.

| Method | NSS | | | 3DMatch | | 3DLoMatch | | KITTI | |
|---|---|---|---|---|---|---|---|---|---|
| | RE (°)↓ | TE (m)↓ | | RE (°)↓ | TE (cm)↓ | RE (°)↓ | TE (cm)↓ | RE (°)↓ | TE (cm)↓ |
| Predator | 13.43 | 0.65 | PEAL | 4.72 | 15.8 | 16.03 | 50.2 | 9.46 | 11.85 |
| + Open3d [20] | 12.76 | 0.64 | + Open3d | 4.72 | 15.8 | 14.23 | 45.1 | 6.21 | 7.72 |
| + DeepMapping2 [18] | 11.54 | 0.64 | + DeepM. | 4.23 | 14.5 | 13.25 | 39.4 | 3.34 | 6.04 |
| + LMPR [33] | 11.35 | 0.62 | + LMPR | 3.98 | 12.6 | 13.07 | 37.3 | 6.79 | 7.89 |
| + LIRTS [79] | 11.42 | 0.61 | + LIRTS | 3.95 | 12.0 | 11.52 | 36.0 | 5.17 | 6.94 |
| + RMPR [70] | 10.87 | 0.62 | + RMPR | 3.57 | 11.6 | 10.18 | 34.4 | 4.69 | 6.38 |
| + **Wednesday** | 2.24 | 0.51 | + **Wednesday** | 2.58 | 9.4 | 7.21 | 29.1 | 2.52 | 5.92 |
| **ODIN + Wednesday** | **2.01** | **0.42** | | **2.32** | **8.4** | **6.44** | **26.5** | **2.18** | **4.76** |

Table 2. **Multiway point cloud registration** on the NSS [61], 3DMatch [84], 3DLoMatch [37] and KITTI [31] datasets. The reported metrics are the average rotation (RE) and translation errors (TE). For each dataset, we choose the best-performing pairwise estimator from the baselines (see Table 1). We run Predator [37] on NSS and PEAL [82] on the other datasets. The best results are in **bold**.

### 3.5. Pose Graph Optimization with Diffusion

Given the global poses estimated in the previous sections, the last step of the algorithm is a joint optimization of positions and orientations. Inspired by [71], we design a denoising network to model the conditional probability $p(\boldsymbol{R}, \boldsymbol{t} \mid \mathcal{K})$ of the samples $(\boldsymbol{R}, \boldsymbol{t})$ given the set of input point clouds $\mathcal{K}$. Probability $p(\boldsymbol{R}, \boldsymbol{t} \mid \mathcal{K})$ is first estimated by training a diffusion model $\mathcal{D}_\theta$ on point clouds with ground truth poses from a training set. At inference time, for a new set of point clouds $\mathcal{K}$, we sample $p(\boldsymbol{R}, \boldsymbol{t} \mid \mathcal{K})$ to estimate the corresponding global pose $\boldsymbol{R}, \boldsymbol{t}$.

The denoising process is conditioned on the input point cloud set $\mathcal{K}$, as $p_\theta(\boldsymbol{R}_{t-1}, \boldsymbol{t}_{t-1} \mid \boldsymbol{R}_t, \boldsymbol{t}_t, \mathcal{K}) =$

$$\mathcal{N}\left(\boldsymbol{R}_{t-1}, \boldsymbol{t}_{t-1}; \sqrt{1-\beta_t}\mathcal{D}_\theta(\boldsymbol{R}_t, \boldsymbol{t}_t, t, \mathcal{K}), (1-\beta_t)\mathcal{K}\right).$$

Denoiser $\mathcal{D}_\theta$ is implemented as a Transformer, which accepts a sequence of noisy poses $\boldsymbol{R}_t^i, \boldsymbol{t}_t^i$, diffusion time $t$, and feature embeddings $\psi(P^i) \in \mathbb{R}^{D_\psi}$ of the input point cloud $\mathcal{K}^i$. The denoiser outputs the tuple of corresponding denoised pose parameters $\mu_{t-1} = (\mu_{t-1}^i)_{i=1}^N$. The feature embeddings come from a pretrained KPConv.
At train time, $\mathcal{D}_\theta$ is supervised by denoising loss $\mathcal{L}_{\text{diff}} =$

$$\mathbb{E}_{\boldsymbol{R}_t, \boldsymbol{t}_t \sim p(\boldsymbol{R}_t, \boldsymbol{t}_t \mid \boldsymbol{R}_0, \boldsymbol{t}_0, \mathcal{K})}^{t \sim \mathcal{U}[1,T]} \left\| \mathcal{D}_\theta(\boldsymbol{R}_t, \boldsymbol{t}_t, t, P) - (\boldsymbol{R}_t.\boldsymbol{t}_t) \right\|^2.$$

The relative poses from the pairwise registration constrain the whole pose graph. The error implied by an edge is $\epsilon(\boldsymbol{R}_{ij}, \boldsymbol{t}_{ij}) = \sqrt{\frac{1}{|\mathcal{C}|}\sum_{(p,q)\in\mathcal{C}_{ij}} \|\boldsymbol{R}p_i + \boldsymbol{t} - q_j\|_2^2}$, where $\mathcal{C}_{ij}$ are the point correspondences in the point clouds. The additional loss we designed to be minimized is $\mathcal{L} = \sum_{(i,j)\in\mathcal{E}} \min(\epsilon(\boldsymbol{R}_{ij}, \boldsymbol{t}_{ij}), \gamma)$, where $\mathcal{E}$ is the pose graph edges and $\gamma$ is a threshold parameter.

The main differences compared to the original method in [71] are the edge loss, measuring global pose consistency, and the employed denoising network architecture.

## 4. Experiments

**Datasets.** To evaluate the proposed algorithms both on pairwise and multiway registration tasks, we use the *3DMatch* [84], *3DLoMatch* [37], *KITTI* [31], and *NSS* [61] datasets. The *3DMatch* [84] dataset contains 62 indoor scenes, with 46 used for training, 8 for validation, and 8 for testing. We use the training data preprocessed by Huang et al. [37] and evaluate on both *3DMatch* and *3DLoMatch* [37] protocols. The point cloud pairs in *3DMatch* have more than 30% overlap, whereas those in *3DLoMatch* have a low overlap of 10% - 30%. The *KITTI* odometry dataset [31] contains 11 sequences of LiDAR-scanned outdoor driving scenarios. We follow [6, 21, 37, 54] and split it into train/val/test sets as follows: sequences 0-5 for training, 6-7 for validation and 8-10 for testing. As in [6, 21, 37, 54], we refine the provided ground truth poses using ICP [10] and only use point cloud pairs that are captured within 10m range of each other. For multiway KITTI, we followed [18] and decreased the frame rate 20 times to avoid saturated results. The *NSS* dataset represents 6 large-scale construction sites and their rescans over time. The data depicts the interior layout construction from creating walls, to adding pipes and air-ducts, and to machinery moving around. It contains spatial and spatiotemporal pairs and has annotations for both pairwise and multiway registration.

**Metrics.** For evaluating pairwise methods, we follow prior work [37, 42, 54, 80]. We compute the Registration Recall (RR), which measures the fraction of successfully registered pairs; relative rotation (RRE); and relative translation errors (RTE). We calculate the average values over all valid pairs and scenes. For multiway registration, we calculate the average rotation (RE; in degrees) and position (TE) errors given the ground truth pose parameters.

**Pairwise Point Cloud Registration.** The results of the standard setting on the **NSS** dataset are in Table 1 (1st col.). ODIN substantially improves compared with the state-of-the-art methods. We improve upon the recent PEAL [82] by a margin of 11% in terms of recall while reducing the trans. error by ≈0.2m and the rot. one by ≈4°. Interestingly, the second best method on this dataset is Predator [37], which also significantly lags behind the proposed ODIN.

The results on **3DMatch** are in Table 1 (2nd). While all methods are accurate, ODIN still manages to reduce the rotation by 1° and translation errors by 0.05 meters compared to the second best algorithm, PEAL. This improvement is

also reflected in the recall, which is improved by 1.7%.

The results on **3DLoMatch** are shown in Table 1 (3rd). On this dataset, we achieve significant improvements compared with other baselines. We improve upon the second most accurate method (GeoTransformer) by reducing the average rotation error by 10° and the translation error by 0.18 meters. This accuracy improvement pushes our recall score up by 7.1% compared to that of GeoTransformer.

The pairwise registration results on the **KITTI** dataset are in Table 1 (4th). While all methods perform very accurately, mainly due to the small baselines between subsequent frames, the proposed ODIN is the best in all metrics. Notably, it almost halves the rotation and position errors of the second most accurate method (PEAL). This clearly shows the advantages of the proposed two-stream architecture with attention masking and diffusion-based denoising.

**Multiway Point Cloud Registration.** We compare Wednesday to [20] implemented in Open3d, DeepMapping2 [18], LMPR [33], LIRTS [79], and RMPR [70]. We train all learning-based methods on the training set of each dataset. To perform pairwise registration, we select the best-performing baseline method on each scene. Thus, we run Predator [37] on NSS and PEAL [82] on the other datasets. Also, we show the results with the proposed ODIN.

The results on all datasets are reported in Table. 2. The proposed Wednesday consistently improves upon *all* state-of-the-art algorithms, often by a substantial margin. For example, the rotations errors on NSS are reduced to 20%. The position errors on 3DMatch and 3DLoMatch are reduced by 2.2 and 5.3 meters, respectively. The rotation errors on KITTI are halved. Using ODIN as a pairwise estimator further reduces the registration errors.

We show results of the proposed method and [70, 79] in Fig. 3. We chose [70, 79] as they are the best-performing alternatives in Table 2, still, they fail entirely. While the proposed method also has inaccuracies compared to the ground truth, it provides significantly better registrations.

**Ablation Studies.** The ablation study of pairwise registration on the 3DLoMatch dataset is in Table 3. We tested ODIN without the proposed procedure of reintegrating the overlap score predictions from previous iterations, and without diffusion-based correlation matrix denoising. It is evident that both advancements have a clear and individual impact on the increased accuracy.

We performed a similar ablation study on Wednesday on the NSS dataset, using ODIN to initialize the pairwise relative poses. The results are in Table 4. The diffusion-based optimization on its own leads to the highest rotations and second highest translation errors, justifying the need for the rest of the proposed pipeline. This is expected as diffusion essentially aims at noise reduction, while the pose graph is not only noisy but contains outliers, necessitating robust estimation. Rotation and translation averaging leads to better results than the diffusion-based process. It is further improved by re-estimating the translations. The best results are obtained when all proposed components are employed.

| Method | RR (%)↑ | RRE (°)↓ | RTE (m)↓ |
|---|---|---|---|
| w/o Overlap and Diffusion | 0.741 | 23.15 | 0.583 |
| w/o Overlap | 0.791 | 14.76 | 0.442 |
| w/o Diffusion | 0.803 | 13.57 | 0.419 |
| **ODIN** | **0.812** | **12.61** | **0.402** |

Table 3. Pairwise registration recall (RR), rotation (RRE) and translation errors (RTE) of the proposed ODIN on the 3DLoMatch dataset without overlap scores or denoising the correlation matrix.

| | D | R+TA | R+TR+TA | R+TA+D | R+TR+TA+D |
|---|---|---|---|---|---|
| RE (°)↓ | 5.21 | 4.05 | 4.05 | 2.06 | **2.01** |
| TE (m)↓ | 0.51 | 0.53 | 0.48 | 0.47 | **0.42** |

Table 4. Multiway registration average rotation and translation errors on the NSS dataset with combinations of the proposed components: (R) rotation and (TA) translation averaging, (TR) translation re-estimation, and (D) diffusion-based pose optimization.

| Method | Runtime (s) ↓ | Inlier Number ↑ |
|---|---|---|
| Exhaustive RANSAC | 0.12 | 18.03 |
| Proposed (Sec. 3.3) | **0.05** | **18.11** |

Table 5. Average runtime (secs) and inlier number of exhaustive RANSAC and the proposed optimal estimator on the NSS dataset.

In Table 5, we demonstrate that the proposed globally optimal translation re-estimation outperforms running exhaustive RANSAC – estimating the translation from each correspondence and selecting the one with the most inliers. The proposed approach is 2.4 times faster while, as expected, it finds more inliers. Please note that this is the run-time on a single point cloud pair, thus the difference is more significant on the entire set of input point clouds.

## 5. Conclusion

We present Wednesday, a novel framework for multiway point cloud mosaicking, or else, aligning a collection of point clouds into a unified coordinate system. It starts with a new pairwise registration method (ODIN) which delivers significantly more accurate results compared to state-of-the-art ones. The pipeline proceeds with rotation and translation averaging to establish the global pose of each point cloud. We also incorporate a globally optimal robust translation re-estimation algorithm to ensure the precision of pairwise translations after acquiring global orientations. Finally, a diffusion-based optimization approach finalizes the output poses. The pipeline leads to substantial improvement over the state-of-the-art algorithms, exemplified by an 80% reduction in rotation error on the NSS dataset. The consistent and significant improvements on all tested large-scale datasets position the proposed algorithm as the new benchmark in both pairwise and multiway point cloud registrations.

## References


[1] Dror Aiger, Niloy J Mitra, and Daniel Cohen-Or. 4-points congruent sets for robust pairwise surface registration. In *ACM SIGGRAPH 2008 papers*, pages 1–10. 2008. 2

[2] Mica Arie-Nachimson, Shahar Z Kovalsky, Ira Kemelmacher-Shlizerman, Amit Singer, and Ronen Basri. Global motion estimation from point matches. In *2012 Second international conference on 3D imaging, modeling, processing, visualization & transmission*, pages 81–88. IEEE, 2012. 2

[3] Federica Arrigoni, Luca Magri, Beatrice Rossi, Pasqualina Fragneto, and Andrea Fusiello. Robust absolute rotation estimation via low-rank and sparse matrix decomposition. In *2014 2nd International Conference on 3D Vision*, pages 491–498. IEEE, 2014. 2

[4] Federica Arrigoni, Beatrice Rossi, and Andrea Fusiello. Spectral synchronization of multiple views in se (3). *SIAM Journal on Imaging Sciences*, 9(4):1963–1990, 2016. 2, 3

[5] First Author, Second Contributor, and Third Other. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the International Conference on Learning Representations*, pages 123–134. ICLR, 2023. 4

[6] Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, and Chiew-Lan Tai. D3feat: Joint learning of dense detection and description of 3d local features. In *CVPR*, 2020. 6, 7

[7] Daniel Barath and Jiří Matas. Graph-cut ransac. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6733–6741, 2018. 1

[8] Daniel Barath, Yaqing Ding, Zuzana Kukelova, and Viktor Larsson. Image stitching with locally shared rotation axis. In *2021 International Conference on 3D Vision (3DV)*, pages 1382–1391. IEEE, 2021. 6

[9] Florian Bernard, Johan Thunberg, Peter Gemmar, Frank Hertel, Andreas Husch, and Jorge Goncalves. A solution for multi-alignment by transformation synchronisation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2161–2169, 2015. 2

[10] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606. Spie, 1992. 2, 7

[11] Uttaran Bhattacharya and Venu Madhav Govindu. Efficient and robust registration on the 3d special euclidean group. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2019. 2

[12] Tolga Birdal and Slobodan Ilic. Point pair features based object detection and pose estimation revisited. In *2015 International conference on 3D vision*, pages 527–535. IEEE, 2015. 2

[13] Tolga Birdal and Slobodan Ilic. Cad priors for accurate and flexible instance reconstruction. In *Proceedings of the IEEE international conference on computer vision*, pages 133–142, 2017. 2

[14] Tolga Birdal and Umut Simsekli. Probabilistic permutation synchronization using the riemannian structure of the birkhoff polytope. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11105–11116, 2019. 2

[15] Tolga Birdal, Umut Simsekli, Mustafa Onur Eken, and Slobodan Ilic. Bayesian pose graph optimization via bingham distributions and tempered geodesic mcmc. *Advances in neural information processing systems*, 31, 2018. 2

[16] Alvaro Parra Bustos and Tat-Jun Chin. Guaranteed outlier removal for point cloud registration with correspondences. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2868–2882, 2017. 1

[17] Avishek Chatterjee and Venu Madhav Govindu. Efficient and robust large-scale rotation averaging. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 521–528, 2013. 5, 6

[18] Chao Chen, Xinhao Liu, Yiming Li, Li Ding, and Chen Feng. Deepmapping2: Self-supervised large-scale lidar map optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9306–9316, 2023. 2, 7, 8

[19] Dmitry Chetverikov, Dmitry Svirko, Dmitry Stepanov, and Pavel Krsek. The trimmed iterative closest point algorithm. In *2002 International Conference on Pattern Recognition*, pages 545–548. IEEE, 2002. 2

[20] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5556–5565, 2015. 2, 3, 7, 8

[21] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *ICCV*, 2019. 6, 7

[22] Zhaopeng Cui and Ping Tan. Global structure-from-motion by similarity averaging. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 864–872, 2015. 3

[23] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In *Proceedings of the European conference on computer vision*, pages 602–618, 2018. 1

[24] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppfnet: Global context aware local features for robust 3d point matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 195–205, 2018. 1

[25] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 4

[26] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 998–1005. Ieee, 2010. 2

[27] Felix Endres, Jürgen Hess, Jürgen Sturm, Daniel Cremers, and Wolfram Burgard. 3-d mapping with an rgb-d camera. *IEEE transactions on robotics*, 30(1):177–187, 2013. 2

[28] Simone Fantoni, Umberto Castellani, and Andrea Fusiello. Accurate and automatic alignment of range surfaces. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 73–80. IEEE, 2012. 2

[29] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1, 2, 6

[30] Alex Flint, Anthony Dick, and Anton Van Den Hengel. Thrift: Local 3d structure recognition. In *Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications*, pages 182–188. IEEE, 2007. 1, 2

[31] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 6, 7

[32] Zan Gojcic, Caifa Zhou, Jan D Wegner, and Andreas Wieser. The perfect match: 3d point cloud matching with smoothed densities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5545–5554, 2019. 1

[33] Zan Gojcic, Caifa Zhou, Jan D Wegner, Leonidas J Guibas, and Tolga Birdal. Learning multiview 3d point cloud registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1759–1769, 2020. 2, 3, 7, 8

[34] Venu Madhav Govindu. Lie-algebraic averaging for globally consistent motion estimation. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, pages I–I. IEEE, 2004. 2

[35] Richard Hartley, Khurrum Aftab, and Jochen Trumpf. L1 rotation averaging using the weiszfeld algorithm. In *CVPR 2011*, pages 3041–3048. IEEE, 2011. 6

[36] Stefan Hinterstoisser, Vincent Lepetit, Naresh Rajkumar, and Kurt Konolige. Going further with point pair features. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 834–848. Springer, 2016. 2

[37] 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. 1, 2, 6, 7, 8

[38] Xiangru Huang, Zhenxiao Liang, Chandrajit Bajaj, and Qixing Huang. Translation synchronization via truncated least squares. *Advances in neural information processing systems*, 30, 2017. 3

[39] Xiangru Huang, Zhenxiao Liang, Xiaowei Zhou, Yao Xie, Leonidas J Guibas, and Qixing Huang. Learning transformation synchronization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8082–8091, 2019. 2, 3

[40] Daniel F Huber and Martial Hebert. Fully automatic registration of multiple 3d data sets. *Image and Vision Computing*, 21(7):637–650, 2003. 2

[41] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, et al. Conceptfusion: Open-set multimodal 3d mapping. 2023. 1

[42] Shengze Jin, Daniel Barath, Marc Pollefeys, and Iro Armeni. Q-REG: End-to-end trainable point cloud registration with surface curvature. *arXiv preprint arXiv:2309.16023*, 2023. 2, 7

[43] Andrew E Johnson and Martial Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on pattern analysis and machine intelligence*, 21(5):433–449, 1999. 1, 2

[44] Marc Khoury, Qian-Yi Zhou, and Vladlen Koltun. Learning compact geometric features. In *Proceedings of the IEEE international conference on computer vision*, pages 153–161, 2017. 1

[45] Hongdong Li and Richard Hartley. The 3d-3d registration problem revisited. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007. 2

[46] Xinyi Li and Haibin Ling. Pogo-net: pose graph optimization with graph neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5895–5905, 2021. 6

[47] Weixin Lu, Guowei Wan, Yao Zhou, Xiangyu Fu, Pengfei Yuan, and Shiyu Song. Deepvcp: An end-to-end deep neural network for point cloud registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12–21, 2019. 1, 2

[48] Weixin Lu, Yao Zhou, Guowei Wan, Shenhua Hou, and Shiyu Song. L3-net: Towards learning based lidar localization for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6389–6398, 2019. 1

[49] Eleonora Maset, Federica Arrigoni, and Andrea Fusiello. Practical and efficient multi-view matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4568–4576, 2017. 2

[50] Nicolas Mellado, Dror Aiger, and Niloy J Mitra. Super 4pcs fast global pointcloud registration via smart indexing. In *Computer graphics forum*, pages 205–215. Wiley Online Library, 2014. 2

[51] Ajmal S Mian, Mohammed Bennamoun, and Robyn Owens. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1584–1601, 2006. 2

[52] Matteo Palieri, Benjamin Morrell, Abhishek Thakur, Kamak Ebadi, Jeremy Nash, Arghya Chatterjee, Christoforos Kanellakis, Luca Carlone, Cataldo Guaragnella, and Ali-akbar Agha-Mohammadi. Locus: A multi-sensor lidar-centric solution for high-precision odometry and 3d mapping in real-time. *IEEE Robotics and Automation Letters*, 6(2):421–428, 2020. 1

[53] Pulak Purkait, Tat-Jun Chin, and Ian Reid. Neurora: Neural robust rotation averaging. In *European Conference on Computer Vision*, pages 137–154. Springer, 2020. 6

[54] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *CVPR*, 2022. 7

[55] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, Slobodan Ilic, Dewen Hu, and Kai Xu. Geotrans-

former: Fast and robust point cloud registration with geometric transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2, 3, 4, 5, 6

[56] Tahir Rabbani, Sander Dijkman, Frank van den Heuvel, and George Vosselman. An integrated approach for modelling and global registration of point clouds. *ISPRS journal of Photogrammetry and Remote Sensing*, 61(6):355–370, 2007. 1

[57] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *ICRA*, 2009. 6

[58] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *IEEE international conference on robotics and automation*, pages 3212–3217. IEEE, 2009. 1, 2

[59] Paul-Edouard Sarlin, Mihai Dusmanu, Johannes L Schönberger, Pablo Speciale, Lukas Gruber, Viktor Larsson, Ondrej Miksik, and Marc Pollefeys. Lamar: Benchmarking localization and mapping for augmented reality. In *European Conference on Computer Vision*, pages 686–704. Springer, 2022. 1

[60] Frank Steinbrucker, Christian Kerl, and Daniel Cremers. Large-scale multi-resolution surface reconstruction from rgb-d sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3264–3271, 2013. 2

[61] Tao Sun, Yan Hao, Shengyu Huang, Silvio Savarese, Konrad Schindler, Marc Pollefeys, and Iro Armeni. Nothing stands still: A spatiotemporal benchmark on 3d point cloud registration under large geometric and temporal change, 2023. 2, 5, 6, 7

[62] Yuichi Taguchi, Yong-Dian Jian, Srikumar Ramalingam, and Chen Feng. Point-plane slam for hand-held 3d sensors. In *2013 IEEE international conference on robotics and automation*, pages 5182–5189. IEEE, 2013. 1

[63] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1

[64] Gk Tejus, Giacomo Zara, Paolo Rota, Andrea Fusiello, Elisa Ricci, and Federica Arrigoni. Rotation synchronization via deep matrix factorization. 2023. 2, 6

[65] Pascal Willy Theiler, Jan Dirk Wegner, and Konrad Schindler. Keypoint-based 4-points congruent sets–automated marker-less registration of laser scans. *ISPRS journal of photogrammetry and remote sensing*, 96:149–163, 2014. 2

[66] Pascal Willy Theiler, Jan Dirk Wegner, and Konrad Schindler. Globally consistent registration of terrestrial laser scans via graph optimization. *ISPRS journal of photogrammetry and remote sensing*, 109:126–138, 2015. 1, 2, 3

[67] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 4

[68] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique signatures of histograms for local surface description. In *European Conference on Computer Vision*, pages 356–369. Springer, 2010. 1, 2

[69] Andrea Torsello, Emanuele Rodola, and Andrea Albarelli. Multiview registration via graph diffusion of dual quaternions. In *CVPR 2011*, pages 2441–2448. IEEE, 2011. 2, 3

[70] Haiping Wang, Yuan Liu, Zhen Dong, Yulan Guo, Yu-Shen Liu, Wenping Wang, and Bisheng Yang. Robust multiview point cloud registration with reliable pose graph initialization and history reweighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9506–9515, 2023. 2, 3, 5, 7, 8

[71] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9773–9783, 2023. 2, 7

[72] Yue Wang and Justin M Solomon. Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3523–3532, 2019. 2

[73] Thomas Whelan, Michael Kaess, John J Leonard, and John McDonald. Deformation-based loop closure for large scale dense rgb-d slam. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 548–555. IEEE, 2013. 2

[74] Ryan W Wolcott and Ryan M Eustice. Visual localization within lidar maps for automated urban driving. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 176–183. IEEE, 2014. 1

[75] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE international conference on computer vision*, pages 1625–1632, 2013. 2

[76] 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. 2

[77] Luwei Yang, Heng Li, Jamal Ahmed Rahim, Zhaopeng Cui, and Ping Tan. End-to-end rotation averaging with multi-source propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11774–11783, 2021. 6

[78] Zi Jian Yew and Gim Hee Lee. 3dfeat-net: Weakly supervised local 3d features for point cloud registration. In *Proceedings of the European conference on computer vision*, pages 607–623, 2018. 1

[79] Zi Jian Yew and Gim Hee Lee. Learning iterative robust transformation synchronization. In *2021 International Conference on 3D Vision (3DV)*, pages 1206–1215. IEEE, 2021. 2, 3, 5, 7, 8

[80] Zi Jian Yew and Gim Hee Lee. Regtr: End-to-end point cloud correspondences with transformers. In *CVPR*, 2022. 6, 7

[81] Hao Yu, Zheng Qin, Ji Hou, Mahdi Saleh, Dongsheng Li, Benjamin Busam, and Slobodan Ilic. Rotation-invariant

transformer for point cloud matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5384–5393, 2023. 1, 2

[82] Junle Yu, Luwei Ren, Yu Zhang, Wenhui Zhou, Lili Lin, and Guojun Dai. Peal: Prior-embedded explicit attention learning for low-overlap point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17702–17711, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[83] Bernhard Zeisl, Kevin Koser, and Marc Pollefeys. Automatic registration of rgb-d scans via salient directions. In *Proceedings of the IEEE international conference on computer vision*, pages 2808–2815, 2013. 1

[84] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1802–1811, 2017. 1, 6, 7

[85] Qian-Yi Zhou and Vladlen Koltun. Dense scene reconstruction with points of interest. *ACM Transactions on Graphics (ToG)*, 32(4):1–8, 2013. 2

[86] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 766–782. Springer, 2016. 2, 3